Modeling metastatic progression from cross-sectional cancer genomics data

Kevin Rupp *[1, 2, 3], Andreas Lösch *[1], Y. Linda Hu *[1], Chenxi Nie[2], Rudolf Schill[1, 2, 3], Maren Klever[4], Simon Pfahler[5], Lars Grasedyck[4], Tilo Wettig[5], Niko Beerenwinkel †[2, 3] , Rainer Spang †[1]

[1] *Faculty of Informatics and Data Science - Statistical Bioinformatics Group, University of Regensburg, Am Biopark 9, 93053 Regensburg, Germany*

[2] *Department of Biosystems and Engineering, ETH Zurich, Schanzenstrasse 44, 4056 Basel, Switzerland*

[3] *SIB Swiss Institute of Bioinformatics, Schanzenstrasse 44, 4056 Basel, Switzerland*

[4] *Institute for Geometry and Applied Mathematics, RWTH Aachen, Templergraben 55, 52062 Aachen, Germany*

[5] *Faculty of Physics, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany*

Abstract

Metastasis formation is a hallmark of cancer lethality. Yet, metastases are generally unobservable during their early stages of dissemination and spread to distant organs. Genomic datasets of matched primary tumors and metastases may offer insights into the underpinnings and the dynamics of metastasis formation. We present metMHN, a cancer progression model designed to deduce the joint progression of primary tumors and metastases using cross-sectional cancer genomics data. The model elucidates the statistical dependencies among genomic events, the formation of metastasis, and the clinical emergence of both primary tumors and their metastatic counterparts. metMHN enables the chronological reconstruction of mutational sequences and facilitates estimation of the timing of metastatic seeding. In a study of nearly 5000 lung adenocarcinomas, metMHN pinpointed TP53 and EGFR as mediators of metastasis formation. Furthermore, the study revealed that post-seeding adaptation is predominantly influenced by frequent copy number alterations. All datasets and code are available on GitHub at https://github.com/cbg-ethz/metMHN.

1 Introduction

Metastasis is the primary cause of cancer-related death. It occurs as tumors evolve, when the primary lesion extends beyond its initial boundaries, invading adjacent healthy tissues, lymph nodes, and blood vessels. Cancer cells can then enter the bloodstream and spread to different locations within the body. At these new sites, the disseminated cells face novel selective pressures, leading to the elimination of many, but not all, cells. The survivors adapt and eventually colonize these foreign tissues, forming metastases [23].The development of cancer, or tumorigenesis, is predominantly driven by the progressive accumulation of genomic alterations, including somatic mutations and copy number alterations in cancer driver genes [40]. These alterations often result in divergent genotypes between a primary tumor and its associated metastasis. Extensive clinical sequencing efforts like the MSK-MET study [26] recently compiled genomic data from primary tumors and metastases. In principle, such datasets may inform about the timing and genetic mechanisms of metastasis formation, but revealing these pieces of information is challenging.

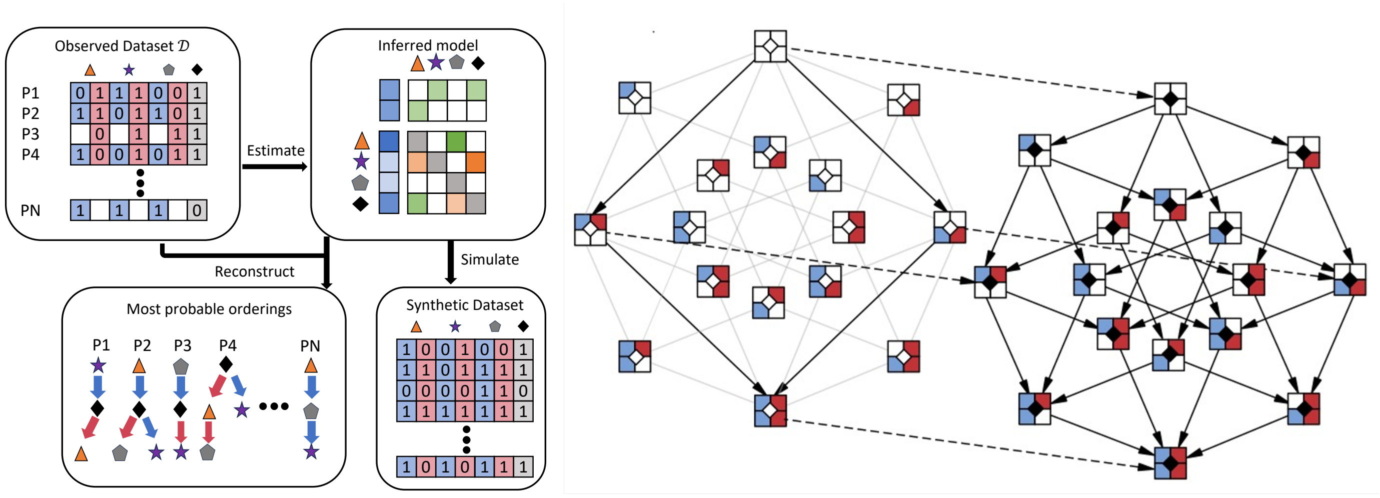
Cancer progression models aim to infer interactions between genomic alterations based on their co-occurrence patterns in cross-sectional data. Such models can then be used to both predict the future progression of tumors as well as to explain the past by inferring the order in which observed alterations accumulated. These models have their roots in the pioneering work of Fearon and Vogelstein [14]. Since then, a variety of models and algorithms have emerged to refine and expand upon this concept. They include Conjunctive Bayesian Networks [2], CAPRI [31], Network Aberration Models [19], HyperTraPS [18] and Mutual Hazard Networks [33]. All of these models only consider the progression of a single sequence and thus can not capture the divergent, branching patterns characteristic of metastatic disease progression. Therefore none of the above mentioned models can leverage the information provided by matched primary tumor and metastasis samples from the same patient. Methods like REVOLVER [7] or TreeMHN [24] can account for this branching behaviour as they model evolution of tumors on a clonal level. However, they require phylogenetic data and are not explicitly designed to model metastatic branching.

Here, we present Mutual Hazard Networks for metastatic disease (metMHN), a cancer progression model that captures the branching progression observed in primary tumors and their metastatic offshoots. The model is designed to infer interactions among genomic alterations and to assess their impact on the propensity for a tumor to seed a metastasis. Additionally, it accounts for metastasis-specific effects on the rates at which genomic alterations accumulate. metMHN

*These authors contributed equally.

†Correspondence: niko.beerenwinkel@bsse.ethz.ch, rainer.spang@ur.de

utilizes both cross-sectional data from matched primary tumors and metastases, and singular observations of only one of the two. It also models how genomic changes affect tumor observability. We demonstrate the utility and robustness of the metMHN model using the lung adenocarcinoma dataset (LUAD) provided by the provided by the Memorial Sloan-Kettering Cancer Center through AACR GENIE [30].



(a)

(b)

Figure 1: (a) Workflow of metMHN. In the top-left section, we show the types of input data that metMHN processes. Each row corresponds to a patient, each column to an event in the primary tumor (blue) or the metastasis (red). Events are represented by symbols and their status is encoded with a '1' for present, '0' for absent, or left blank if a tumor is unobserved. On the right, we present the primary output of metMHN: A network of interactions between events in matrix form. In the lower section, we show the most probable chronological ordering in which events accumulated in observed data points as inferred by metMHN. The progression trajectory of the primary tumor is indicated by blue arrows, while the trajectory of the metastasis is marked by red arrows. (b) The metMHN process and its state space: Black-bordered squares represent full states: the two compartments on the left detail the status of the primary tumor, the two on the right correspond to the metastasis, and the central diamond symbolizes the seeding event. The diagram is divided into two subspaces, with the left half constituting the subspace $\mathcal{S}_0$ and the right half comprising the subspace $\mathcal{S}_1$. Transitions between states that occur at non-zero rates are shown as solid black arrows. Transitions that are not possible in $\mathcal{S}_0$ but are possible in $\mathcal{S}_1$ are indicated by greyed-out arrows. Dotted arrows highlight transitions that influence the seeding event specifically.

2 Methods

metMHN extends the Mutual Hazard Network (MHN) framework, originally introduced by Schill et al. in 2020 [33] and further developed by Schill et al. in 2023 [34], which models the progression of primary tumors. We first establish the notation employed by MHNs and then introduce metMHN.

2.1 Mutual Hazard Networks

MHNs [33] model the evolution of primary tumors as a continuous-time Markov chain (CTMC) $\{X(t), t \geq 0\}$ on the binary state space $\{0,1\}^n$. A state $x \in \{0,1\}^n$ corresponds to a set of progression events, such as mutations or copy number alterations, where $x_i = 1$ indicates that event $i \in \{1, ..., n\}$ is present, whereas $x_i = 0$ indicates its absence. Let $\mathbf{p}(t) \in [0,1]^{2^n}$ denote the probability distribution over states at time t, where the states are ordered lexicographically. The evolution of the probability distribution over time is governed by the Kolmogorov forward equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = Q\mathbf{p}(0) \quad \text{solved by}$$

$$\mathbf{p}(t) = \exp(tQ)\mathbf{p}(0)\,. \tag{1}$$

Here $\mathbf{p}(0)$ denotes the distribution over states at the start of the progression. It is assumed that all tumors start in a wild type state, where no event has occurred yet, thus $\mathbf{p}(0) = (1, 0, \dots, 0)^T$. $Q \in \mathbb{R}^{2^n \times 2^n}$ denotes the transition rate matrix on the state space. Events are assumed to accumulate irreversibly and one at a time. Therefore, the only non-zero off-diagonal entries of Q are the transition rates from states $x = (\ldots, x_{i-1}, 0, x_{i+1}, \ldots)$ to $x_{+i} = (\ldots, x_{i-1}, 1, x_{i+1}, \ldots)$ that differ by exactly one event i. The transition rates are parameterized by a much smaller matrix $\Theta \in \mathbb{R}^{n\times n}_{\geq 0}$ as

$$Q_{x_{+i},x} = \Theta_{i,i} \prod_{x_j=1} \Theta_{i,j} \,. \tag{2}$$

Here $\Theta_{i,i}$ denotes the base rate with which event i spontaneously occurs in a tumor and $\Theta_{i,j}$ the multiplicative effect of the presence of event j on the rate of event i. The age of a tumor at the time of its diagnosis is unknown. In [33] it is assumed to be exponentially distributed with mean 1 and independent of the state of the tumor. Marginalizing over t in Equation (1) yields the time-marginal distribution

$$\mathbf{p} := \int_0^\infty \exp(tQ)\mathbf{p}(0)\mathrm{d}t = (I - Q)^{-1}\mathbf{p}(0)\,, \tag{3}$$

where I denotes the identity matrix. Let $\mathbf{p}_x$ denote the probability of observing a tumor in state x. Then the average log-likelihood for a dataset $\mathcal{D}$ of tumor states is defined as

$$l_{\mathcal{D}}(\Theta) = \frac{1}{|\mathcal{D}|} \sum_{x\in\mathcal{D}} \log \mathbf{p}_x \,. \tag{4}$$

The matrix Q does not need to be stored explicitly, because it can be written as a sum of tensor products. By using tensor operations, $\mathbf{p}$ can be calculated efficiently and Θ can be learned with a time and space complexity only exponential in the number of events that have occurred for each tumor in the dataset, rather than exponential in $2n$ [32, 5]. Recently [22, 16, 28] reduced the complexity further to n^3 using modern tensor formats and thus made MHN applicable to even larger state spaces.

Clearly, a tumor can only appear in a dataset after it has been clinically detected. This detection, in turn, is influenced by the tumor's genotype, as certain mutations can induce growth or alter the tumor's morphology. Such changes may result in symptoms that lead to the tumor's discovery, followed by its diagnosis, surgical removal, and eventual sequencing. Therefore the rate of observation should be dependent on the state of the tumor. In [34], the observation of a tumor was introduced as a separate event with its own set of parameters $\Omega \in \mathbb{R}^n_{>0}$. The observation of a state x happens at a rate $u_x = \prod_{x_j=1} \Omega_j$, where Ω_j is a multiplicative effect of the presence of event j on the rate of observation. On the other hand multiplicative effects of the observation on other events are set to 0. Thus, as soon as the observation event occurs, progression is halted. States where the observation occurred are thus absorbing states of the Markov chain. Then the probability distribution at observation is equal to the stationary distribution $\mathbf{p}(\infty)$ and given by

$$\mathbf{p}(\infty) = U(U - Q)^{-1}\mathbf{p}(0) = (I - QU^{-1})^{-1}\mathbf{p}(0)\,, \tag{5}$$

with $U = \mathrm{diag}((u_x)_x) \in \mathbb{R}^{2^n\times 2^n}_{>0}$ and Q and $\mathbf{p}(0)$ defined as in Equation (3) [34].

2.2 metMHN

We now present metMHN, an extension of the original MHN framework, meticulously tailored to analyze the dynamics of metastatic cancers.

2.2.1 Dynamics on the combined state space

With metMHN, we model the joint progression of primary tumors and metastases as a Markov process on the combined event space of both tumor entities (see Figure 1b). Formally, we consider a CTMC $\{X(t), t \geq 0\}$ on the state space $\mathcal{S} := \{\{0,1\} \times \{0,1\}\}^n \times \{0,1\}$. A state $x \in \mathcal{S}$ is represented by a bit string of length $2n+1$. Each of the n progression events is encoded by two bits. The first bit x_{i_P} indicates the status of event $i \in \{1, \ldots, n\}$ in the primary tumor, and the second bit x_{i_M} indicates the status of event i in the metastasis. We use the notations $\mathrm{PT}(x) = (x_{i_\mathrm{P}})$ and $\mathrm{MT}(x) = (x_{i_\mathrm{M}})$ for i in $\{1, \ldots, n\}$ to refer to the genotypes of the primary tumor and the metastasis respectively. The $(n+1)^{\text{th}}$ event is encoded by one bit only and indicates the status of the seeding event. In the model context, the seeding event denotes that the progression of the metastasis has become decoupled from the progression of the primary tumor. Analogous to MHN we parameterize all transition rates by a low-dimensional set of parameters $\Theta \in \mathbb{R}^{(n+1)\times(n+1)}$, where $\Theta_{i,i}$ refers to the base rate of event i and $\Theta_{i,j}$ to the effect of event j on the rate of event i. Before and after the seeding of a metastasis we assume different transition dynamics, which we describe in the following paragraphs.

Prior to seeding, the (soon-to-be) metastasis is identical to the primary tumor. Thus, events occur simultaneously in the primary tumor and the metastasis. Formally, we can describe these dynamics by a CTMC on the subspace

$\mathcal{S}_0 := \{\{0,1\}\times\{0,1\}\}^n \times \{0\} \subset \mathcal{S}$ with transition rate matrix $Q_0 \in \mathbb{R}^{2^{2n+1}\times 2^{2n+1}}$. Let $x = (\dots, x_{(i-1)_\mathrm{M}}, 0, 0, x_{(i+1)_\mathrm{P}}, \dots, 0)$ and $x_{+i_\mathrm{P}+i_\mathrm{M}} := (\dots, x_{(i-1)_\mathrm{M}}, 1, 1, x_{(i+1)_\mathrm{P}}, \dots, 0)$ be states that differ by exactly one event i. Transitions from states x to states $x_{+i_\mathrm{P}+i_\mathrm{M}}$ happen at rate

$$Q_0(\Theta)_{x_{+i_\mathrm{P}+i_\mathrm{M}},x} = \Theta_{i,i} \prod_{\substack{x_{j_\mathrm{P}}=x_{j_\mathrm{M}}=1\\ j\leq n}} \Theta_{i,j}\,. \tag{6}$$

All other transitions within $\mathcal{S}_0$ are prohibited (rate 0).

After seeding, the primary tumor and the metastasis are separate tumors and we assume that both accumulate mutations independently of each other. Formally, we describe the post-seeding dynamics by a CTMC on the subspace $\mathcal{S}_1 = \{\{0,1\}\times\{0,1\}\}^n \times \{1\} \subset \mathcal{S}$. We introduce two transition rate matrices Q_P and $Q_\mathrm{M} \in \mathbb{R}^{2^{2n+1}\times 2^{2n+1}}$. Q_P holds the rates for transitions that change only the primary tumor part of a state x: Transitions from states $x = (\dots, x_{(i-1)_\mathrm{M}}, 0, x_{i_\mathrm{M}}, x_{(i+1)_\mathrm{P}}, \dots, 1)$ to states $x_{+i_\mathrm{P}} = (\dots, x_{(i-1)_\mathrm{M}}, 1, x_{i_\mathrm{M}}, x_{(i+1)_\mathrm{P}}, \dots, 1)$ occur at rate

$$Q_\mathrm{P}(\Theta)_{x_{+i_\mathrm{P}},x} = \Theta_{i,i} \prod_{\substack{x_{j_\mathrm{P}}=1\\ j\leq n}} \Theta_{i,j}\,. \tag{7}$$

Note that transition rates in Q_P only depend on the primary tumor genotype $\mathrm{PT}(x)$ and not on the full state x. Since events must occur one at a time, all other transitions on $\mathcal{S}_1$ that affect the primary tumor occur at rate 0. Q_M holds the rates for transitions that change only the metastasis part of a state x. We assume that metastatic tumors spread to foreign sites and face novel selective pressures that can differ drastically from the original site. We account for this by explicitly modeling effects from the seeding event on the progression events. Progression events occur in the metastasis at a rate given by the product of their base rates, the effects of events that are present in the metastasis and the effect of the new environment. Hence, transitions from states $x = (\dots, x_{(i-1)_\mathrm{M}}, x_{i_\mathrm{P}}, 0, x_{(i+1)_\mathrm{P}}, \dots, 1)$ to states $x_{+i_\mathrm{M}} = (\dots, x_{(i-1)_\mathrm{M}}, x_{i_\mathrm{P}}, 1, x_{(i+1)_\mathrm{P}}, \dots, 1)$ occur at rate

$$Q_\mathrm{M}(\Theta)_{x_{+i_\mathrm{M}},x} = \Theta_{i,i} \left(\prod_{\substack{x_{j_\mathrm{M}}=1\\ j\leq n}} \Theta_{i,j} \right) \Theta_{i,n+1}\,. \tag{8}$$

All other transitions on $\mathcal{S}_1$ that affect the metastasis are prohibited (rate 0). The full transition rate matrix on $\mathcal{S}_1$ is then given by the sum of Q_P and Q_M.

By construction, the last event that occurs jointly and at the same time in a primary tumor and metastasis is the seeding event. Let $Q_\mathrm{S} \in \mathbb{R}^{2^{2n+1}\times 2^{2n+1}}$ denote the transition rate matrix that holds the rates for all transitions from states $x = (x_{1_\mathrm{M}}, \dots, x_{n_\mathrm{M}}, 0)$ in $\mathcal{S}_0$ to their corresponding states $x_{+\mathrm{S}} = (x_{1_\mathrm{M}}, \dots, x_{n_\mathrm{M}}, 1)$ in $\mathcal{S}_1$. Such transitions occur at rate

$$Q_\mathrm{S}(\Theta)x_{+\mathrm{S}}, x = \Theta_{n+1,n+1} \prod_{\substack{x_{j_\mathrm{P}}=x_{j_\mathrm{M}}=1\\ j\leq n}} \Theta_{n+1,j}\,. \tag{9}$$

See Figure 1b for an illustration of the state space for $n = 2$. The transition rate matrix on the full state space $\mathcal{S}$ is then

$$Q = Q_0 + Q_\mathrm{S} + Q_\mathrm{P} + Q_\mathrm{M} \tag{10}$$

and we denote the probability distribution over states at time t by $\mathbf{p}(t)$. Following [22] we also provide formulas for the matrices $Q_0, Q_\mathrm{S}, Q_\mathrm{P}, Q_\mathrm{M}$ as sums of tensor products in the supplement. By using these tensor structures in conjunction with the methods outlined in [32], the model parameters can be learned with a time and space complexity only exponential in the number of events that have occurred for each sample in the dataset, rather than exponential in $2(2n+1)$.

2.2.2 Modeling consecutive observations

Following [34] we model the observation of tumors explicitly as events. Since we model two tumors that at some point evolve independently and can also be observed separately, we have to include two distinct observation events. Thus we now model a CTMC on the extended state space $\mathcal{S}_D := \mathcal{S} \times \{0,1\}^2$. Let $\bar{\mathbf{p}}(t)$ denote the probability distribution over states on the extended state space at time t. We assume that each event has a multiplicative effect on the rate of observation of the tumor it occurred in. Since the events that lead to the detection of a primary tumor can be vastly different from the effects that lead to the detection of a metastasis, we introduce two separate parameter vectors $\Omega_\mathrm{P}, \Omega_\mathrm{M} \in \mathbb{R}^{n+1}_{>0}$ that contain the effects of progression events in the primary tumor and the metastasis on the rates of their respective observation event.

The primary tumor and the metastasis observation rates are defined as

$$(u_\mathrm{P})_x = \begin{cases} \prod\limits_{\substack{x_{j_\mathrm{P}}=1 \\ j\le n}} (\Omega_\mathrm{P})_j \,, & \text{if } x_{n+1} = 0\,, \\ (\Omega_\mathrm{P})_{n+1} \prod\limits_{\substack{x_{j_\mathrm{P}}=1 \\ j\le n}} (\Omega_\mathrm{P})_j \,, & \text{otherwise}\,, \end{cases} \tag{11}$$

$$(u_\mathrm{M})_x = \begin{cases} 0\,, & \text{if } x_{n+1} = 0\,, \\ (\Omega_\mathrm{M})_{n+1} \prod\limits_{\substack{x_{j_m}=1 \\ j\le n}} (\Omega_\mathrm{M})_j \,, & \text{otherwise}\,. \end{cases} \tag{12}$$

Let $U_\mathrm{P}, U_\mathrm{M} \in \mathbb{R}^{2^{2n+1}\times 2^{2n+1}}$ denote the diagonal matrices that hold the observation rates for primary tumors and metastases respectively and $U_\mathrm{S} = U_\mathrm{P} + U_\mathrm{M}$. We define that a metastasis is not observable prior to the seeding. Therefore, we set the rates of observation of metastases for such states to 0. We are interested in the distribution of the full system at the time of first observation, which can be triggered by either primary tumor or metastasis. We calculate this analogously to [34] as the stationary distribution $\bar{\mathbf{p}}$ on the extended state space $\mathcal{S}_D$ where each of the observation events halts the progression of the entire system. Each state where either observation occurred becomes an absorbing state. Thus the entire probability mass is located on the sets of states $O_\mathrm{P} = \mathcal{S} \times (1,0)$ (primary tumor is observed) and $O_\mathrm{M} = \mathcal{S} \times (0,1)$ (metastasis is observed). Analogous to Equation (5), we therefore have

$$\bar{\mathbf{p}}|_{O_\mathrm{P}} = U_\mathrm{P}(U_\mathrm{S} - Q)^{-1}\mathbf{p}_0 \text{ and} \tag{13}$$

$$\bar{\mathbf{p}}|_{O_\mathrm{M}} = U_\mathrm{M}(U_\mathrm{S} - Q)^{-1}\mathbf{p}_0\,. \tag{14}$$

In most cases, there is a considerable time lag between the observation of a primary tumor and the observation of its metastatic offspring. To account for this, we model two consecutive observations. Consider the case where the primary tumor is observed first with genotype $x^\mathrm{P} \in \{0,1\}^n$ and the metastasis is only observed at a later point in time with genotype $x^\mathrm{M} \in \{0,1\}^n$. In this case the metastasis is unobservable at the time of primary tumor observation, and thus we are interested in the metastasis marginal probability $\bar{\mathbf{p}}^\mathrm{Po}$ of only observing a primary tumor x^P, given by

$$\bar{\mathbf{p}}^\mathrm{Po}_{x^\mathrm{P}} = \sum_{\substack{x\in O_\mathrm{P} \\ \mathrm{PT}(x) = x^\mathrm{P}}} (\bar{\mathbf{p}}|_{O_\mathrm{P}})_x\,. \tag{15}$$

Note that each tumor in a dataset is observed exactly once and no information about its subsequent progression is available. Therefore we do not track the progression of the primary tumor after its observation. Instead from here on, we only model the progression of the still unobserved metastasis. To do so, we first calculate the distribution of metastasis genotypes at the time of primary tumor observation conditioned on the observed primary tumor genotype, which is given by

$$\bar{\mathbf{p}}^{\mathrm{M}|\mathrm{Po}}_x = \begin{cases} \frac{\left(\bar{\mathbf{p}}|_{O_\mathrm{P}}\right)_x}{\bar{\mathbf{p}}^\mathrm{Po}_{x^\mathrm{P}}}\,, & \text{if } \mathrm{PT}(x) = x^\mathrm{P}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{16}$$

In words, we set the probability of all states where the primary tumor genotype is not equal to the observation to 0, and then renormalize the resulting vector to obtain the desired conditional distribution. Next analogously to [34] we propagate the distribution of unobserved metastases forward in time, until the metastasis is observed. This yields

$$\bar{\mathbf{p}}^{\mathrm{Mo}|\mathrm{Po}} = U_\mathrm{M}(U_\mathrm{M} - Q_\mathrm{M})^{-1}\bar{\mathbf{p}}^{\mathrm{M}|\mathrm{Po}}\,. \tag{17}$$

Finally, the probability to observe a primary tumor and metastasis pair in state x, given that the primary tumor was observed first is

$$\bar{\mathbf{p}}^{\mathrm{Po}>\mathrm{Mo}}_x = \bar{\mathbf{p}}^{\mathrm{Mo}|\mathrm{Po}}_x \; \bar{\mathbf{p}}^\mathrm{Po}_{x^\mathrm{P}}\,. \tag{18}$$

By an analogous calculation the probability to observe a primary tumor and metastasis pair in state x, given that the metastasis was observed first is given by

$$\bar{\mathbf{p}}^{\mathrm{Mo}>\mathrm{Po}}_x = \bar{\mathbf{p}}^{\mathrm{Po}|\mathrm{Mo}}_x \; \bar{\mathbf{p}}^\mathrm{Mo}_{x^\mathrm{M}}\,. \tag{19}$$

If the order of observation is not recorded then we evaluate the total probability to observe state x as

$$\bar{\mathbf{p}}^\mathrm{tot}_x = \bar{\mathbf{p}}^{\mathrm{Po}>\mathrm{Mo}}_x + \bar{\mathbf{p}}^{\mathrm{Mo}>\mathrm{Po}}_x\,. \tag{20}$$

Equations (18), (19), (20) give the probabilities of observing pairs of genotypes. However, often only a single genotype is available, whereas the other is missing. Such individual data points are incorporated by first calculating the full joint distributions over all states and then by marginalizing over the missing genotypes. First consider the case, where only a primary tumor is observed with genotype x^{P}, then marginalization over the unobserved metastasis genotypes yields

$$\bar{\mathbf{p}}^{\mathrm{Mm}}_{x^{\mathrm{P}}} = \sum_{\substack{y \in \mathcal{S} \times (1,1) \\ \mathrm{PT}(y) = x^{\mathrm{P}}}} \bar{\mathbf{p}}^{\mathrm{tot}}_{y} \, . \tag{21}$$

If a metastasis was observed but not sequenced, then we do not need to sum over all states, but only over states in $\mathcal{S}_1$. Conversely, if evidence for the complete absence of metastases is available, then we only need to sum over states in $\mathcal{S}_0$. Next, consider the case where only a metastasis is observed with genotype x^{M}, then marginalizing over the unobserved primary tumor genotypes yields

$$\bar{\mathbf{p}}^{\mathrm{Pm}}_{x^{\mathrm{M}}} = \sum_{\substack{y \in \mathcal{S}_1 \times (1,1) \\ \mathrm{MT}(y) = x^{\mathrm{M}}}} \bar{\mathbf{p}}^{\mathrm{tot}}_{y} \, . \tag{22}$$

Since a metastasis is observed, we know that seeding must have occurred and therefore we only need to sum over states in $\mathcal{S}_1$.

2.2.3 Parameter estimation

The average log-likelihood of a dataset $\mathcal{D}$ containing primary tumor and metastasis pairs as well as single genotypes is given by

$$l_{\mathcal{D}}(\Theta, \Omega_M, \Omega_P) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \log(\mathbf{p}_d) \tag{23}$$

where

$$\mathbf{p}_d = \begin{cases} \bar{\mathbf{p}}^{\mathrm{Mm}}_d \, , & \text{if } d \text{ is a single primary tumor} \, , \\ \bar{\mathbf{p}}^{\mathrm{Pm}}_d \, , & \text{if } d \text{ is a single metastasis} \, , \\ \bar{\mathbf{p}}^{\mathrm{Po>Mo}}_d \, , & \text{if } d \text{ is paired, primary obs. first} \, , \\ \bar{\mathbf{p}}^{\mathrm{Mo>Po}}_d \, , & \text{if } d \text{ is paired, metastasis obs. first} \, , \\ \bar{\mathbf{p}}^{\mathrm{tot}}_d \, , & \text{if } d \text{ is paired, obs. order unknown} \, . \end{cases} \tag{24}$$

We infer the parameters $\Theta, \Omega_{\mathrm{M}}, \Omega_{\mathrm{P}}$ from data via maximum likelihood estimation. We follow [34] and utilize the penalization

$$\begin{aligned} \mathrm{penal}(\Theta, \Omega_{\mathrm{M}}, \Omega_{\mathrm{P}}) = & \sum_{i \neq j}^{n+1} \sqrt{\theta_{i,j}^2 + \theta_{j,i}^2 - \theta_{i,j}\theta_{j,i}} \\ & + \sum_{j=1}^{n+1} \left(|(\omega_{\mathrm{P}})_j| + |(\omega_{\mathrm{M}})_j| \right) \end{aligned} \tag{25}$$

with $\theta_{i,j} = \log(\Theta_{i,j})$, $(\omega_{\mathrm{M}})_j = \log((\Omega_{\mathrm{M}})_j)$, $(\omega_{\mathrm{P}})_j = \log((\Omega_{\mathrm{P}})_j)$. The penalty promotes sparsity as the logarithmic parameters are shrunk to 0. Additionally, it promotes symmetry as effects between events i and j are grouped and selected together. We then optimize

$$l_{\mathcal{D}}(\Theta, \Omega_M, \Omega_P) - \lambda \, \mathrm{penal}(\Theta, \Omega_M, \Omega_P) \tag{26}$$

via gradient ascent. The hyper parameter λ is selected via 5-fold cross validation.

3 Results

To further our understanding of metastatic spread in lung adenocarcinomas, we trained metMHN on 4,852 paired and unpaired samples from the LUAD cohort of the MSK-IMPACT study. Next, we describe the dataset and then present our key findings.

3.1 Data preparation

We retrieved the AACR GENIE 14.1 data release [30] through synapse.org [12]. Our selection included all samples assayed at the Memorial Sloan-Kettering Cancer Center annotated with the ONCOTREE code 'LUAD' (Lung Adenocarcinoma). For primary tumors without corresponding metastasis samples, we retrieved information about their metastatic status from [26] and excluded samples where the status of the metastasis was unknown. The final dataset consisted of 453 matched primary tumor (PT)/metastasis (MT) samples, 2,127 unpaired MT samples, 595 PT samples without corresponding metastases (seeding=0), and 1,677 PT samples with metastases that were not sequenced (seeding=1). The three most highly mutated paired samples were excluded from our analysis due to computational challenges in processing them with metMHN. In total, our study included 2,725 PT and 2,580 MT samples from 4,852 patients. Metadata for each sample also included the age of the corresponding patient at which the sample was reported. This data informs the model about the order of observation of primary tumors and metastases in the same patients. When multiple PT or MT samples were present, we chose the PT sample with the youngest sampling age and the MT sample with the oldest sampling age.

Genomic data consisted of somatic mutation data and segmented log R ratio (LRR) copy number data derived from single-region bulk sequencing using the targeted MSK-IMPACT panel [11]. We annotated mutation data using OncoKB [8] and filtered for variants likely to be functional, as outlined in [34]. Our analysis was restricted to genes consistently included in all versions of the MSK-IMPACT panel [12]. Specifically, we examined mutations in the top 20 most frequently mutated genes. In the case of copy number alterations, we initially normalized segmented copy number data using mecan4CNA [15]. Amplifications were identified with LRR values corresponding to relative copy number gains ≥ 0.5. Conversely, deletions were marked by LRR values corresponding to relative copy number losses ≤ -0.5. We determined the precise minimal intervals necessary for a copy number event classification in 8 instances, based on the minimal commonly altered regions per chromosome arm and gene extents. For amplifications, we required full gene extents to be covered by an alteration, whereas for deletions we allowed for shorter intervals. In total, our study considered 28 distinct genomic events, including mutational events ('M'), copy number amplification ('Amp') and deletion ('Del') events. Binary event input data, alongside exact interval definitions for copy number events, records of the selected patients and samples and preparation scripts are accessible at https://github.com/cbg-ethz/metMHN.

3.2 Effects between genomic events and seeding

On the dataset described above, we fit metMHN and tuned the hyperparameter λ in a 5-fold cross-validation (Figure 2). Reassuringly, the LUAD model confirms several interactions well-documented in the literature. Specifically, it identifies the strong, antagonistic relationship (evidenced by a bidirectional negative edge) between the principal drivers KRAS (M) and EGFR (M) [37, 35]. Our model infers that EGFR suppresses further mutational co-drivers, which suggests that it might often be sufficient for progression. Instead, EGFR-driven LUADs frequently exhibit disruption of cell cycle regulation through copy number losses in RB1 and CDKN2A, two patterns also described in [25].

The model further highlights synergistic interactions that reflect established oncogenic processes, such as the rate increases observed between STK11 (M) and KEAP1 (M), and between TP53 (M) and RB1 (M) [41, 6, 27]. metMHN also infers multiple positive interactions between gene mutations and corresponding copy number alterations, exemplified by the interaction between EGFR (M) and amplification of EGFR/7p, as well as between STK11 (M) and deletion of STK11/19p — a pattern commonly seen across various cancers [1]. Additionally, the model reflects that several mutational events capable of activating the (RTK)-RAS-RAF-MEK signaling pathway—namely, KRAS (M), EGFR (M), NF1 (M), BRAF (M), and MET (M)—tend to promote the observation of primary tumors and suppress each other's occurrence [20].

3.3 metMHN identifies drivers of metastasis

We next examined the interactions between genomic events and metastatic seeding. The outgoing edges from the seeding event (rightmost column in Figure 2) represent the cancer cell's adaptive response to the changing selective pressures encountered during its journey from the primary tumor to the metastatic site. The incoming edges into the seeding event (bottom row in Figure 2) indicate how particular mutations within the primary tumor may accelerate or impede the metastatic seeding rate, thereby pinpointing genetic elements that either drive or hinder metastasis development.

metMHN identifies mutations and amplifications in EGFR, along with TP53 mutations and deletions, and MET mutations, as accelerators of metastasis formation, as indicated by positive edges (i.e., promoting effects) from these events to the seeding event (Figure 2). These findings are substantiated by experimental evidence which indicate that activation of EGFR [36, 10], inactivation of TP53 [38, 29], and activation of MET [42, 9] enhance the metastatic capacity of lung cancer cells. Beyond these events, metMHN also revealed that various other copy number alterations positively influence the seeding process. Although widespread aneuploidy is typically regarded as a hallmark of advanced cancer stages [3], specific copy number changes, like CDKN2A deletions, have been documented to sometimes occur early in

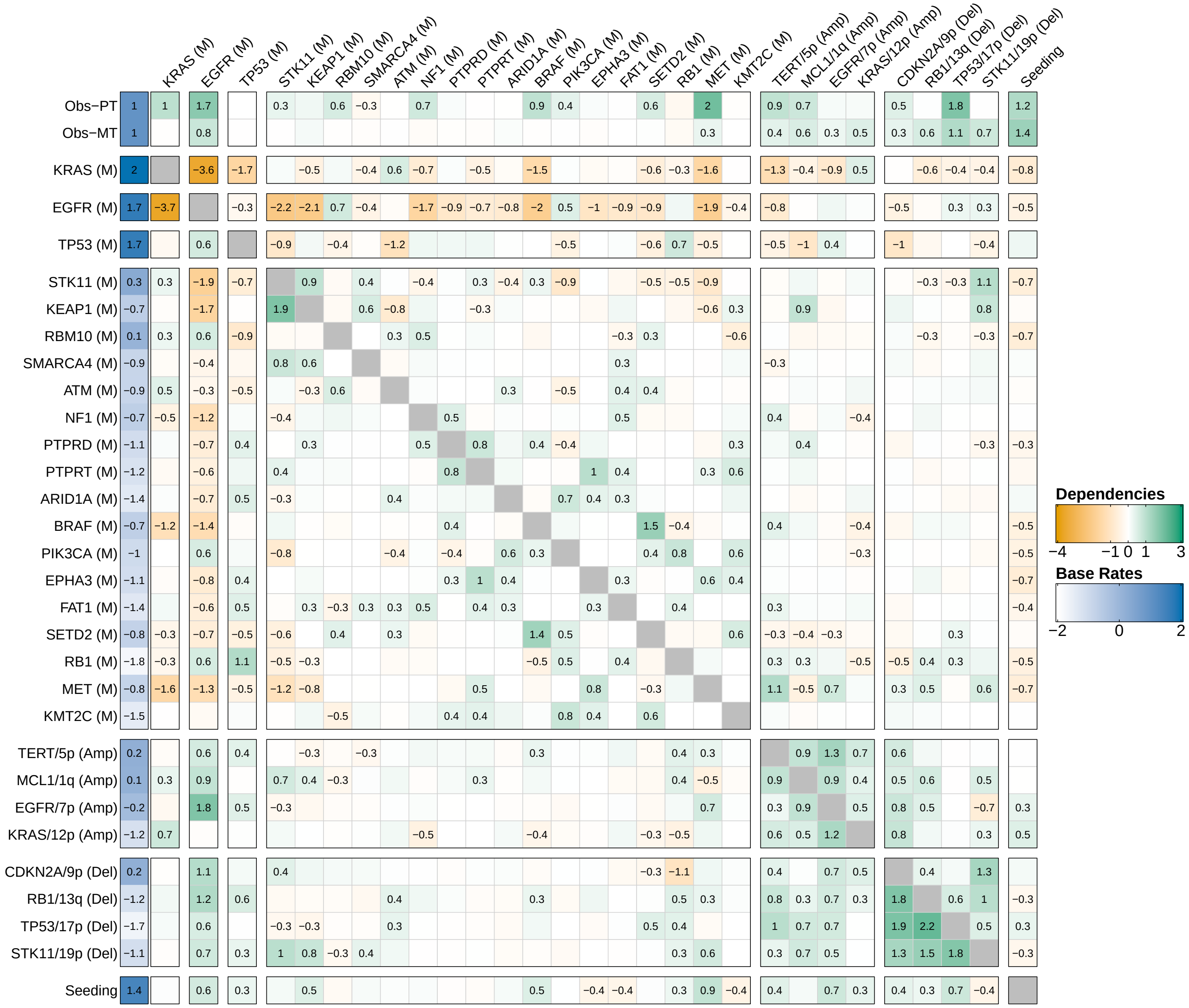


Figure 2: Interactions between progression events in lung adenocarcinomas. The log-effects on observation (clinical detection) of the primary tumor and metastasis ω_P and ω_M are plotted in the first two rows, the remaining matrix shows the log-interaction strengths among genomic events θ. The base rates of all events are plotted on the left (in blue). The effects an event i exerts on other events j are collected in the i^th column (outgoing edges). Vice versa the effects, that events j exert on event i are collected in the i^th row (incoming edges). Effects of genomic events on seeding are shown in the bottom row. Vice versa effects from seeding on genomic events are shown in the rightmost column.

lung adenocarcinoma development [25, 39]. In this context we also note metMHN's inference that copy number events generally do not substantially affect the primary tumor observation rate but indeed promote metastasis observability.

Interestingly, the effects promoting metastasis were relatively modest when compared to the base rate of seeding. This observation suggests that certain genetic or non-genetic drivers of the metastatic process might not be accounted for in the model. Alternatively, this could also indicate that primary tumor cells may inherently possess a propensity to metastasize, as suggested by [21]. Lastly, metMHN suggests that upon the seeding of metastases, the accumulation rates of many mutational events tend to decrease. This pattern could imply that once the metastatic process is initiated and in progress, there is diminished pressure for further mutational driver alterations, compared to the initial stages of primary tumorigenesis [13].

3.4 Relative timing of progression events and seeding

We computed the most likely chronological sequences of events for 313 paired data points and 2,127 unpaired metastases, where we limited our analysis to cases where calculations were feasible. For the paired data points the orderings are branched, as exemplified in Figure 3a. Prior to seeding, events happen jointly in the primary tumor. Upon seeding, the trajectory splits into a primary tumor branch and a metastasis branch (blue lower and red upper branches in Figure 3a, respectively). The unpaired metastases' orderings are linear.

Next, we analyzed the distribution of event positions, relative to trajectory lengths (Figure 3b): The plots show for every event how often it occurred for each relative time point, where the left end of the axes corresponds to the beginning and the right to the end of progression. Well-established and highly frequent mutational drivers of LUAD progression, such as KRAS (M), EGFR (M) and TP53 (M) appear consistently early as initiating events. We find similar patterns for less frequent mutational events, such as MET (M) and SETD2 (M). Some events rarely appear as initiators, but still mostly occur in the early half of any sequence, such as STK11 (M) and BRAF (M). For example, RB1 (M) rarely happens spontaneously, which is reflected by its low base rate. However, it is promoted by both EGFR (M) and TP53 (M) and thus tends to happen subsequently, see Figure 2 and Figure 3d. Crucially, metastatic seeding was observed to happen at varying stages, with the majority of trajectories showing genomic progression both before and after seeding. On the late end of the spectrum we mainly find copy number events. After the first such event happens, it usually promotes other copy number events (see Figure 2), leading to compounding rate increases for copy number events towards the end of a typical trajectory, possibly reflecting genomic instability in late stage cancers [3].

Next, we stratified the inferred metastasis trajectories by the 3 most prevalent initial events. Specifically, trajectories starting with TP53 (M), KRAS (M) and EGFR (M) at the first position accounted for 1,766 patients or 72.38% of the analyzed metastases (Figure 3d). Remarkably, the subset of trajectories initiated by TP53 (M) (left side) included a significant number of tumors which seeded immediately after. These tumors then predominantly acquired copy number events. In a minority of cases, additional mutation events such as STK11 (M) and KEAP1 (M) occurred before seeding. Trajectories that began with KRAS (M) (center) generally showed later seeding, frequently after the accumulation of other mutational co-drivers, including TP53 (M), STK11 (M), KEAP1 (M), RBM10 (M), and ATM (M). These trajectories too typically concluded with a series of copy number events. Conversely, trajectories initiated by EGFR (M) (right side) exhibited distinctly different progression patterns. Contrary to those beginning with KRAS (M), these trajectories rarely accumulated additional mutational events before seeding, with TP53 (M) being an exception. Post-seeding, the progression was once again dominated by copy number changes. However, these events followed characteristic sequences, often starting with EGFR/7p (Amp) and CDKN2A/9p (Del), then proceeding to TP53/17p (Del) and STK11/19p (Del), and culminating with the clinical detection of the tumor.

3.5 metMHN is consistent with clonality information

A key quality of metMHN is its ability to quantify the timing of seeding relative to other progression events. To validate this, we compared it with an orthogonal readout of metastatic development relative to mutational events: A mutation that predates the seeding of a primary tumor clone is expected to be clonal, i.e., exhibit a high variant allele frequency (VAF, close to 0.5) in subsequent metastases [4]. In contrast, mutations arising post-seeding in metastases are more likely to be subclonal and thus exhibit lower VAFs. Therefore, we used per-gene mean VAFs in metastasis samples as a proxy for the relative timing (pre- or post-seeding) of the occurrence of mutations in the respective gene. To account for a bias in VAF distributions, we restricted VAF measurements to cases in which the respective gene was not copy number altered. We compared for each mutation its mean VAF with the model-derived probability that the event occurred prior to seeding. To this end, we approximated this probability through simulations using Gillespie's algorithm [17]. We found that mutational events with high pre-seeding probabilities in metastases corresponded to elevated VAFs in metastasis samples as evidenced by a Pearson correlation coefficient of 0.55 ($p = 0.01$) see Figure 3c and Figure 1 in the supplement. In summary, while metMHN builds on co-occurrence patters and does not leverage VAF information, they nevertheless produce results consistent with clonality information.

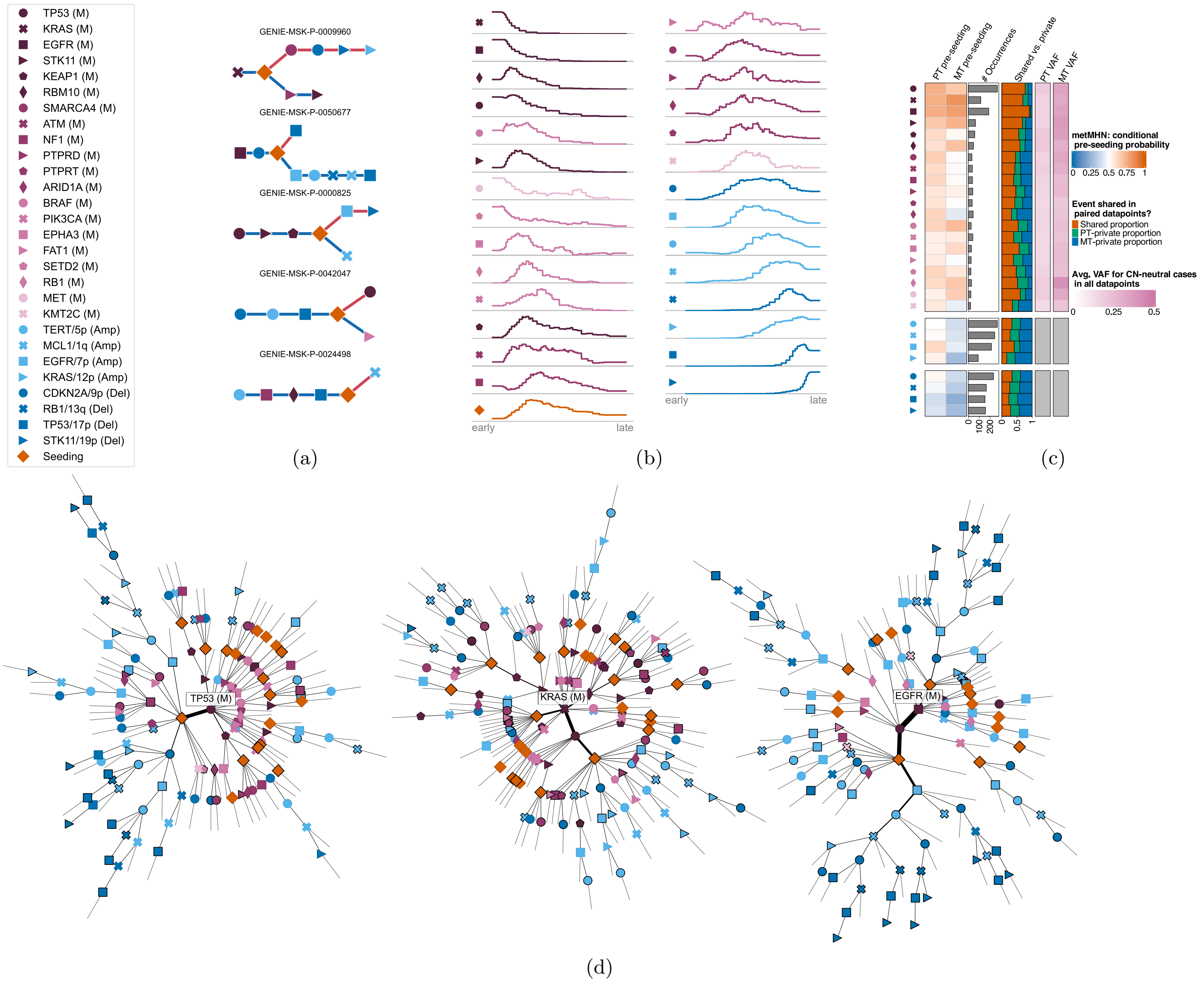


Figure 3: (a) Event orders for 5 patients as inferred by metMHN. Events accumulate from left to right. Blue edges represent the primary tumor development, red edges the one of the metastasis. Distances between events do not correspond to real or estimated time. (b) Distribution of relative positions in trajectories. The left end of the axes corresponds to the beginning, and the right to the end of progression. (c) Pre-seeding probabilities estimated by metMHN and empirical evidence from paired samples. The first and second column show the pre-seeding probabilities estimated by metMHN conditioned on the event being observed in the primary tumor (column 1) or the metastasis (column 2). Column 3 shows the number of occurrences for each event in the paired data, column 4 shows the proportions of shared versus private occurrences for each event in the paired data. Columns 5 and 6 show the mean variant allele frequencies in the primary tumor and the metastasis respectively. (d) Most probable event orderings for observed metastases genotypes as inferred by metMHN, stratified by TP53 (M) (left), KRAS (M) (middle) and EGFR (M) (right) as their first event. Each branch extending out from a tree's root represents a group of metastases for which the events were inferred to occur in the order of the branch. Edge widths scale proportionally to the dataset's count of metastases commencing with that particular sequence of events, and branches are trimmed at an edge threshold of 3. Black-bordered nodes indicate observed genotypes.

4 Discussion

We have presented metMHN, an efficient analytical model for cancer progression, specifically designed to investigate the forking progression paths of primary tumors and their metastatic offspring. Distinguishing itself from specialized phylogenetic methods operating on rare multi-region sequencing data, metMHN capitalizes on the extensive cross-sectional data available from clinical targeted sequencing and is able to infer relationships between events that are shared across individual samples. Our comprehensive analysis, encompassing data from nearly 5000 lung cancer patients, corroborates well-established relationships among key genomic drivers. In addition, metMHN successfully identifies specific events in primary tumors that may accelerate the development of metastases and quantifies how the dynamics of event accumulation change upon metastatic branching. Moreover, metMHN allows for the reconstruction as well as for the simulation of disease histories yielding further insight into the dynamics of metastatic cancers. This dual capability of metMHN not only deepens our comprehension of the key events that propel cancer progression but also provides a quantitative perspective on how these interactions manifest into distinct histories of tumor progression.

Every model's efficacy is inherently tied to the quality of its training data. While metMHN uses comprehensive cross-sectional data from bulk tissue, this approach has its limitations, particularly in resolving the clonal structures of heterogeneous tumors. In metMHN, binary states represent the tumor as a whole. Consequently, two tumors with identical mutations will be interpreted identically by the model, even if, in one case, the mutations exist within the same clone, and in the other, they are in separate clones. Another challenge arises when the training data does not accurately represent the patient population. For instance, an under-representation of metastatic tumors in the training data could lead to an underestimation of the base rate for the seeding event, falsely suggesting they occur later in the progression than they actually do, while an over-representation of these cases would have the opposite effect. In contrast, phylogenetic methods, which reconstruct tumor evolution on an individual basis, are less susceptible to biases in datasets. These methods also offer the advantage of resolving clonal structures, presenting a more detailed picture of tumor evolution. However, the scarcity of data, especially in multi-region sequencing studies, limits their ability to represent patient populations comprehensively.

In summary, metMHN complements phylogenetic analyses and stands out as the only cancer progression models capable of fully utilizing the largest clinical genomic datasets currently available. metMHN models offer a distinct advantage: They provide a quantitative and dynamic description of metastatic cancer progression. This unique approach enables them to contribute valuable insights into the complexities of metastatic spread, enriching our understanding of cancer progression with their analytical perspective.

5 Competing interests

No competing interest is declared.

6 Author contributions statement

KR, RSC, NB and RSP conceptualized and initiated the project. KR, RSC and YLH developed the model. KR, YLH, CN implemented the algorithms. SP, MK, TW and LG provided numerical foundations for model analysis. AL prepared the input data. AL, YLH, KR analysed the LUAD data. KR, AL, YLH, and RSP drafted the manuscript. All authors critically read and improved upon the draft.

7 Acknowledgments

We thank Nikolaus Schultz from MSKCC, Thomas Ratzke, Stefan Vocht and Stefan Hansch for helpful discussions. This work is supported by the Deutsche Forschungsgemeinschaft (DFG: TRR 305-A02), GR-3179/6-1 “Tensorapproximationsmethoden zur Modellierung von Tumorprogression” Swiss National Science Foundation grant 179518 and Swiss Cancer League grant KFS-2977-08-2012.

References

[1] Tommaso Becchi et al. “A pan-cancer landscape of pathogenic somatic copy number variations”. In: *Journal of Biomedical Informatics* 147 (Nov. 2023), p. 104529. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2023.104529.

[2] Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. “Conjunctive Bayesian networks”. In: *Bernoulli* 13.4 (2007), pp. 893–909. DOI: 10.3150/07-BEJ6133.

[3] Uri Ben-David and Angelika Amon. "Context is everything: aneuploidy in cancer". In: *Nature Reviews Genetics* 21.1 (Sept. 2019), pp. 44–62. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0171-x.

[4] Nicolai J. Birkbak and Nicholas McGranahan. "Cancer Genome Evolutionary Trajectories in Metastasis". In: *Cancer Cell* 37.1 (Jan. 2020), pp. 8–19. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2019.12.004.

[5] Paul E. Buis and Wayne R. Dyksen. "Efficient Vector and Parallel Manipulation of Tensor Products". In: *ACM Trans. Math. Softw.* 22.1 (1996). ISSN: 0098-3500. DOI: 10.1145/225545.225548.

[6] Ling Cai et al. "A Pan-Cancer Assessment of RB1/TP53 Co-Mutations". In: *Cancers* 14.17 (Aug. 2022), p. 4199. ISSN: 2072-6694. DOI: 10.3390/cancers14174199.

[7] Giulio Caravagna et al. "Detecting repeated cancer evolution from multi-region tumor sequencing data". In: *Nature Methods* 15.9 (Sept. 2018), pp. 707–714. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0108-x.

[8] Debyani Chakravarty et al. "OncoKB: A Precision Oncology Knowledge Base". In: *JCO Precision Oncology* 1 (Nov. 2017), pp. 1–16. ISSN: 2473-4284. DOI: 10.1200/po.17.00011.

[9] Chih-Cheng Chang et al. "Regulation of metastatic ability and drug resistance in pulmonary adenocarcinoma by matrix rigidity via activating c-Met and EGFR". In: *Biomaterials* 60 (Aug. 2015), pp. 141–150. ISSN: 0142-9612. DOI: 10.1016/j.biomaterials.2015.04.058.

[10] Ting-Fang Che et al. "Mitochondrial translocation of EGFR regulates mitochondria dynamics and promotes metastasis in NSCLC". In: *Oncotarget* 6.35 (Oct. 2015), pp. 37349–37366. ISSN: 1949-2553. DOI: 10.18632/oncotarget.5736.

[11] Donavan T. Cheng et al. "Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT)". In: *The Journal of Molecular Diagnostics* 17.3 (May 2015), pp. 251–264. ISSN: 1525-1578. DOI: 10.1016/j.jmoldx.2014.12.006.

[12] The AACR Project GENIE Consortium. *Release 14.1-public*. 2023. DOI: 10.7303/SYN52918985.

[13] Dagmara Dymerska and Anna A. Marusiak. "Drivers of cancer metastasis – Arise early and remain present". In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1879.2 (Mar. 2024), p. 189060. ISSN: 0304-419X. DOI: 10.1016/j.bbcan.2023.189060.

[14] Eric R. Fearon and Bert Vogelstein. "A genetic model for colorectal tumorigenesis". In: *Cell* 61.5 (1990), pp. 759–767. ISSN: 0092-8674. DOI: https://doi.org/10.1016/0092-8674(90)90186-I.

[15] Bo Gao and Michael Baudis. "Minimum error calibration and normalization for genomic copy number analysis". In: *Genomics* 112.5 (Sept. 2020), pp. 3331–3341. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2020.05.008.

[16] Peter Georg. "Tensor Train Decomposition for solving high-dimensional Mutual Hazard Networks". In: (2022). DOI: 10.5283/EPUB.53004.

[17] Daniel T. Gillespie. "Exact stochastic simulation of coupled chemical reactions". In: *The Journal of Physical Chemistry* 81.25 (Dec. 1977), pp. 2340–2361. ISSN: 0022-3654. DOI: 10.1021/j100540a008.

[18] Sam F. Greenbury, Mauricio Barahona, and Iain G. Johnston. "HyperTraPS: Inferring Probabilistic Patterns of Trait Acquisition in Evolutionary and Disease Progression Pathways". In: *Cell Systems* 10.1 (2020), 39–51.e10. ISSN: 2405-4712. DOI: https://doi.org/10.1016/j.cels.2019.10.009.

[19] Marcus Hjelm, Mattias Höglund, and Jens Lagergren. "New Probabilistic Network Models and Algorithms for Oncogenesis". In: *Journal of Computational Biology* 13.4 (2006). PMID: 16761915, pp. 853–865. DOI: 10.1089/cmb.2006.13.853. eprint: https://doi.org/10.1089/cmb.2006.13.853.

[20] Robin Imperial et al. "Comprehensive pancancer genomic analysis reveals (RTK)-RAS-RAF-MEK as a key dysregulated pathway in cancer: Its clinical implications". In: *Seminars in Cancer Biology* 54 (Feb. 2019), pp. 14–28. ISSN: 1044-579X. DOI: 10.1016/j.semcancer.2017.11.016.

[21] Christoph A. Klein. "Cancer progression and the invisible phase of metastatic colonization". In: *Nature Reviews Cancer* 20.11 (Oct. 2020), pp. 681–694. ISSN: 1474-1768. DOI: 10.1038/s41568-020-00300-6.

[22] Maren Klever et al. "Low-rank tensor methods for Markov chains with applications to tumor progression models". In: *Journal of Mathematical Biology* 86.1 (Dec. 2022), p. 7. ISSN: 1432-1416. DOI: 10.1007/s00285-022-01846-9.

[23] Arthur W. Lambert, Diwakar R. Pattabiraman, and Robert A. Weinberg. "Emerging Biological Principles of Metastasis". In: *Cell* 168.4 (Feb. 2017), pp. 670–691. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.11.037.

[24] Xiang Ge Luo, Jack Kuipers, and Niko Beerenwinkel. "Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees". In: *Nature Communications* 14.1 (June 2023), p. 3676. ISSN: 2041-1723. DOI: 10.1038/s41467-023-39400-w.

[25] Rahul Nahar et al. "Elucidating the genomic architecture of Asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing". In: *Nature Communications* 9.1 (Jan. 2018). ISSN: 2041-1723. DOI: 10.1038/s41467-017-02584-z.

[26] Bastien Nguyen et al. "Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients". In: *Cell* 185.3 (Feb. 2022), 563–575.e11. ISSN: 0092-8674. DOI: 10.1016/j.cell.2022.01.003.

[27] Michael Offin et al. "Concurrent RB1 and TP53 Alterations Define a Subset of EGFR-Mutant Lung Cancers at risk for Histologic Transformation and Inferior Clinical Outcomes". In: *Journal of Thoracic Oncology* 14.10 (Oct. 2019), pp. 1784–1793. ISSN: 1556-0864. DOI: 10.1016/j.jtho.2019.06.002.

[28] Simon Pfahler et al. *Taming numerical imprecision by adapting the KL divergence to negative probabilities.* 2023. arXiv: 2312.13021 [stat.CO].

[29] Emily Powell, David Piwnica-Worms, and Helen Piwnica-Worms. "Contribution of p53 to Metastasis". In: *Cancer Discovery* 4.4 (Apr. 2014), pp. 405–414. ISSN: 2159-8290. DOI: 10.1158/2159-8290.cd-13-0136.

[30] Trevor J. Pugh et al. "AACR Project GENIE: 100, 000 Cases and Beyond". In: *Cancer Discovery* 12.9 (July 2022), pp. 2044–2057. ISSN: 2159-8290. DOI: 10.1158/2159-8290.cd-21-1547.

[31] Daniele Ramazzotti et al. "CAPRI: efficient inference of cancer progression models from cross-sectional data". In: *Bioinformatics* 31.18 (May 2015), pp. 3016–3026. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv296. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/18/3016/49034748/bioinformatics_31_18_3016.pdf.

[32] Rudolf Schill. *Mutual Hazard Networks: Markov chain models of cancer progression.* Dec. 2022. DOI: 10.5283/epub.53417.

[33] Rudolf Schill et al. "Modelling cancer progression using Mutual Hazard Networks". In: *Bioinformatics* 36.1 (June 2020), pp. 241–249. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz513. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/1/241/48981322/bioinformatics_36_1_241.pdf.

[34] Rudolf Schill et al. "Overcoming Observation Bias for Cancer Progression Modeling". In: *bioRxiv* (2023). DOI: 10.1101/2023.12.03.569824. eprint: https://www.biorxiv.org/content/early/2023/12/05/2023.12.03.569824.full.pdf.

[35] Ferdinandos Skoulidis and John V. Heymach. "Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy". In: *Nature Reviews Cancer* 19.9 (Aug. 2019), pp. 495–509. ISSN: 1474-1768. DOI: 10.1038/s41568-019-0179-8.

[36] Meng-Feng Tsai et al. "EGFR-L858R mutant enhances lung adenocarcinoma cell invasive ability and promotes malignant pleural effusion formation through activation of the CXCL12-CXCR4 pathway". In: *Scientific Reports* 5.1 (Sept. 2015). ISSN: 2045-2322. DOI: 10.1038/srep13574.

[37] Arun M Unni et al. "Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma". In: *eLife* 4 (June 2015). ISSN: 2050-084X. DOI: 10.7554/elife.06907.

[38] Shu-Ping Wang et al. "p53 controls cancer cell invasion by inducing the MDM2-mediated degradation of Slug". In: *Nature Cell Biology* 11.6 (May 2009), pp. 694–704. ISSN: 1476-4679. DOI: 10.1038/ncb1875.

[39] Thomas B. K. Watkins et al. "Pervasive chromosomal instability and karyotype order in tumour evolution". In: *Nature* 587.7832 (Sept. 2020), pp. 126–132. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2698-6.

[40] Robert A. Weinberg. *The biology of cancer.* eng. Second edition. New York: Garland Science, Taylor & Francis Group New York, 2014. ISBN: 0815342195; 9780815342199; 0815342209; 9780815342205. DOI: 10.1201/9780429258794.

[41] Corrin A. Wohlhieter et al. "Concurrent Mutations in STK11 and KEAP1 Promote Ferroptosis Protection and SCD1 Dependence in Lung Cancer". In: *Cell Reports* 33.9 (Dec. 2020), p. 108444. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2020.108444.

[42] Jun Yin et al. "HGF/MET Regulated Epithelial-Mesenchymal Transitions And Metastasis By FOSL2 In Non-Small Cell Lung Cancer". In: *OncoTargets and Therapy* Volume 12 (Nov. 2019), pp. 9227–9237. ISSN: 1178-6930. DOI: 10.2147/ott.s217595.