Bioinformatics, 2024, **40**, i140–i150
https://doi.org/10.1093/bioinformatics/btae250
ISMB 2024

Modeling metastatic progression from cross-sectional cancer genomics data

Kevin Rupp[1,2,3], Andreas Lösch[1], Yanren 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, Regensburg 93053, Germany
[2]Department of Biosystems Science and Engineering, ETH Zurich, Basel 4056, Switzerland
[3]SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland
[4]Institute for Geometry and Applied Mathematics, RWTH Aachen, Aachen 52062, Germany
[5]Faculty of Physics, University of Regensburg, Regensburg 93053, Germany

*Corresponding authors. Department of Biosystems Science and Engineering, ETH Zurich, Schanzenstrasse 44, Basel 4056, Switzerland. E-mail: niko.beerenwinkel@bsse.ethz.ch (N.B.) and Faculty of Informatics and Data Science—Statistical Bioinformatics Group, University of Regensburg, Am Biopark 9 Regensburg 93053, Germany. E-mail: rainer.spang@ur.de (R.S.)

Abstract

Motivation: 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.

Results: 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.

Availability and implementation: All datasets and code are available on GitHub at https://github.com/cbg-ethz/metMHN.

Keywords: cancer progression models, Mutual Hazard Networks, Markov chains, metastasis, cancer genomics, lung cancer.

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 (Lambert *et al.* 2017). This last step, the establishment of a (detectable) metastasis at a distant site, is what is commonly referred to as metastatic seeding. 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 (Weinberg 2014). These alterations often result in divergent genotypes between a primary tumor and its associated metastasis. Extensive clinical sequencing efforts like the MSK-MET study (Nguyen *et al.* 2022) 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 (1990). Since then, a variety of models and algorithms have emerged to refine and expand upon this concept. They include Conjunctive Bayesian Networks (Beerenwinkel *et al.* 2007), CAPRI (Ramazzotti *et al.* 2015), Network Aberration Models (Hjelm *et al.* 2006), HyperTraPS (Greenbury *et al.* 2020), and Mutual Hazard Networks (Schill *et al.* 2020). All of these models only consider the progression of a single sequence and thus cannot 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 (Caravagna *et al.* 2018) or TreeMHN (Luo *et al.* 2023) can account for this branching behavior 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 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 (LUAD) dataset provided by the Memorial Sloan-Kettering Cancer Center through AACR GENIE (Pugh *et al.* 2022).

Materials and methods

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

Mutual hazard networks

MHNs (Schill *et al.* 2020) model the progression of primary tumors as a continuous-time Markov chain (CTMC) $\{X(t), t \geq 0\}$ on the binary state space $\{0,1\}^n$. Specifically, a state $x \in \{0,1\}^n$ represents a cancer genome, where $x_i = 1$ indicates that event $i \in \{1, \dots, n\}$ (e.g. a somatic driver mutation or copy number alteration) was detected in the cancer genome, whereas $x_i = 0$ indicates that it was not detected. metMHN thus models the progression of consensus mutational profiles, without accounting for subclonal structure. 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

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = Q\mathbf{p}(t) \text{ solved by } \mathbf{p}(t) = \exp(tQ)\mathbf{p}(0). \tag{1}$$

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 $x = (\dots, x_{i-1}, 0, x_{i+1}, \dots)$ to $x_{+i} = (\dots, x_{i-1}, 1, x_{i+1}, \dots)$ 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. No assumption is made about the biological mechanisms underlying such rate changes. However, within the context of this specific analysis, rate changes between mutational events may represent evolutionary dependencies (Mina *et al.* 2022) and positive rate changes between copy number events progressively increasing levels of chromosomal instability (Potapova *et al.* 2013). The age of a tumor at the time of its diagnosis is unknown. In Schill *et al.* (2020), it is assumed to be exponentially distributed with mean 1 and independent of the state of the tumor. Marginalizing over t in the solution of 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

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

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 $2n$ (Buis and Dyksen 1996, Schill 2022). Recently (Klever *et al.* 2022, Georg 2022, Pfahler *et al.* 2023) 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 Schill *et al.* (2023), 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) (Schill *et al.* 2023).

metMHN

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 Fig. 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 $i \in \{1, \dots, n\}$ 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 $\{1, \dots, n\}$ 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

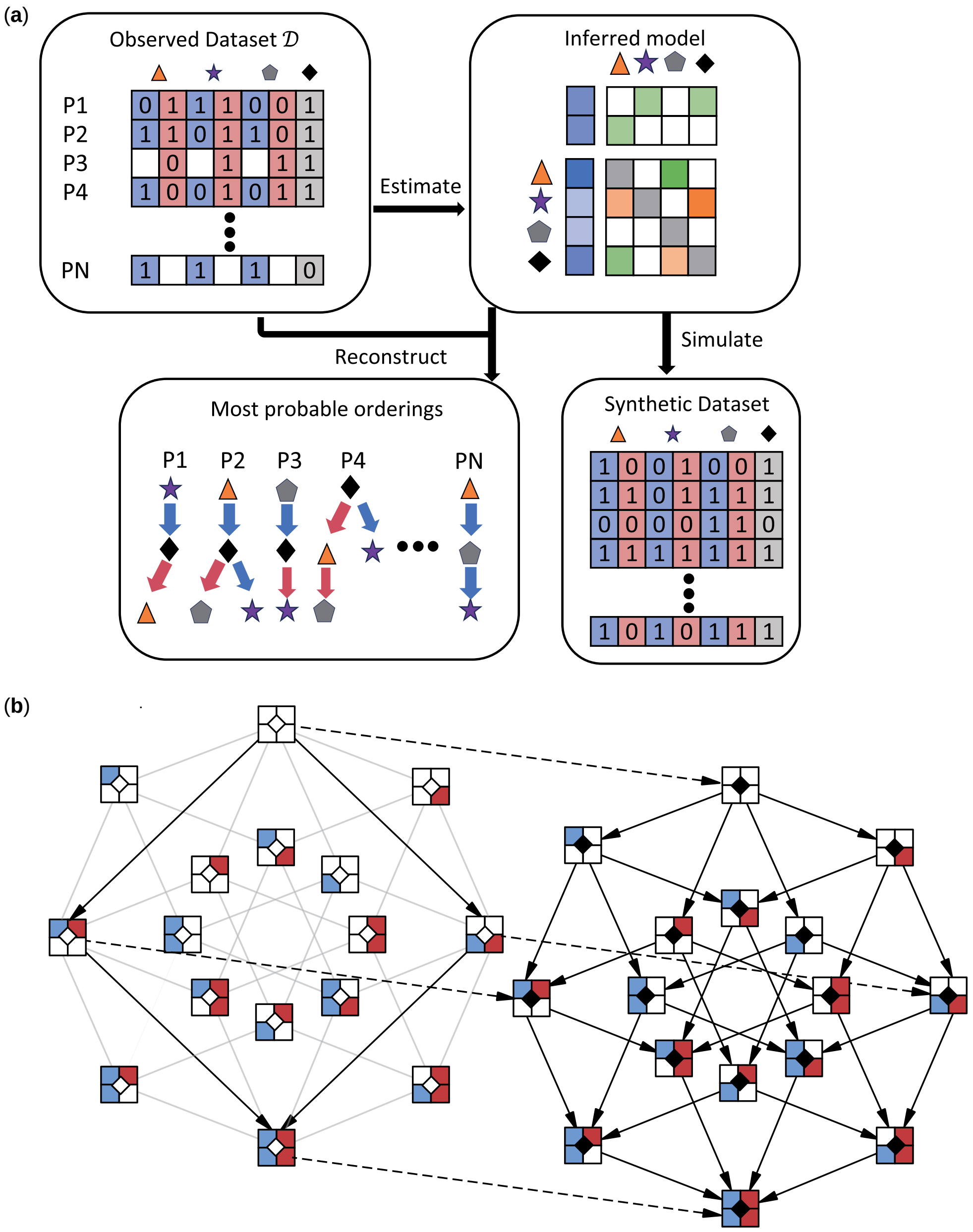

Figure 1. (a) Workflow of metMHN. In the top-left section, we show the types of input data that metMHN processes. Each row corresponds to a patient, each column to an event in the primary tumor (blue) or the metastasis (red). Events are represented by symbols and their status is encoded with a ‘1’ for 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.

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 $x = (\ldots, x_{(i-1)_\mathrm{M}}, 0, 0, x_{(i+1)_\mathrm{P}}, \ldots, 0)$ and $x_{+i_\mathrm{P}+i_\mathrm{M}} := (\ldots, x_{(i-1)_\mathrm{M}}, 1, 1, x_{(i+1)_\mathrm{P}}, \ldots, 0)$ 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

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

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

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

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

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

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

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

See Fig. 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 Klever *et al.* (2022), we also provide formulas for the matrices $Q_0, Q_\mathrm{S}, Q_\mathrm{P}, Q_\mathrm{M}$ as sums of tensor products in Supplementary Section S1. By using these tensor structures in conjunction with the methods outlined in Schill (2022, Appendix A), 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 $2(2n+1)$.

Modeling consecutive observations

Following Schill *et al.* (2023), 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 Schill *et al.* (2023) 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 $\mathrm{O}_\mathrm{P} = \mathcal{S} \times (1,0)$ (primary tumor is observed) and $\mathrm{O}_\mathrm{M} = \mathcal{S} \times (0,1)$ (metastasis is observed). Analogous to Equation (5), we therefore have

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

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

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

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

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 Schill *et al.* (2023) we propagate the distribution of unobserved metastases forward in time, until the metastasis is observed. This yields

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

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

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

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

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

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

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

Equations (18–20) 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 $\mathcal{S}_1$.

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

$$l_{\mathcal{D}}(\Theta, \Omega_{\mathrm{M}}, \Omega_{\mathrm{P}}) = \frac{1}{|\mathcal{D}|}\sum_{\mathrm{d} \in \mathcal{D}} \log(\mathbf{p}_{\mathrm{d}}) \tag{23}$$

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 (Schill *et al.* 2023) and utilize the penalization

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

with $\theta_{i,j} = \log(\Theta_{\mathrm{i,j}})$, $(\omega_{\mathrm{M}})_j = \log((\Omega_{\mathrm{M}})_j)$, $(\omega_{\mathrm{P}})_j = \log((\Omega_{\mathrm{P}})_j)$. 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

$$l_{\mathcal{D}}(\Theta, \Omega_{\mathrm{M}}, \Omega_{\mathrm{P}}) - \lambda \mathrm{penal}(\Theta, \Omega_{\mathrm{M}}, \Omega_{\mathrm{P}}) \tag{26}$$

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

Results

We first assessed metMHN's ability to recover parameters in simulations. The exact simulation setup and the results are shown in Supplementary Section S2. Next, to further our understanding of metastatic spread in lung adenocarcinomas, we trained metMHN on 4852 paired and unpaired samples from the LUAD cohort of the MSK-IMPACT study. In the following section we describe the dataset and then present our key findings.

Data preparation

We retrieved the AACR GENIE 14.1 data release (Pugh *et al.* 2022) through synapse.org (The AACR Project Genie Consortium 2023). 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 Nguyen *et al.* (2022) and excluded samples where the status of the metastasis was unknown. The final dataset consisted of 453 matched primary tumor (PT)/metastasis (MT) samples, 2127 unpaired MT samples, 595 PT samples without corresponding metastases (seeding = 0), and 1677 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 2725 PT and 2580 MT samples from 4852 patients. Metadata for each sample also included the age of the corresponding patient at which the sample was reported (see Supplementary Fig. S7). These data inform 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 (Cheng *et al.* 2015). We annotated mutation data using OncoKB (Chakravarty *et al.* 2017) and filtered for variants likely to be functional, as outlined in Schill *et al.* (2023). Our analysis was restricted to genes consistently included in all versions of the MSK-IMPACT panel (The AACR Project Genie Consortium 2023). 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 (Gao and Baudis 2020). 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.

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 (Fig. 2). Reassuringly, the LUAD model confirms several interactions well-documented in the literature. Specifically, it identifies the strongly mutually suppressive relationship (evidenced by a bidirectional negative edge) between the principal drivers KRAS (M) and EGFR (M) (Unni *et al.* 2015; Skoulidis and Heymach 2019). 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 Nahar *et al.* (2018).

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) (Offin *et al.* 2019, Wohlhieter *et al.* 2020, Cai *et al.* 2022). 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 (Becchi *et al.* 2023). 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 (Imperial *et al.* 2019).

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 Fig. 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 Fig. 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 (ie, promoting effects) from these events to the seeding event (Fig. 2). These findings are substantiated by experimental evidence which indicate that activation of EGFR (Che *et al.* 2015, Tsai *et al.* 2015), inactivation of TP53 (Wang *et al.* 2009, Powell *et al.* 2014), and activation of MET (Chang *et al.* 2015, Yin *et al.* 2019) 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 (Ben-David and Amon 2019), specific copy number changes, like CDKN2A deletions, have been documented to sometimes occur early in lung adenocarcinoma development (Nahar *et al.* 2018, Watkins *et al.* 2020). 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

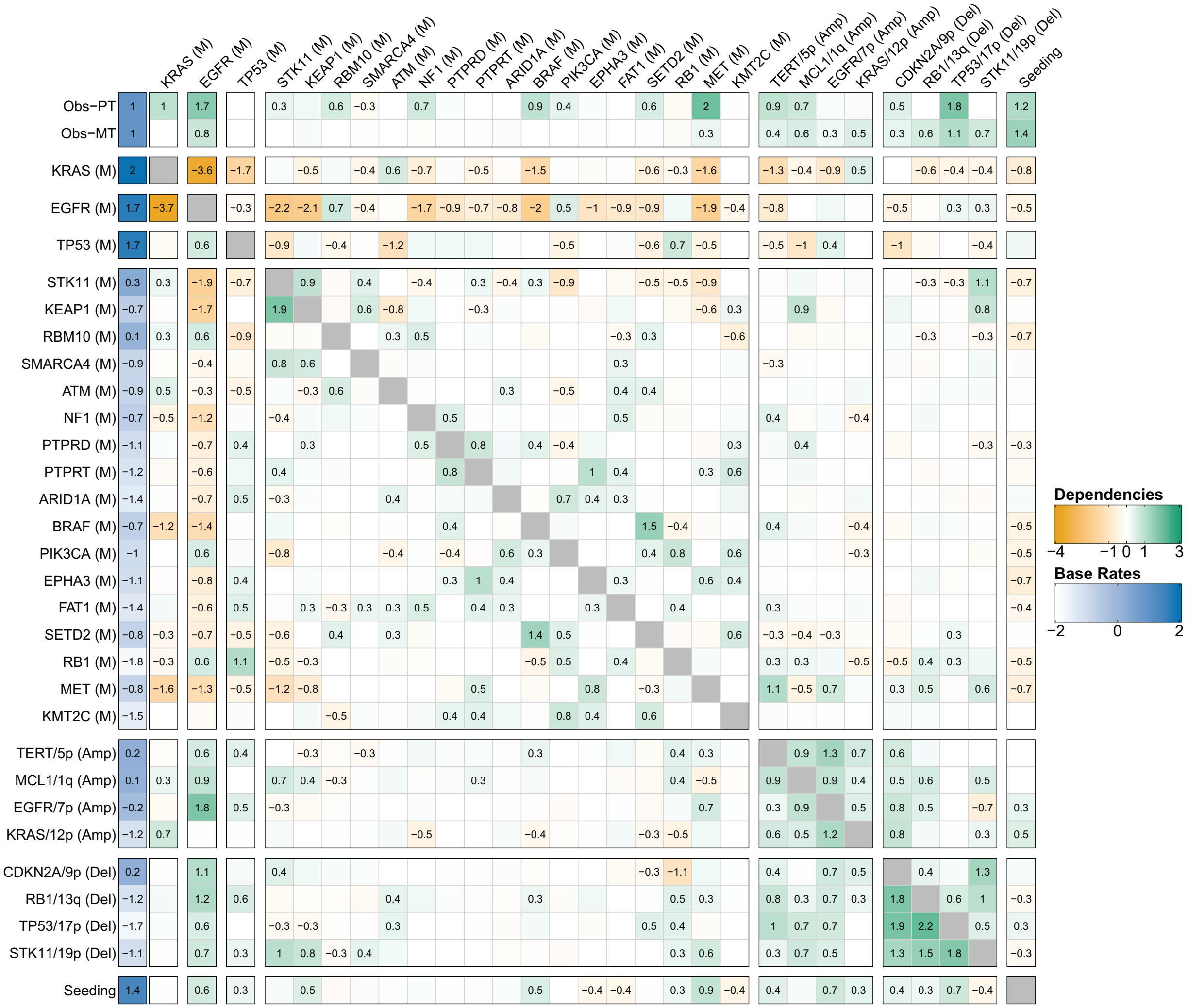

Figure 2. Interactions between progression events in lung adenocarcinomas. The log-effects on observation (clinical detection) of the primary tumor and metastasis ω_P and ω_M are plotted in the first two rows, the remaining matrix shows the log-interaction strengths among genomic events θ. Promoting effects are colored in green and suppressive effects are colored in orange. 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.

primary tumor cells may inherently possess a propensity to metastasize, as suggested by Klein (2020). 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 (Dymerska and Marusiak 2024).

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 Fig. 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 Fig. 3a, respectively). The unpaired metastases' orderings are linear.

Next, we analyzed the distribution of event positions in metastasis genotypes, relative to trajectory lengths (Fig. 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 Fig. 2 and Supplementary Fig. S5. 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 Fig. 2), leading to compounding rate increases for copy number events towards the end of a typical trajectory, possibly

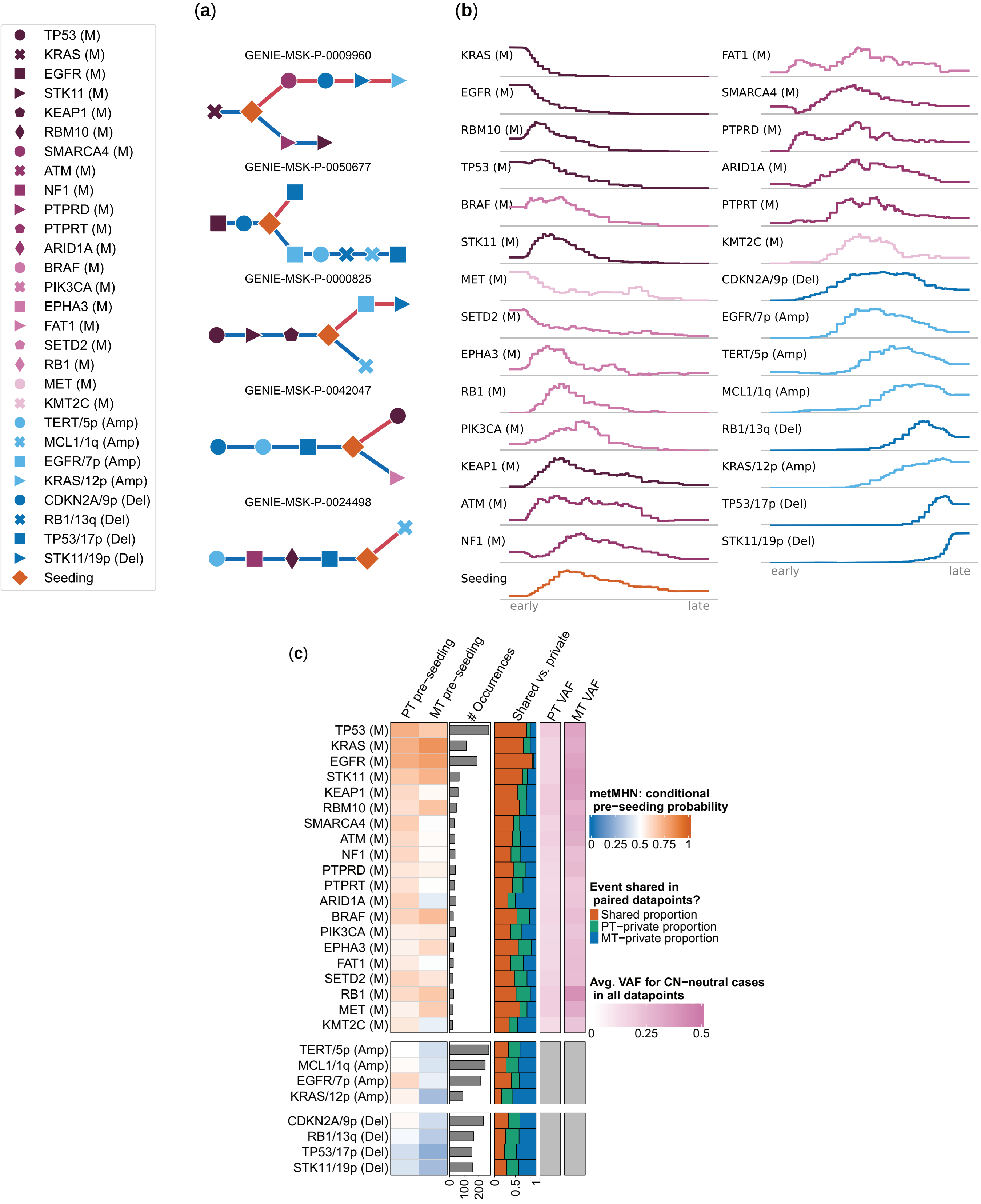

Figure 3. (a) Event orders for five patients as inferred by metMHN. Events accumulate from left to right. Blue edges represent the primary tumor development, red edges the one of the metastasis. Distances between events do not correspond to real or estimated time. (b) Distribution of relative positions in trajectories of metastasis genotypes. The left end of the axes corresponds to the beginning, and the right to the end of progression. (c) Empirical evidence from paired samples and pre-seeding probabilities estimated by metMHN through simulation. 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.

reflecting genomic instability in late stage cancers (Ben-David and Amon 2019).

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 1766 patients or 72.38% of the analyzed metastases (see Supplementary Fig. S5). Remarkably, the subset of trajectories initiated by TP53 (M) 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) 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.

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 (Birkbak and McGranahan 2020). 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 (Gillespie 1977). 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 ($P = .01$) see Fig. 3c and Supplementary Fig. S6. For a more detailed analysis of timing trends between pairs of genomic events, we provide a comparison of metMHN inferences with trends in phylogenetic analyses of TCGA primary tumors (Raynaud *et al.* 2018) in Supplementary Section S3. In summary, while metMHN builds on co-occurrence patters and does not leverage VAF information, they nevertheless produce results consistent with clonality information.

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. 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.

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. In terms of what we define as seeding event, the most accurate biological interpretation would be the onset of genetic divergence between the metastasis-seeding cell and its most recent detectable ancestor in the primary tumor (Sun and Nikolakopoulos 2021). Phylogenetic methods which use multi-regional samples have an advantage in accurately timing this event. Furthermore, this notion of seeding does not necessarily correspond to the seeding cell leaving the primary tumor, nor does it necessarily correspond to the establishment of the seeding cell at its metastatic site (Sun and Nikolakopoulos 2021).

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.

The complexity of cancer progression can exceed the capabilities of metMHN, for example, when patients present with numerous metastatic lesions or harbor disseminated cells that have yet to form detectable metastases. Various factors, including treatment modalities, genetic predispositions, age, inflammation, and other comorbidity conditions may further influence disease progression.

In summary, metMHN is a cancer progression model providing a quantitative and dynamic description of tumor development and metastatic seeding. It can be learned from currently available large clinical genomic datasets comprising cross-sectional bulk sequencing data.

Acknowledgements

We thank Nikolaus Schultz from MSKCC, Thomas Ratzke, Stefan Vocht, and Stefan Hansch for helpful discussions.

Author contributions

Kevin Rupp, Rudolf Schill, Niko Beerenwinkel, and Rainer Spang conceptualized and initiated the project. Kevin Rupp, Rudolf Schill, and Yanren Linda Hu developed the model. Kevin Rupp, Yanren Linda Hu, and Chenxi Nie implemented the algorithms. Simon Pfahler, Maren Klever, Tilo Wettig, and Lars Grasedyck provided numerical foundations for model analysis. Andreas Lösch prepared the input data. Andreas Lösch, Yanren Linda Hu, and Kevin Rupp analyzed the LUAD data. Kevin Rupp, Andreas Lösch, Yanren Linda Hu, and Rainer Spang drafted the manuscript. All authors critically read and improved upon the draft.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interests

None declared.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft DFG with TRR 305-A02 and GR-3179/6–1, the Swiss National Science Foundation grant 179518, and the Swiss Cancer League grant KFS-2977–08-2012 .

Data availability

The original genomic and clinical data for the MSK-LUAD cohort underlying this article are available at synapse.org (The AACR Project Genie Consortium 2023). The processed event data in metMHN input format as well as an exact record of used samples are available in the metMHN repository at https://github.com/cbg-ethz/metMHN. Subclonal reconstructions on TCGA data used in the Suplementary analysis are avlable as stated in Raynaud *et al.* 2018.

References

Becchi T, Beltrame L, Mannarino L *et al.* A pan-cancer landscape of pathogenic somatic copy number variations. *J Biomed Inform* 2023;**147**:104529.

Beerenwinkel N, Eriksson N, Sturmfels B. Conjunctive bayesian networks. *Bernoulli* 2007;**13**:893–909.

Ben-David U, Amon A. Context is everything: aneuploidy in cancer. *Nat Rev Genet* 2019;**21**:44–62.

Birkbak NJ, McGranahan N. Cancer genome evolutionary trajectories in metastasis. *Cancer Cell* 2020;**37**:8–19.

Buis PE, Dyksen WR. Efficient vector and parallel manipulation of tensor products. *ACM Trans Math Softw* 1996;**22**:18–23.

Cai L, DeBerardinis RJ, Xiao G *et al.* A Pan-Cancer assessment of RB1/TP53 Co-Mutations. *Cancers (Basel)* 2022;**14**:4199.

Caravagna G, Giarratano Y, Ramazzotti D *et al.* Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat Methods* 2018;**15**:707–14.

Chakravarty D, Gao J, Phillips S *et al.* OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* 2017;**1**:1–16.

Chang CC, Hsieh TL, Tiong TY *et al.* Regulation of metastatic ability and drug resistance in pulmonary adenocarcinoma by matrix rigidity via activating c-Met and EGFR. *Biomaterials* 2015;**60**:141–50.

Che TF, Lin CW, Wu YY *et al.* Mitochondrial translocation of EGFR regulates mitochondria dynamics and promotes metastasis in NSCLC. *Oncotarget* 2015;**6**:37349–66.

Cheng DT, Mitchell TN, Zehir A *et al.* Memorial sloan Kettering-Integrated mutation profiling of actionable cancer targets (MSK-IMPACT). *J Mol Diagn* 2015;**17**:251–64.

Dymerska D, Marusiak AA. Drivers of cancer metastasis—arise early and remain present. *Biochim Biophys Acta - Rev Cancer* 2024;**1879**:189060.

Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990;**61**:759–67.

Gao B, Baudis M. Minimum error calibration and normalization for genomic copy number analysis. *Genomics* 2020;**112**:3331–41.

Georg P. *Tensor train decomposition for solving high-dimensional mutual hazard networks*, PhD thesis, Universität Regensburg 2022; https://doi.org/10.5283/epub.53004.

Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 1977;**81**:2340–61.

Greenbury SF, Barahona M, Johnston IG. HyperTraPS: inferring probabilistic patterns of trait acquisition in evolutionary and disease progression pathways. *Cell Syst* 2020;**10**:39–51.e10.

Hjelm M, Höglund M, Lagergren J. New probabilistic network models and algorithms for oncogenesis. *Journal of Computational Biology* 2006;**13**:853–65.

Imperial R, Toor OM, Hussain A *et al.* Comprehensive pancancer genomic analysis reveals (RTK)-RAS-RAF-MEK as a key dysregulated pathway in cancer: its clinical implications. *Semin Cancer Biol* 2019;**54**:14–28.

Klein CA. Cancer progression and the invisible phase of metastatic colonization. *Nat Rev Cancer* 2020;**20**:681–94.

Klever M, Georg P, Grasedyck L *et al.* Low-rank tensor methods for Markov chains with applications to tumor progression models. *J Math Biol* 2022;**86**:7.

Lambert AW, Pattabiraman DR, Weinberg RA. Emerging biological principles of metastasis. *Cell* 2017;**168**:670–91.

Luo XG, Kuipers J, Beerenwinkel N. Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees. *Nat Commun* 2023;**14**:3676.

Mina M, Iyer A, Ciriello G. Epistasis and evolutionary dependencies in human cancers. *Curr Opin Genet Dev* 2022;**77**:101989.

Nahar R, Zhai W, Zhang T *et al.* Elucidating the genomic architecture of asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing. *Nat Commun* 2018;**9**:216.

Nguyen B, Fong C, Luthra A *et al.* Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell* 2022;**185**:563–75.e11.

Offin M, Chan JM, Tenet M *et al.* Concurrent RB1 and TP53 alterations define a subset of EGFR-mutant lung cancers at risk for histologic transformation and inferior clinical outcomes. *J Thorac Oncol* 2019;**14**:1784–93.

Pfahler S, Georg P, Schill R *et al.* Taming numerical imprecision by adapting the KL divergence to negative probabilities, arXiv 2023;2312.13021.

Potapova TA, Zhu J, Li R. Aneuploidy and chromosomal instability: a vicious cycle driving cellular evolution and cancer genome chaos. *Cancer Metastasis Rev* 2013;**32**:377–89.

Powell E, Piwnica-Worms D, Piwnica-Worms H. Contribution of p53 to metastasis. *Cancer Discov* 2014;**4**:405–14.

Pugh TJ, Bell JL, Bruce JP, AACR Project GENIE Consortium, Genomics and Analysis Working Group *et al.* AACR project GENIE: 100, 000 cases and Beyond. *Cancer Discovery* 2022;**12**:2044–57.

Ramazzotti D, Caravagna G, Olde Loohuis L *et al.* CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* 2015;**31**:3016–26.

Raynaud F, Mina M, Tavernari D *et al.* Pan-cancer inference of intratumor heterogeneity reveals associations with different forms of genomic instability. *PLoS Genet* 2018;**14**:e1007669.

Schill R. Mutual Hazard Networks: Markov chain models of cancer progression. PhD thesis, Universität Regensburg 2022; https://doi.org/10.5283/epub.53417.

Schill R, Solbrig S, Wettig T *et al.* Modelling cancer progression using mutual hazard networks. *Bioinformatics* 2020;**36**:241–9.

Schill R, Klever M, Lösch A *et al.* Overcoming observation bias for cancer progression modeling. *bioRxiv*2023; 2023.12.03.569824.

Skoulidis F, Heymach JV. Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nat Rev Cancer* 2019;**19**:495–509.

Sun R, Nikolakopoulos AN. Elements and evolutionary determinants of genomic divergence between paired primary and metastatic tumors. *PLoS Comput Biol* 2021;**17**:e1008838.

The AACR Project Genie Consortium. Release 14.1-public, 2023. https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn52918985&type=ENTITY (20 December 2023, date last accessed).

Tsai MF, Chang TH, Wu SG *et al.* EGFR-L858R mutant enhances lung adenocarcinoma cell invasive ability and promotes malignant pleural effusion formation through activation of the CXCL12-CXCR4 pathway. *Sci Rep* 2015;**5**:13574.

Unni AM, Lockwood WW, Zejnullahu K *et al.* Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma. *Elife* 2015;**4**:e06907.

Wang SP, Wang WL, Chang YL *et al.* p53 controls cancer cell invasion by inducing the MDM2-mediated degradation of slug. *Nat Cell Biol* 2009;**11**:694–704.

Watkins TBK, Lim EL, Petkovic M *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* 2020;**587**:126–32.

Weinberg RA. *The biology of cancer*, 2nd ed. New York, New York: Garland Science, Taylor & Francis Group, 2014.

Wohlhieter CA, Richards AL, Uddin F *et al.* Concurrent mutations in STK11 and KEAP1 promote ferroptosis protection and SCD1 dependence in lung cancer. *Cell Rep* 2020;**33**:108444.

Yin J, Hu W, Fu W *et al.* HGF/MET regulated epithelial-mesenchymal transitions and metastasis by FOSL2 in non-small cell lung cancer. *Onco Targets Ther* 2019;**12**:9227–37.

Bioinformatics, 2024, 40, 140–150
https://doi.org/10.1093/bioinformatics/btae250
ISMB 2024