Elements and Evolutionary Determinants of Genomic Divergence Between Paired Primary and Metastatic Tumors

Ruping Sun\textsuperscript{a,b,*}, Athanasios N. Nikolakopoulos\textsuperscript{a,b,*}

\textsuperscript{a}Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA
\textsuperscript{b}Masonic Cancer Center, University of Minnesota, Minneapolis, MN, USA

\textbf{ARTICLE INFO}

\textit{Keywords:}
Metastatic-Primary Genomic Divergence
Between-Tumor Heterogeneity
Metastatic Dissemination Time
Tumor Growth Mode
Clonal Evolution
Metastatic Potential
Most Recent Detectable Variant
Sequencing Detectability

\textbf{ABSTRACT}

Can metastatic-primary (M-P) genomic divergence measured from next generation sequencing reveal the natural history of metastatic dissemination? This remains an open question of utmost importance in facilitating a deeper understanding of metastatic progression, and thereby, improving its prevention. Here, we utilize mathematical and computational modeling to tackle this question as well as to provide a framework that illuminates the fundamental elements and evolutionary determinants of M-P divergence. Our framework facilitates the integration of sequencing detectability of somatic variants, and hence, paves the way towards bridging the measurable between-tumor heterogeneity with analytical modeling and interpretability. We show that the number of somatic variants of the metastatic seeding cell that are experimentally undetectable in the primary tumor, can be characterized as the path of the phylogenetic tree from the last appearing variant of the seeding cell back to the most recent detectable variant. We find that the expected length of this path is principally determined by the decay in detectability of the variants along the seeding cell’s lineage; and thus, exhibits a significant dependence on the underlying tumor growth dynamics. A striking implication of this fact, is that dissemination from an advanced detectable subclone of the primary tumor can lead to an abrupt drop in the expected measurable M-P divergence, thereby breaking the previously assumed monotonic relation between seeding time and M-P divergence. This is emphatically verified by our single cell-based spatial tumor growth simulation, where we find that M-P divergence exhibits a non-monotonic relationship with seeding time when the primary tumor grows under branched and linear evolution. On the other hand, a monotonic relationship holds when we condition on the dynamics of progressive diversification, or by restricting the seeding cells to always originate from undetectable subclones. Our results highlight the fact that a precise understanding of tumor growth dynamics is the \textit{sine qua non} for exploiting M-P divergence to reconstruct the chronology of metastatic dissemination. The quantitative models presented here enable further careful evaluation of M-P divergence in association with crucial evolutionary and sequencing parameters.

\section{1. Introduction}

Reconstructing the chronology of metastatic spread in human cancers (Hu et al., 2019; Reiter et al., 2018, 2020; Yates et al., 2017) has profound clinical importance (Birkbak and McGranahan, 2020; Durrett et al., 2011) and significant potential in improving cancer treatment. It is an indispensable first step towards a deeper understanding of metastatic tumor progression. Evaluation of the genomic divergence between paired metastatic and primary tumors (M-P divergence) has become the \textit{de facto} starting point for tackling this challenging task. However, despite the imperative need for accurate translation of M-P divergence measurements to metastatic seeding times, disparate interpretations of such measurements still abide. Should we read a small M-P divergence as an indication of “early” metastatic seeding (Hosseini et al., 2016; Hu et al., 2019), or rather, as evidence that metastatic potential is acquired at a late stage (Yachida et al., 2010; Turajlic and Swanton, 2016; Yates et al., 2017) of primary tumor growth? This remains an open and challenging question. Before unraveling this puzzle, however, one needs to carefully characterize what exactly \textit{is being captured} on the trees of tumor evolution by M-P divergence measurements arising from next generation sequencing.

Somatic evolution takes years to decades before the founder cell of the primary tumor gets established (Jones et al., 2008). As it is a descendant of the founder that will be equipped with the ability to leave the primary site through the
initial steps of the invasion-metastasis cascade (Lambert et al., 2017), the dissemination event must happen late, in absolute terms. However, with the establishment of the founder granted, the connection between the mode of primary tumor expansion (Davis et al., 2016; Noble et al., 2019) and the relative timing of metastatic seeding remains unclear. Moreover, the distinction between the initiation of the founder and subsequent expansion is also blurred; for example, the definition of the tumor founder can be fuzzy, given the fact that it is still challenging to identify the true “clonal” variants using a limited number of tumor samples (Opasic et al., 2019; Werner et al., 2017).

While the expansion dynamics of the primary tumor in vivo are largely undetectable, several hypothetical modes of growth have been proposed. Namely, effective neutrality mode, where clonal composition progressively diversifies (Sottoriva et al., 2015); branching evolution mode, where several subclones appear late and dominate the growth (Turajlic et al., 2019); as well as, linear evolution mode with continuous clonal replacement (Davis et al., 2016). The latter two modes differ from progressive diversification due to the presence of ongoing clonal sweeps at regional (branching evolution) or broad (linear evolution) scales, which leads to patterns of sub-clonal convergence rather than divergence. These modes represent possible tumor growth dynamics, which are determined by evolutionary and ecological forces, such as selection upon (epi)genetic alterations (Sun et al., 2017b; Dinh et al., 2020), spatial structure (Noble et al., 2019), as well as, tumor micro-environments (Marusyk et al., 2014; Anderson et al., 2006; Edwards et al., 2020). Accordingly, different tumor types and even different stages of a tumor may involve various growth dynamics (Fukatsu et al., 2013; Jolly and Van Loo, 2018; Yang et al., 2019).

To correctly translate experimental measurements of M-P divergence into cancer cell’s migration history, it is necessary to decompose its intrinsic elements and understand their relation to the metastatic dissemination time under different tumor growth dynamics. A recently published, deme (or compartment) based spatial tumor growth model suggests that small divergence reflects early metastatic seeding (Hu et al., 2019). A caveat, however, is that deme-based models (by treating 5,000-10,000 genetically homogeneous cells as a single unit) are tailored to colorectal cancers and can translate only approximately the fitness of single cells (individuals) into deme-based growth rates (Wakano and Lehmann, 2014); furthermore the currently deployed strategy of randomly picking one compartment at a time for division (Sottoriva et al., 2013) imposes additional limitations on the expressiveness and the biological plausibility of the model. By exploring limited conditions, the deme-based model mainly depicts the patterns of progressive diversification (Noble et al., 2019; Sun et al., 2017b). Hence, other growth dynamics, such as branching and linear evolution, have not been investigated for this question. More importantly, despite some qualitative discussions (Turajlic and Swanton, 2016; Hu et al., 2019), a careful mathematical analysis relating measurable M-P divergence to the underlying phylogenetic tree patterns under different tumor growth dynamics, is still missing in the literature.

Genomic heterogeneity among tumor cells is notably fueled by the passenger somatic variants, which occur several orders of magnitude more frequently than the driver variants during cell divisions (Bozic et al., 2016). On the other hand, the bulk sequencing of tumor samples can only reveal variants that are present in a detectable fraction of cancer cells. Therefore, in conjunction with the underlying evolutionary forces, the observed M-P divergence is actively shaped by the experimental limitations in detectability of somatic variants as well. Here, we take the first step towards a quantitative characterization of measurable M-P divergence as a function of seeding time and sequencing detectability; our analysis starts from the establishment of the founder, and our approach involves both mathematical modeling of the fundamental M-P divergence patterns under distinct tumor growth dynamics, as well as computational modeling of spatial growth kinetics subject to elaborate local density constraints.

2. Results
2.1. The elements of M-P divergence on the clonal evolutionary tree

We assume that the primary tumor growth begins with a single founder cell (Nowell, 1976). We focus on monoclonal seeding (El-Kebir et al., 2018) at a time during the growth of the primary, i.e., individual progeny cells of the founder disseminate when the primary tumor reaches a size of $N_d$ (see Methods). The terms “dissemination” and “seeding” are used interchangeably in this paper as we do not model the processes in between. The primary and the single-clone-seeded metastatic tumor are detected until they reach size $N$, respectively. On the phylogenetic tree built from metastatic and primary tumor sample(s), the M-P divergence typically involves the metastatic specific Branch ($B_m$), i.e., number of (#) somatic variants ($v$) found in the metastasis ($m$) but not in the primary ($p$): $B_m = \# \{ v | f_{v}^{m} > \gamma, f_{v}^{p} < \alpha \}$ and the primary specific branch $B_p = \# \{ v | f_{v}^{p} > \gamma, f_{v}^{m} < \alpha \}$, respectively (Turajlic and Swanton, 2016; Reiter et al., 2017; Hu et al., 2019) (Figure 1A). Here $\gamma$ is the variant allele frequency threshold above which substantial presence of a variant may be deemed. We assume that the variant detection is fair,
i.e., any variant in the tumor with a frequency above a detectability threshold $\alpha$ will be detected, e.g., through extensive multi-region sequencing (Uchi et al., 2016). The variants are assumed to be heterozygous, i.e., the cancer cell fraction thresholds will be at $2\gamma$ and $2\alpha$, respectively. While the somatic variants usually refer to point mutations and small indels, the framework could be extended for large structural changes and copy number alterations as well. $B_{md}$ can be further divided into variants accumulated before ($B_{md}^k$) or after dissemination. We focus on $B_{md}$ as it is related to the primary tumor growth and the genetic determinants of dissemination (Birkbak and McGranahan, 2020). $B_{md}$ refers to those variants that are NOT detectable in the primary tumor. $B_p$, on the other hand, includes all the variants with a frequency above $\gamma$ in the primary tumor except those appearing on the metastatic seeding cell. To illustrate $B_{md}$ and $B_p$ on the single-cell-based clonal evolutionary tree of primary tumors, we refer to the progenitor cells whose descendant population occupies a fraction of the primary greater than $2\alpha$ as “Detectable Ancestors at $\alpha$” (DA$\alpha$). $B_{md}$ is the branch length (measured in variants) from the metastatic seeding cell to its most recent DA at $\alpha$ (MRDA$\alpha$). $B_p$ is the total branch length of the genealogies of DA$\gamma$, after subtracting the branch between the founder and the MRDA$\gamma$ of the seeding cell (Figure 1A). Accordingly, the variants accrued in the DA$\alpha$ are Detectable Variants at $\alpha$ (DV$\alpha$). $B_{md}$ and $B_p$ involve the most recent detectable variants at $\alpha$ (MRDV$\alpha$) and $\gamma$ (MRDV$\gamma$), respectively.

To realize the definitions of $B_{md}$ and $B_p$, we implemented Comet, a single-cell-based spatial tumor growth model to simulate virtual primary and metastatic tumor pairs and collect the corresponding virtual sequencing data (see Section 4, and Methods). The phylogenetic distance that defines $B_{md}$ on the evolutionary trees of the virtual primary tumor explains more than 80% of the variance of $B_m$ measured from virtual sequencing data. By contrast, the tumor size (as a surrogate of time) when the seeding cell disseminates does not explain reliably the measured divergence (Figure 2A). Notably, although $B_m$ seems to be positively associated with tumor size when the primary tumor grows rapidly, i.e., with a small death-to-birth ratio ($\rho = 0.72$) (Diaz et al., 2012) and weak subclonal selection ($s < 0.05$), it displays a characteristic “zig-zag” pattern when the tumor expands under high death rate ($\rho = 0.99$) (Bozic et al., 2009) and strong selection ($s \geq 0.1$) (Figure 2B). On the other hand, $B_p$ displays little variability when tumor proliferates quickly, whereas it exhibits a “decreasing staircase” pattern when the tumor grows under strong selection (Supplementary Figure 1B). The phylogenetic definition of $B_p$ in the single cell trees explains 90% of the variance of the measured $B_p$ under models involving strong selection (Supplementary Figure 1A). Moreover, the ranges of $B_m$ and $B_p$ become wider as the tumor grows larger. The time-series that depict $B_m$ and $B_p$ across seeding times are heteroscedastic under selective growth modes, i.e., their variability is different across time (Median $p$-value < 0.01 for $s \geq 0.1$, White’s test for heteroscedasticity (Doornik, 1996), Supplementary Figure 2).

### 2.2. Non-monotonic dependence of M-P divergence on dissemination time

The zig-zag pattern of $B_m$ suggests that the number of metastasis-specific variants accumulated before dissemination, $B_{md}$, is not monotonically related to the metastatic dissemination time. Assume that seeding of a cell that has $k$ variants results in $B_{md}^k$. We want to calculate $B_{md}^{k+1}$ when this cell stays in the primary tumor and one of its progeny cells, which has just acquired one more variant, seeds the metastasis instead (i.e., the new seeding cell has $k + 1$ variants in total). Assume that the progenitor cell which stays in the primary eventually grows into a subclone that occupies a fraction $f$ of the entire tumor. If $f$ is too small to be detectable ($f < 2\alpha$), $B_{md}^{k+1}$ would increase from $B_{md}^k$ by one. However, if $f$ is large enough to be detectable ($f \geq 2\alpha$), the progenitor cell with $k$ variants would become a detectable ancestor (DA$\alpha$), and so do all of its ancestors. In this case it would be the most recent detectable ancestor at $\alpha$ (MRDA$\alpha$) of the new seeding cell and $B_{md}^{k+1}$ would only refer to the newborn variant (Figure 1B). More generally, the birth of a detectable subclone ensures detectability of the variants harbored by the subclone founder, and when the seeding cell belongs to its progeny, the variants in MRDA$\alpha$ will be subtracted from $k$ to yield the corresponding $B_{md}^k$. Therefore,

$$B_{md}^k = k - V_{MRDA}(k, \alpha) \quad (1)$$

where $V_{MRDA}(\cdot, \cdot)$ is a function that takes as an input a given seeding cell with $k$ variants and a detectability frequency threshold $\alpha$, and outputs the number of variants in its MRDA$\alpha$. Similarly, if we use $B_p^k$ to express the number of detectable primary-specific variants that are absent in the seeding cell, we have

$$B_p^k = V_{PRIMARY}(\gamma) - V_{MRDA}(k, \gamma) \quad (2)$$

where $V_{PRIMARY}(\gamma)$ denotes the number of variants with frequency above a threshold $\gamma$. 

---

Sun and Nikolakopoulos Page 3 of 20
Figure 1: Elements of Metastatic-Primary genomic divergence and their potential pattern in tumor evolution. (A) A sample tree (schematic) based on somatic variant analysis of paired Metastatic (M) and Primary (P) tumor samples depicts the two branches involving the number of metastatic specific \( B_m \) (dashed green line) and primary specific somatic variants \( B_p \) (dashed blue line), respectively. We focus primarily on the variants accumulated in the metastatic seeding cell (green circle) before its dissemination \( B_{md} \), as it is directly related to the primary tumor growth and the genetic determinants of metastatic dissemination. Below the sample tree shown a genealogy tree focusing on the lineage of the seeding cell, where we note the branches that represent \( B_{md} \) and \( B_p \), respectively. Detectable ancestors in the primary at a frequency of \( \gamma (DA\gamma) \) and the most recent detectable ancestor at \( \alpha (MRDA\alpha) \) of the seeding cell are also indicated, respectively. (B) If a cell with \( k \) variants “cancels” its dissemination (i.e., stays in the primary tumor) and its progeny cell with \( k+1 \) variants migrates instead, the resulting \( B_{md} \) can drop if the progenitor cell with \( k \) variants eventually grows into a detectable subclone in the primary tumor. In this case, it becomes the most recent detectable ancestor of the seeding cell. More generally, \( B_{md} \) can decrease if an ancestor along the genealogy the the seeding cell becomes detectable due to subclone expansion. (C) We hypothesize that \( B_{md} \) and \( B_p \) are affected by the kinetics of primary tumor growth. Under the growth pattern of progressive subclonal divergence, late seeding would lead to larger \( B_{md} \) than early seeding as only the early subclones are detectable; \( B_p \) stays constant as the detectable portion of the seeding cell lineage is restricted. By contrast, detectable subclones can appear late leading to the pattern of sub-clonal convergence (due to selection or other ecological forces), under which scenario seeding from early or late detectable subclones may lead to similar \( B_{md} \), but differs in \( B_p \).

While the total number of variants in the seeding cell indeed increases with dissemination time, the number of detectable ones is heavily influenced by the probability of subclonal expansions; a fact that further complicates the study of the relation between \( B_{md}^k \) and \( B_p^k \) and seeding time (Figure 1C).

### 2.2.1. Mathematical insights on the expected \( B_{md} \) patterns

Let \( d_{cj} \) denote the probability variant \( c_j \) ends up being detectable in the primary tumor. In Methods we show that under the infinite allele model, the probability \( B_{md}^k \) takes the value \( i \) can be expressed in a surprisingly simple form in terms of \( d_{cj} \), as

\[
\Pr[B_{md}^k = i] = d_{ck-i} - d_{ck-i+1}.
\]

Equation (3) essentially relates the length of \( B_{md} \) with the probabilistic decay in detectability along the lineage of the seeding cell. Parameter \( k \) refers to the number of variants that are present in the seeding cell, effectively acting as a
Figure 2: Single-cell-based spatial simulation of paired metastatic and primary tumors validates the phylogenetic definition of \( B_m \) and emphasizes that \( B_m \) can appear sensitive or insensitive to seeding time based on the kinetics of primary tumor growth. (A) \( B_m \), the sub branch length (measured in cell generations) from the metastatic seeding cell to its most recent detectable ancestor (MRDA) captures the majority of the variance of \( B_m \) measured from the virtual tumors (green density curves). By contrast, the tumor size at metastatic dissemination, as a surrogate of dissemination time (gray density curves) does not faithfully explain the measured \( B_m \). (B) Two virtual tumors with distinct kinetics exemplify the conditional dependence of \( B_m \) on metastatic dissemination time. Upper panel: the genealogy trajectories (gray lines) of metastatic seeding cells (green dots) along the expansion of the primary tumor in a three dimensional lattice. The Euclidean distance from a cell to the center of the lattice is shown against the tumor size fraction when the corresponding cell disseminates. The tumor size fraction is plotted at its cube root scale to reflect the clock of actual time. Ancestor cells that are detectable at a frequency greater than 0.01 are marked as blue dots; Middle panel: \( B_m \) is plotted against the tumor size fraction when the seeding cell (green dots) disseminates, the running mean (black dots) and standard deviation (gray bars) is also shown; Lower panel: \( B_m \) is shown against the \( B_m \), the sub-branch length from the seeding cell to its MRDA. The slope of the linear regression is consistent with the passenger variant rate (0.15) used in the model, CI: 95% confidence interval.

clock for the dissemination time. Calculating \( B_m \) for different values of \( k \) enables us to reason about the expected \( B_m \) under settings of early (small \( k \)) or late (large \( k \)) dissemination. It also allows us to reason about the monotonicity of the sequence \{ \( B_m \) \}_{k=0,1,...} \), which is crucial for accurate estimation of seeding time, based on specific genomic divergence measurements. In particular, notice that a strictly increasing \{ \( B_m \) \}_{k=0,1,...} \) would allow, in principle, to unambiguously relate divergence measurements with seeding times; on the contrary, if the monotonicity of \{ \( B_m \) \}_{k=0,1,...} \) is not guaranteed, an unequivocal interpretation of generic divergence measurements cannot be guaranteed in advance—at least not without further assumptions regarding the growth dynamics of the tumor under study, or additional information.
that could allow one to break this ambiguity.

To further explore the implications of equation (3) in the analytically tractable case of neutral growth, building on the results of (Bozic et al., 2016), we derive an approximate form for the expected $B_{md}$ as a function of the seeding time $K$, the detectability threshold $f$, the mutation rate $u$, and the death-to-birth ratio of the tumor cells $\rho$

$$\mathbb{E}[B_{md};K, f, \rho] = \sum_{j=1}^{K} \left( \frac{u}{u - \log(1 + (\rho - 1)f)} \right)^{g(c_{k-j})} - \left( \frac{u}{u - \log(1 + (\rho - 1)f)} \right)^{g(c_{k-j+1})}$$

(4)

where $g(\cdot)$ is an approximate mapping that associates the variants that are particular to the seeding cell, with their absolute order of appearance in the overall primary tumor (specified in Section 7.1.2).

Importantly, we find that under neutral growth the resulting sequence $\{\mathbb{E}[B_{md};K, f, \rho]\}_{k=0,1,...}$ is in fact monotonically increasing with time. In particular, one can discern two distinct phases, the shape of which is modulated by the parameters of the model; an initial phase where the increase in expected $B_{md}$ is almost imperceptible (the length of this phase is controlled by $\rho$); and a later phase of steady increase in expected $B_{md}$ at almost constant rate (that is not particularly sensitive to $\rho$). Supplementary Figure 4 demonstrates the related patterns. Observe that larger values of $\rho$ are indeed associated with larger periods of time for which $B_{md}$ is expected to be very small. However, invariably, as more and more variants are being introduced in the primary tumor, their detectability probability plummets, thereby leading to a predictable increase in expected $B_{md}$ in late seeding settings. This is a direct consequence of the exponential decay in detectability probabilities of newly emerging variants down the phylogenetic tree. Intuitively, when seeding happens late under neutral growth, the later accumulating variants on the seeding cell are almost surely undetectable, and the depth of the MRDA $f$ converges to a fixed value that is no longer affected by $K$.

The situation, however, is markedly different in the non-neutral growth setting. To see this consider the case where a new type of advantageous cell is introduced to the primary tumor at time $T_1$. Also assume that the overall population of cells that carry the beneficial variant in the primary tumor (the one that characterizes the advantageous type) reaches a fraction $f_1$. When dissemination happens at time $k < T_1$ the seeding cell originates necessarily from the original type of cells, and thus, the above reasoning still applies. On the other hand, if dissemination happens at time $k > T_1$, the value of $B_{md}$ will be predominantly determined by whether the seeding cell carries the beneficial variant or not. Indeed, if the seeding cell originates from the beneficial subpopulation, the expected $B_{md}$ will be inevitably smaller since only variants that are introduced after the beneficial one (variants arising after time $T_1$) could end up being undetectable.

The detectability of the beneficial type itself ensures such outcome.

Thus, to reason about the effect of the introduction of new types on the expected $B_{md}$, we need to explicitly account for the probability the seeding cell originates from the subpopulation of advantageous cells as a function of seeding time, $k$. We denote such probability $\phi^k_0$ (similarly we use $\phi^k_1 = 1 - \phi^k_0$ to denote the probability of seeding from the original type of cells), and express $\{\mathbb{E}[B_{md};k]\}_{k=1,...}$ as

$$\mathbb{E}[B_{md};K, f_1, f_0] \approx \phi^k_1 \times \mathbb{E}[B_{neutral};K - T_1, f_1, f_0, \rho_0] + \phi^k_0 \times \mathbb{E}[B_{neutral};K, f_1, f_0, \rho_0]$$

(5)

In Methods we explore several different choices of $\{\phi^k_0\}_{k=1,...}$, $\{\phi^k_1\}_{k=1,...}$ that can capture different assumptions regarding the underlying seeding potential of the different types of cells that are present in the population. Notably, our findings suggest that the monotonicity of the sequence $\{\mathbb{E}[B_{md};k]\}_{k=1,...}$ is expected to break by the introduction of newly emerging advantageous types. Such effect is unavoidable when dissemination capability is positively associated with fitness advantage, however we find that the expected drop in $B_{md}$ remains significant even when seeding happens uniformly at random from all living cells (Figure 6; third column).

### 2.3. Single-Cell Spatial Tumor Growth Simulation Reveals that Dependence of M-P Divergence on Seeding Time is Conditional on the Underlying Modes of Tumor Growth

To corroborate the mathematical insights in a spatial growth setting, we develop Comet; a state-of-the-art simulator of paired primary and metastatic tumor growth. From a technical point of view Comet extents the single cell based spatial tumor growth model (Iwasaki and Innan, 2017). Various parameter settings of driver genomic variants and cell death-to-birth rates enable our model to encompass three distinct and representative tumor growth dynamics:

---

2 subject to a modification of the detectability threshold which now needs to account for the fact that the relative frequency that a variant needs to achieve to become detectable in the overall tumor at detection time is now $f/(1 - f_1)$. 

---
progressive diversification, branching, and linear evolution (see Methods, Table 1). In addition to driver events, we simulate passenger variants that occur at a given rate $u$, and record variant lineages as the tumor expands. To track the changes of M-P divergence along with the expansion, our model allows monoclonal dissemination and metastatic seeding as the population size of the primary tumor increases (every 25,000 cells). Both virtual primary and metastatic tumors reach a final volume of $10^7$ cells (Dinh et al., 2020). We perform multi-region sampling of the virtual tumors and calculate the variant allele frequencies by assuming a negative binomial distribution of the sequencing depth (Anders and Huber, 2010) for each sample. $B_{md}$ and $B_p$ is measured from the virtual sequencing data accordingly.

Our model makes it possible to study both subclonal diversification and subclonal convergence in the spatial tumor growth setting (Figure 4 and Figure 5A). We note that while the parameters used are model-dependent, the clonal evolutionary patterns realized by different levels of positive selection and death-to-birth ratio are compatible with those caused by other evolutionary and ecological forces. For example, competition between tumor and normal cells, which is not modelled here, can amplify clonal selection (Noble et al., 2019). Weak selection ($s < 0.05$) coupled with small death-to-birth ratio ($\rho = 0.72$) leads to a pattern of progressive diversification which has been discussed previously (Sun et al., 2017a). Interestingly, increasing the selection coefficient further facilitates late subclonal sweeps at different scales (Figure 4A, upper panel). Moderately strong selection ($s = 0.1$) leads to a branched evolution pattern where a few detectable subclones jointly constitute the whole tumor. High death-to-birth ratio ($\rho = 0.99$) coupled with strong
Figure 4: Virtual tumors with three representative clonal kinetics and the changing patterns of M-P divergence along with the metastatic seeding time. Upper panel: the genealogy trajectories (gray lines) of metastatic seeding cells (green dots) during the expansion of the primary tumor in a three-dimensional lattice. The Euclidean distance from a cell to the center of the lattice is shown against the fraction of the final primary tumor size when the corresponding seeding cell is born and disseminated. The tumor size fraction is plotted at its cube root scale to reflect the clock of actual time. Blue dots represent progenitor cells that are detectable at a frequency higher than 0.01. Middle panel: $B_m$ is plotted against the tumor size fraction when the seeding cell disseminates, where black dots and gray bars indicate the running mean and standard deviation of $B_m$, respectively. Lower panel: change of $B_p$ with different seeding time, blue dots and bars represent the running mean and standard deviation of $B_p$, respectively.

The results of the spatial tumor growth model are aligned with our mathematical insights, suggesting that the non-monotonicity of M-P divergence is ubiquitous in representative tumor growth models, regardless of the specific parametric settings. When the clonal composition of the primary tumor progressively diversifies, $B_m$ increases with seeding time (Figure 4, left column), and $B_p$ remains small. However, $B_m$ starts to become less sensitive to metastatic dissemination time when the tumor grows under deep branching evolution (Figure 4, middle column). Under linear growth dynamics, $B_m$’s non-monotonic relationship with seeding time is evident (right column). Notably, the mathematically predicted zig-zag pattern of the expected $B_{md}$ is evident in all virtual tumors. Under the dynamics of clonal convergence, $B_p$ has the tendency to transit to lower levels if the seeding cell disseminates later. The “decreasing staircase” pattern of $B_p$ is particularly evident for the linear growth mode. In the virtual primary tumors, the between-region genetic divergence (measured by the Fixation index (Bhatia et al., 2013)) initially increases with the selection coefficient when $s < 0.05$, in agreement with our previous findings using a deme-based model (Sun et al., 2017b). However, higher and more influential selection levels ($s \geq 0.1$) reduce the genetic diversity in the primary tumor and create the dynamics of subclonal convergence (Figure 5A, Supplementary Figure 3). In the meantime, more variants...
of the seeding cell become detectable in the primary tumor on average (i.e., less variants contribute in $B_m$) as we raise the selection coefficient. When $s \geq 0.1$, more than 50% of the seeding cell lineage on average is detectable in the primary tumor, and $B_m$ starts to become much less sensitive to seeding time (Figure 5A).

In summary, the metastatic seeding time is a radically different function of $B_m$ and $B_p$ under distinct dynamics of primary tumor growth (Figure 5B). Under the mode of progressive diversification ($s \leq 0.05$), $B_m$ is the most informative indicator. For linear evolution ($s \geq 0.5$), the associated step-wise drops in $B_p$ may be leveraged to inform the timing. Under branched evolution, however, the situation is significantly more perplexed; large $B_m$ and $B_p$ readings point to late seeding, whereas mid-range readings suggest a non-trivial—possibly non-linear—relationship between $B_p$, $B_m$ pairs and metastatic dissemination time. Furthermore, the non-monotonicity of M-P divergence over time leads to increased uncertainty in mapping certain pairs of $B_m$ and $B_p$ onto metastatic initiation times (variegated pattern, Figure 5B). In the latter setting, additional indicators may be necessary to allow an accurate statistical reconstruction of the metastatic dissemination chronology.
3. Discussions

To reliably bridge the scales between the divergence at the genomic level and the metastatic spread at the cellular level, we focus on if and how dissemination time and tumor growth dynamics remodel the between-tumor heterogeneity. Our results pave the way towards building effective statistical models that connect M-P divergence to a metastatic timeline. Firstly, the growth-dependent relationship between $B_m$, $B_p$ and seeding time indicates that knowing the primary tumor growth dynamics is the prerequisite for translating M-P divergence. For example, a small $B_m$ and $B_p$ (Hu et al., 2019; Yates et al., 2017) may suggest early seeding for a tumor grown under progressive diversification, but late dissemination under linear evolution (Figure 5B). For neoplasms characterized by deep branching evolution, the timing of metastatic spread may be more difficult to identify. This is particularly relevant for tumors showing invasive glandular structure, which have been associated with deep branching mode (Noble et al., 2019). Secondly, the analytically predicted—and computationally verified—$B_m$, $B_p$ patterns indicate that more elaborate statistical models, that take into account non-monotonic dependencies, are necessary to utilize such measurements effectively. Lastly, the difference between $B_m$ and $B_p$, with the latter also being measured based on a lowered substantial presence threshold ($\alpha$), in theory cancels out the drop effect upon the introduction of a new detectable subclone and may inform better the seeding time. However this quantity can be heavily affected by the noise in variant detection at low frequencies. As the M-P divergence alone, may not always carry enough information to accurately infer the seeding time (under branching growth mode in particular), additional indicators, as well as other clinical information may be leveraged to increase the confidence of the inference (Cheek and Antal, 2018).

We mathematically showed that the M-P divergence also reflects the interplay between clonal dynamics and the subclonal preference of seeding (Reiter et al., 2018), which may imply the genetic determinants of metastatic dissemination (Nguyen and Massagué, 2007). Based on our discovery, one may quantitatively test hypotheses about such determinants using multi-region sequencing of multiple metastases and paired primary tumor of the same patient. For example, always seeding from a rare subclone that has metastatic-specific driver variants may lead to a $B_m$ that is more monotonically related to seeding time compared to planting from a major subclone which appears late and possesses a growth advantage. Moreover, the step-wise decrease of $B_p$, if seen in patients with multiple metastases, could suggest an increased probability of seeding from actively growing subclones (RS et al., 1988; Ramaswamy et al., 2003). On the other hand, the actual pattern of M-P divergence could also hint on an evolutionary bottleneck caused by microenvironmental factors. For example, seeding cells that carry fewer unique variants may experience a safer passage to distant sites, by reducing the chance of being eliminated by immune surveillance (Lambert et al., 2017).

The measured M-P divergence from Next Generation Sequencing is not only determined by evolutionary forces but also affected by the detection regime. Usually, only a handful of samples of each lesion are collected for sequencing (Birkbak and McGranahan, 2020), and their spatial location in the corresponding tumor is unknown. Depending on the sampling location and clonal structure of the tumors, $B_m$ and $B_p$ can be artificially enlarged or reduced. For example, if we reduce the number of virtual samples from ten to five, while the effect of evolutionary determinants is still evident, the same M-P divergence could map to a different seeding time due to the change in sampling alone (Supplementary Figure 8). In addition, primary tumor samples often come from formalin-fixed, paraffin-embedded tissues that have been stored for much longer than the paired metastatic samples, leading to unequal DNA quality between the two (Savas et al., 2016). Between-sample heterogeneity can be further over- or under-estimated due to low tumor purity. As the ploidy and purity are often jointly estimated (Ha et al., 2014), the copy number evolution in the metastatic progression of certain tumor types (Bakhoum et al., 2018; Sansregret and Swanton, 2017) may confound the quantification of genomic divergence between samples. Therefore, an assessment of the variability in heterogeneity due to suboptimal detection is necessary before comparing the real data with computational models, and is thus, beyond the focus of this study. Given the increasing emphasis on obtaining longitudinal samples and single-cell sequencing, we envision that some hypotheses generated by our models may be better tested by combining these two, complementary sequencing data types.

In our framework, time zero is set at the birth of the tumor founder cell, for which the full metastatic competence may not necessarily have been bestowed. We make minimal assumptions about the determinants of seeding potential by only requiring that it is the progeny cell of the founder that spreads, owing to genetic or non-genetic factors. If one assumes that the founder is a fully transformed (malignant) cancer cell (Hu et al., 2019), which grows under progressive diversification in vivo (Sottoriva et al., 2015), early seeding is more likely given a small M-P divergence. However, this is circular reasoning as the starting cell already has the full potential of metastasis, and short waiting time before seeding is not surprising. Another related consideration is that the actual primary tumor size, as a surrogate of time,
should be determined by all leaves originated from the founder cell, not just those of the more malignant subclone (Magrì and Bardelli, 2019). Our model also considers several distinct but representative growth dynamics of the founder cell’s expansion. This flexible setting helps us to chart the landscape of connections between M-P divergence and metastatic seeding times.

Since this is the first attempt to analytically characterize M-P divergence patterns, we focus on key but inevitably simplified considerations; nevertheless, our framework can serve as a foundation for future data-driven efforts aiming at understanding more complex metastatic dynamics. The actual mode of primary tumour growth may vary when the physiological conditions of the micro-environments vary, resulting in complex dynamics that our model does not capture. For example, clonal competition could increase, and the tumour may slow down in terms of growth on account of crowding over the available resources (Alzahrani et al., 2014). As a result, M-P divergence generated by a real patient tumor is even less likely to be monotonic as the tumor expands. On the other hand, while we do leave the dynamics of metastatic tumor unexplored by focusing on $B_{md}$, the modes of the primary tumor can also be applied to the secondary tumor. We expect that the length of $B_m$ rather than $B_p$ is affected by the growth pattern of metastatic tumor given its fitness in the distant site and possible iatrogenic perturbations (Pich et al., 2019). We analyzed monoclonal seeding solely, which accounts for the majority of distant metastasis (El-Kebrir et al., 2018), but the analytical thinking is extendable to reason about more complex poly-clonal or consecutive seeding (Gundem et al., 2015; Heyde et al., 2019). While our analysis here is applied to the specific case of metastatic progression, the usefulness of our approach extends to the case of comparing longitudinal tumor samples, as well. Finally, our findings emphasize that investigating tumor growth dynamics in detail has important analytical purchase, particularly for the efforts that seek to translate heterogeneity among longitudinal samples to the evolutionary timeline of progression.

4. Acknowledgments

This work was funded by the Department of Laboratory Medicine and Pathology and the Masonic Cancer Center at the University of Minnesota. This work was also funded by the Karen Wyckoff Rein in Sarcoma Foundation. This study uses the computing resources in Minnesota Supercomputing Institute. We thank Dr. Jasmine Foo and Zicheng Wang for commenting on the analytical framework. We also thank Dr. Boyang Liu and Dr. Jie Ding for helpful discussions. We thank our family members for their support.

5. Author Contributions

Conceptualization, R.S. and A.N.N.; Methodology, R.S. and A.N.N.; Software, R.S. and A.N.N.; Formal Analysis, R.S. and A.N.N.; Investigation, R.S. and A.N.N.; Data Curation, R.S. and A.N.N.; Writing—Original Draft, R.S. and A.N.N.; Writing—Review & Editing, R.S. and A.N.N.; Visualization, R.S. and A.N.N.; Funding Acquisition, R.S.; Resources, R.S.; Supervision, R.S..

6. Declaration of Interests

The authors declare that there is no conflict of interest.

7. Methods

7.1. Mathematical Modeling

7.1.1. Tree-based characterization of $B_{md}$ and $B_p$

Consider the phylogenetic tree of the surviving variants in the primary tumor, i.e., the tree rooted on the variant that characterizes the founder cell, as well as the additional variants that are accumulated in the primary tumor as it evolves over time. Let us index such variants according to their order of appearance in the primary tumor, as $\mathcal{V} = (v_0, v_1, \ldots, v_{k_{primary}})$. At a certain point in time a cell departs from the primary tumor and becomes the founder cell of a metastatic tumor. We call this event, metastatic seeding, and we refer to the time this event takes place, as seeding time; henceforth denoted as $k_{seed}$. Notice that the variants that are present on the seeding cell, $C$, at seeding time, are by definition, also present in the primary tumor, i.e., $C \subseteq \mathcal{V}$. However, the final frequencies reached by a certain subset of such variants upon observation of the primary tumor, can potentially fall below the detectability threshold of current sequencing technology. This implies that only a subset of the variants, $C_{det} \subseteq C$, will be registered, in practice, as
Modeling Metastatic Primary Genomic Divergence

being part of the primary tumor. On the other hand, all such variants will be detectable on the metastatic tumor, since they will be present in 100% of the metastatic cells as part of the founder of the metastasis. Here are are interested in characterizing the number of variants in \( C \), that end up not being detectable in the primary tumor, \( B_{md}^k \), as well as the number of frequent variants in \( V \) that are absent from the metastatic tumor, \( B_p^k \).

Let \( c \) be a seeding cell and also let \( C = (c_0, c_1, \ldots, c_{k_{seed}}) \) be its associated set of variants, indexed according to the order of appearance in the tumor. Let \( d_{c_j} \) denote the probability variant \( c_j \) is detectable in the primary tumor, and also let \( B_{md}^k \) be the random variable that counts the number of variants in \( C \) that are not detectable in the primary tumor at observation time. Under the infinite allele model if a seeding-cell specific mutation \( c_j \), becomes detectable in the primary the same will be true for all \( c_i < c_j \). Thus, \( \{d_{c_j}\}_{j=0,1,...} \) will be a decreasing sequence that starts from \( d_{c_0} = 1 \) and goes to 0 with \( j \). Notice that for \( B_{md}^k \) to take the value \( j \), the final frequencies of variants \( c_{k_{seed}}, c_{k_{seed}-1}, \ldots, c_{k_{seed}-j+1} \) are necessarily below the detectability threshold, while \( c_{k_{seed}-j} \) is above. This leads to the probability distribution

\[
\Pr[B_{md}^k = k_{seed} - j] = \begin{cases} 
    d_{c_j} - d_{c_{j+1}}, & j < k_{seed} \\
    d_{c_{k_{seed}}}, & j = k_{seed}
\end{cases}
\]  

(6)

For a detailed derivation of (6) see Supplementary Methods 2.1.

We now turn our attention on \( B_p^k \). Let \( M \) be the set of variants that reach a frequency of at least \( \gamma < f_p < 1 \) in the primary tumor. By definition we have

\[
B_p^k(\gamma) = |M| - V_c^{k_{seed}}
\]  

(7)

where \( V_c^{k_{seed}} = |C_{seed} \cap M| \) is the number of variants of the seeding cell that are also included in \( M \). Taking into consideration the phylogenetic view detailed above, in conjunction with the definition of \( M \), reveals that \( V_c^{k_{seed}} \) is, in fact, equivalent to \( V_{MRDA}(k, \gamma) \). This allows us to express (cf. (1)) the sought quantity (also explicitly considering the reference frequency \( \gamma \)) as

\[
B_p^k(\gamma) = |M| - k + B_{md}^k(\gamma)
\]  

(8)

In practice, \( |M| \) can be inferred from multi region sampling of the primary tumor at detection time (Opasic et al., 2019) and is, therefore, assumed known for the scope of our analysis. Importantly, equation (8) reveals that \( B_p^k(\gamma) \) decays linearly with seeding time, \( k_{seed} \), while it is positively correlated with \( B_{md}^k(\gamma) \). Intuitively this suggests that sudden drops in the sequence \( \{B_{md}^k(\gamma)\}_{k=0,1,...} \) are expected to be mirrored by similar drops in \( \{B_p^k(\gamma)\}_{k=0,1,...} \), whereas ranges where \( \{B_{md}^k(\gamma)\}_{k=0,1,...} \) is increasing linearly with time, would be mapped to \( B_p^k(\gamma) \) subsequences that remain constant.

Having characterized the general form of \( B_{md}^k \) and \( B_p^k \), we will now investigate how can the \( B_{md}^k \) probability mass function can be approximated for the analytically tractable case of the exponential growth model without spatial structure. We first restrict our attention to the case of neutral growth, while also considering the accumulation of passenger variants that may become detectable in the observed tumor. We will then proceed to the investigation of the effects of the introduction of advantageous types of cells in the population.

7.1.2. \( B_{md}^k \) under neutral growth

Following (Bozic et al., 2016), we model tumor growth using a multi-type branching process that starts with a single type-0 cell. The cells divide with rate \( \lambda \) and die with rate \( \mu \). Upon each division, one of the daughter cells acquires a new passenger variant with probability \( u \), which denotes the birth of a new type. Such type, while not being differentiated in terms of \( \lambda \) and \( \mu \) from type-0, it is uniquely characterized by the newly-acquired genetic variant. To emphasize the fact that from a fitness point-of-view such new types are not advantageous with respect to type-0, we henceforth refer to them as subtypes of type-0. Any new variant that appears in the population can be lost due to stochastic fluctuations. The probability that its lineage will not survive is \( \rho = \mu / \lambda \), the ratio of the death and the birth rates.

Now consider a seeding cell \( c \) and let \( (c_0, c_1, \ldots, c_{k_{seed}}) \) be the set of variants that it carries, indexed according to their order of appearance in the tumor as above. Within the aforementioned tumor growth setting we are interested in deriving an approximation of the probability any given variant that is present in the seeding cell will be detectable in the final primary tumor. In particular, we are interested in an expression that also takes into account the limitations of
sequencing technology. Therefore, we assume that in order to be detectable a variant needs to be present in a fraction of cells of the overall population larger than \( f = 2\alpha \). Bozic et al. (2016) established that the final frequencies of passenger variants in the population is primarily affected by the order by which these variants appeared in the tumor. Therefore, a necessary step towards obtaining the sought approximation utilizing the results of (Bozic et al., 2016) is to derive a mapping between the relative ordering of the variants that are particular to the seeding cell, to their absolute order of appearance in the overall tumor population (see example Figure 3).

Note that such mapping is exactly specified by the topology of the resulting phylogenetic tree that arises by a realization of the aforementioned branching process. It is, therefore, inherently random in itself. However, for the purpose of our analysis, opting for simplicity, we treat this mapping as deterministic, utilizing basic properties of branching processes with exponentially distributed lifetimes. Concretely, consider the lineage of the seeding cell \( c \), i.e., the path in the phylogenetic tree, from the root (founder variants, i.e., \( c_0 \)) to the leaf (the last appearing variant, i.e., \( c_{\text{seed}} \)). The number of nodes in this path is by definition \( k_{\text{seed}} + 1 \) and they depict the different variants that are present on the seeding cell \( c \). Now, let’s associate each node with a positive real number referring to the time of appearance of the corresponding variant in the population. Due to the assumption of exponentially distributed waiting times between cell divisions, it follows that the time difference between the appearance of two consecutive variants along this path is going to also be exponentially distributed with rate \( u \times \lambda \). Thus, the distribution of the time needed for variant \( c_j \) to appear in the population, can be expressed as the sum of \( j \) independent exponentially distributed random variables with identical rates equal to \( u \times \lambda \). Therefore the successive expected times of appearance of the variants along the seeding cell’s lineage are given by \( t_0 = 0 \) (by assumption the time starts when the founder cell is born), and \( t_j = \frac{1}{u \times \lambda} \), for \( j = 1, \ldots, k_{\text{seed}} \).

Now, the expected number of cells in the population at a given time (conditioned on long-term tumor survival) can be calculated (see e.g., (Reiter et al., 2018)) as

\[
N(t) = \frac{1}{\rho} e^{\rho t} - \left( \frac{1}{\rho} - 1 \right) e^{-\lambda t}. \tag{9}
\]

Thus, substituting the expected times of appearance of the variants of the seeding cell, \( t_j \), to equation (9) we can get an estimate of the expected number of cells in the population, at the time each of the seeding-cell-specific variants appeared. Utilizing also the fact that the expected number of passenger variants with surviving lineage that are present in a population of \( N \) cells is \( N \times u \) (Bozic and Nowak, 2014), we derive the final expression that maps the \( j \)-th seeding-cell-specific variant, \( c_j \), to its absolute ordering in the overall tumor population, \( v_{g(c_j)} \):

\[
g(c_j) = u \left( e^{\frac{u}{2} - \frac{1}{u}} \right)^{j} \sinh \left( \frac{\frac{1}{u} - \frac{2}{u}}{\rho} \right) + 2 \frac{\sinh \left( \frac{\frac{1}{u} - \frac{2}{u}}{\rho} \right)}{\rho}. \tag{10}
\]

Having obtained the approximate mapping \( g(\cdot) \), we can utilize the results of (Bozic et al., 2016), and express the sought probability \( d_{c_j} \) as

\[
d_{c_j} = \Pr[\text{the frequency of cells that carry variant } c_j \text{ is at least } f] \approx \left( \frac{u}{u - \log(1 + (\rho - 1)f)} \right)^{g(c_j)}. \tag{11}
\]

Hence, the final expression of the expected \( B_{\text{md}}^{k} \), is specialized for this setting, as

\[
\mathbb{E}[B_{\text{md}}^{k}; K, f, \rho] = \sum_{j=1}^{K} \left[ \left( \frac{u}{u - \log(1 + (\rho - 1)f)} \right)^{g(c_{K-j})} - \left( \frac{u}{u - \log(1 + (\rho - 1)f)} \right)^{g(c_{K-(j+1)})} \right]. \tag{12}
\]

\( \mathbb{E}[B_{\text{md}}^{k}; K, f, \rho] \) and \( \text{Std}[B_{\text{md}}^{k}; K, f, \rho] \) over seeding times for different values of \( \rho \) and \( f \) can be found in the Supplementary Figures 4 and 5.

**7.1.3. On the effects of the introduction of advantageous types**

We study how the introduction of advantageous types affects the monotonicity of \( \{B_{\text{md}}^{k}\}_{k=0, \ldots, k_{\text{seed}}} \). Our approach is based on a straightforward generalization of the neutral-growth setting, and our analysis assumes phylogenetic independence between the passenger variants accumulated on the original type-1 cell, and is also conditioned on explicit
knowledge of the final population frequencies achieved by the different beneficial variants (in practice, the subclonal architecture of a tumor can be inferred from deep sequencing data (Miller et al., 2014; Nik-Zainal et al., 2012)).

For simplicity we consider the case where there are two types of cells in the population: (a) type-0 cells, which are characterized by death-to-birth ratio $\rho_0$ and are originated at time $T_0 = 0$ (i.e., type-0 is the type of the founder cell); and, (b) type-1 cells, which are characterized by death-to-birth ratio $\rho_1 < \rho_0$ and are originated in the population at time $T_1$, reaching a final frequency $f_1$ at the time of detection. When seeding happens at time $k$, the seeding cell originates from the type-1 cell population with probability $\phi_1^k$ and from the type-0 cell population with probability $\phi_0^k = 1 - \phi_1^k$. Of course, $\phi_1^k = 0$ for $k < T_1$, since there exist no type-1 cells in the population. Note that different choices of $(\phi_0^k)_{k=1,\ldots}, (\phi_1^k)_{k=1,\ldots}$ can capture different assumptions regarding the underlying seeding potential of the different types of cells that are present in the population. Here we consider three representative cases:

(a) **Seeding from the most advanced type.** When seeding happens before the introduction of the advanced type, the seeding cell is a type-0 cell, whereas when seeding happens after the introduction of type-1, the seeding cell is a type-1 cell. Concretely,

$$\begin{align*}
\phi_1^k &= \begin{cases} 0, & k < T_1 \\
1, & k \geq T_1 \end{cases}, & \phi_0^k &= 1 - \phi_1^k.
\end{align*}$$

Such approach models the assumption that disseminated and circulating tumor cells are associated with advanced stage cancers and the rate of shedding generally increases with tumor size (Alečković et al., 2019; Nguyen and Massagué, 2007).

(b) **Seeding proportional to fitness.** If $\lambda_0$ is the birth rate of type-0 cells and $\lambda_1$ the birth rate of type-1 cells,

$$\begin{align*}
\phi_1^k &= \begin{cases} 0, & k < T_1 \\
\frac{\lambda_1}{\lambda_0 + \lambda_1}, & k \geq T_1 \end{cases}, & \phi_0^k &= 1 - \phi_1^k.
\end{align*}$$

The coupling of dissemination and growth advantage is partly supported by the evidence that seeding cells more likely come from the tumor invasive front, where blood vessels are enriched (Kather et al., 2017).

(c) **Seeding uniformly over all living cells in the tumor.** In this approach the probability the seeding cell comes from a specific type is proportional to the number of cells in the population of such type. Concretely

$$\begin{align*}
\phi_1^k &= \begin{cases} 0, & k < T_1 \\
\frac{N_1}{N_0 + N_1}, & k \geq T_1 \end{cases}, & \phi_0^k &= 1 - \phi_1^k,
\end{align*}$$

where $N_0$, $N_1$ the population of cells of type 0 and 1, respectively, at seeding time. This scenario does not assume dissemination advantage of the new type, and the seeding is purely stochastic. This case is also considered here to account for the fact that few convincing recurrent genetic mutations have been identified so far that characteristically empower the metastatic dissemination in human cancers (Lambert et al., 2017).

We can now focus on $B_{md}^k$. Consider first the case where the seeding cell originates from the type-0 cell population. When $k < T_1$ this event occurs with probability 1, whereas when $k \geq T_1$ it occurs with probability $\phi_0^k$. Notice that while different cells in the population can accumulate different passenger variants as the type-0 population grows over time, all cells share the same fitness characteristics, i.e., their death-to-birth ratio is equal to $\rho_0$. Therefore, from a modeling perspective we fall within the setting analyzed in Section 7.1.2, and the expected $B_{md}^k$ can thus be given by (12) parametrized by $\rho = \rho_0$ and $f = \frac{2\alpha}{1-f_1}$. Similarly, if we condition on the seeding cell originating from type-1 we can calculate the expected $B_{md}$ based on (12) upon using the parametrization $\rho = \rho_1$, $f = \frac{2\alpha}{f_1}$ and $K = k_{\text{seed}} - T_1$. Therefore, a joint expression for approximating the expected $B_{md}$ in the case of two types is given by

$$\mathbb{E}[B_{md}^k; k, T_1, f_1] \approx \phi_1^k \times \mathbb{E}[B_{md}^{\text{neutral}}, k - T_1, \frac{2\alpha}{f_1}, \rho_1] + (1 - \phi_1^k) \times \mathbb{E}[B_{md}^{\text{neutral}}, k, \frac{2\alpha}{1-f_1}, \rho_0].$$

Note that the above approach can be extended to three or more types, in a straightforward way, and it thus omitted for the sake of brevity.
A fundamental implication of the above analysis is that the monotonicity of the sequence \( \{ \mathbb{E}[B_{md}; k] \}_{k=1}^{\infty} \) is expected to break by the introduction of newly emerging advantageous types. Such effect is, of course, inescapable when dissemination is assumed to be intertwined with fitness advantage (as in cases (a) and (b) described above), however the expected drop remains prominent even when no such assumption is enforced (Figure 6; top row, third column). Introduction of advantageous types affects \( B_p \) in a way that mirrors the related effects on \( B_{md} \). Notice that when new types lead to rapid drops in \( B_{md} \), \( B_p \) tracks these drops with abrupt decays (as seen in the bottom row of Figure 6), whereas when \( B_{md} \) increases linearly with seeding time, the corresponding \( B_p \) regions appear flat. This is in accordance with our discussion in Section 7.1.1, and, in particular, with equation (8). The variability due to the uncertainty in seeding origination, under scenarios (b) and (c), is increasing as new detectable types are introduced in the population—leading to the characteristic heteroscedastic patterns, also observed in Section 2.1. This is expected, as seeding from a new detectable type leads to a markedly smaller \( B_{md} \), compared to seeding from previous types that are also present in the population. Note that this variability arises from the seeding origination distribution over time \((\phi^k_0, \phi^k_1, \ldots)\) and is, of course, absent under seeding scenario (a) which is deterministic (Figure 6; first column). For additional plots reporting \( B_{md} \) and \( B_p \) patterns under different parametric choices, see Supplementary Figures 6 and 7.
7.2. Computational Modeling

The primary goal of the above analysis was to reason mathematically about the expected behavior of $B_{md}$ and $B_{p}$ across seeding times, as well as to derive example plots of such patterns for the case of analytically tractable growth models. Interestingly, such patterns remain true even under the more realistic setting, when tumor growth dynamics are governed by spatial 3D tumor structure (birth rate limitations imposed by local cell density etc). To study the behavior of the sought quantities at such more realistic settings we recur to our computational modeling framework, the details of which are presented below.

7.2.1. Birth death process on a three-dimensional lattice

We extend an existing single-cell-based spatial tumor growth model, named Tumopp (Iwasaki and Innan, 2017), to simulate paired primary and metastatic tumors. Tumopp provides a flexible yet computationally efficient framework for investigating various parameter settings for tumor growth, making it an ideal foundation for our implementation. We name our simulation framework Comet (COmputational Modeling of Evolving Tumors). In our model, tumor expansion starts with a single founder cell located in the center of a three dimensional lattice with the Moore neighborhood. Cells divide at rate $\lambda$ and die at rate $\mu$ per unit time. The waiting time between two consecutive division events is assumed to be exponentially distributed with mean $1/\lambda$. We set $\lambda = 0.25$ by assuming that the average cell cycle is four days. The death rate is chosen according to experimentally estimated death-to-birth ratios $\rho = \mu/\lambda$ in human cancers. For the slow growing scenario (e.g., early primary tumor), we set $\mu = 0.2475$ to achieve a $\rho = 0.99$ (Bozic et al., 2009), and for the fast growing scenario (e.g., metastatic tumor), we set $\mu = 0.18$ to achieve a $\rho = 0.72$ (Diaz et al., 2012). When a cell divides, one of its daughter cells stays at the original position on the grid, whereas the other fills one the neighboring empty sites uniformly at random. When a cell dies, it is removed from the lattice. The local density is taken into consideration such that the actual birth rate of the $j$-th cell is $\lambda_j = \lambda \times w_j$, where $w_j$ is the proportion of empty neighboring sites for this cell. This setting naturally reflects the spatial constraints in solid tumors and could lead to a pattern of peripheral growth consistent with experimental observations (Waclaw et al., 2015).

7.2.2. Simulating various tumor expansion kinetics and metastatic seeding

Comet allows passenger variants to occur at a rate of $u$ per cell birth. We randomly assign a virtual genomic coordinate for each variant by assuming that the variant resides in the broadly defined sequencing targetable region, such as the whole exome including UTRs ($5 \times 10^6$bp). Beneficial (or driver) variants which increase the birth rate by $s$, i.e., $\lambda'_{daughter} = \lambda'_{mother} \times (1 + s)$, occur at a rate $u_b$ per cell birth. The virtual genomic coordinate for a driver variant is sampled from a pool of $2 \times 10^6$bp in size. Although a smaller number of positions that confer fitness advantage have been assumed for point mutations (Reiter et al., 2018), here we assume that a variant captures any type of somatic alterations, including the breakpoints of structural and copy number variants, as long as they can be resolved at the single nucleotide resolution. These settings are all configurable and prepare our model to be comparable to other types of genomic divergence in the future. We employ various levels of the selection coefficient, in conjunction with the two levels of $\rho$, to model distinct growth dynamics. The selection coefficient is sampled from the Gaussian distribution with mean $s$ and standard deviation $s/2$. When $s = \{0, 0.02, 0.05\}$ and $\rho = 0.72$, the virtual tumor grows with a pattern of progressive diversification; $s = \{0.1, 0.2\}$ and $\rho = \{0.72, 0.99\}$ leads to a branching evolution and $s = \{0.5, 1\}$ and $\rho = 0.99$ realize the dynamics of linear evolution (Figure 4). Although the parameter setting is model specific, the resulting clonal dynamics may also be realized by other forces that are not modeled here, such as micro-environmental forces (Anderson et al., 2006) and purifying selection (Cvijović et al., 2018).

When the population size of the primary tumor increases by $N_I$ (25,000) cells ($N_{md} = I \times N_I$, where $I = 1, 2, \ldots$), we allow the next cell that is born, to disseminate and seed a metastatic tumor. This strategy takes into account both the impact of growth advantage and the existent population size of a subclone on metastatic dissemination potential. In other words, it falls between scenarios (b) seeding proportional to fitness and (c) seeding uniformly over all living cells in the tumor as discussed in the mathematical analysis in Section 7.1.3. Once a cell disseminates, it is removed from the lattice of the primary tumor, and is placed at the center of a new lattice devoted to the metastatic tumor. We apply small $\rho = 0.72$ to metastatic tumor growth (Diaz et al., 2012) and we let it inherit the rest of the parameter values of the corresponding primary tumor. We do not explore the dynamics of metastatic tumor expansion, as we focus on variants accrued in the seeding cell before dissemination, as well as variants in the primary tumor.
7.2.3. Lineage tracing and virtual multi-region sampling

Comet assigns a unique id to a cell when it is born, and tracks its location on the lattice, the id of its parent, the time of birth, the time when it divides or dies, as well as, the passenger and driver alterations (if any) occurred at cell’s birth. Such detailed information allows us to reconstruct the full genealogy map and all the variants in every cell after simulation. Both the virtual primary and metastatic tumors reach a final volume of \( N \) (\( 10^7 \)) cells. We use this number to strike a balance between computational efficiency in the single-cell-based setting, and practical relevance of the corresponding final physical size of the tumor (Dinh et al., 2020). Using this setting, we get a sequence of 400 virtual metastases as the primary tumor expands.

We randomly sample ten spherical regions (with a radius of 6.3 lattice units, leading to approximately 1,000 cells each) from the final primary and metastatic tumors, respectively. The variant repertoire of each region is reconstructed based on the genealogy relationship and variant history. A virtual sequencing experiment is performed to generate allele frequencies for each variant. The sequencing depth count is assumed to follow a negative binomial distribution

\[
D \sim \text{NegBinom}(m_d, \sigma_d), \quad \text{with} \quad m_d = 100, \quad \text{and} \quad \sigma_d = 10.
\]

The resulting read counts of the variant allele having a true allele frequency of \( p \) follow a binomial distribution

\[
M \sim \text{Binom}(D, F).
\]

7.2.4. Virtual sequencing data and tumor genealogy analysis

The allele frequency of a variant \( v \) in a region \( r \) is defined as \( f_v^r = M_v^r / D_v^r \). We evaluate the average between-region genomic divergence in the primary tumor based on the Fixation index (Fst) estimate (Sun et al., 2017b) and the Fst between region \( r \) and \( r' \) is computed as follows:

\[
\text{Fst}_{rr'} = \frac{\sum_{v=1}^{V} \left[ (f_v^r - f_v^{r'})^2 - \frac{f_v^r(1-f_v^r)}{D_v^r-1} - \frac{f_v^{r'}(1-f_v^{r'})}{D_v^{r'}-1} \right]}{\sum_{v=1}^{V} \left[ f_v^r(1-f_v^r) + f_v^{r'}(1-f_v^{r'}) \right]}.
\] (17)

The genetic variance components (numerator and denominator) are averaged separately to obtain a ratio combining the Hudson Fst estimate (Bhatia et al., 2013) across all \( V \) variants in the two regions. To obtain a global allele frequency of the variant \( v \) in the entire tumor, we calculate a pooled allele frequency across all the sampled regions for primary \( (f_v^p = \sum M_v^p / \sum D_v^p) \) and metastatic tumor \( (f_v^m) \), respectively (Hu et al., 2019). The number of metastatic specific variants is calculated from the virtual sequencing data as \( B_m = \# \{ v | f_v^m > \gamma, f_v^p < \alpha \} \). Primary specific variants are calculated as \( B_p = \# \{ v | f_v^p > \gamma, f_v^m < \alpha \} \). In this study, we use substantial presence threshold \( \gamma = 0.2 \) and a sequencing detectability threshold \( \alpha = 0.01 \) to demonstrate our findings.

The clonal evolutionary tree of a virtual tumor is being analyzed as a directed graph with vertices denoting cells and edges representing birth events. The branch length between two cells (e.g., a seeding cell and its most recent detectable ancestor at frequency \( \alpha \)) is calculated as the number of edges between them. The total branch length of a group of cells (e.g., all detectable ancestors at frequency \( \gamma \) in the primary tumor) is the sum of all edges in the sub-graph representing the genealogy of these cells.

8. Code Availability

Code for our single-cell based spatial tumor growth simulation studies and mathematical analysis is available at

https://github.com/SunPathLab/Comet
Modeling Metastatic Primary Genomic Divergence

http://genome.cshlp.org/content/24/11/1881.full; doi:10.1011/gr.180281.114.

Heyde, A., Reiter, J.G., Naxerova, K., Nowak, M.A., 2019. Consecutive seeding and transfer of genetic diversity in metastasis. Proceedings of the National Academy of Sciences of the United States of America 116, 14129–14137. URL: https://www.pnas.org/content/116/28/14129.abstract; doi:10.1073/pnas.1819408116.

Hosseini, H., Obradovic, M.M., Hoffmann, M., Harper, K.L., Sosa, M.S., Werner-Klein, M., Nanduri, L.K., Werno, C., Ehrl, C., Maneck, M., Patwary, N., Haunschild, G., Guzvic, M., Reimelt, C., Gruvogli, M., Eichner, N., Weber, F., Hartkopf, A.D., Taran, F.A., Brucker, S.Y., Fehm, T., Rack, B., Buechholz, S., Spang, R., Meister, G., Aguirre-Ghiso, J.A., Klein, C.A., 2016. Early dissemination seeds metastasis in breast cancer. Nature 540, 552–558. URL: https://www.nature.com/articles/nature20785; doi:10.1038/nature20785.

Hu, Z., Ding, J., Ma, Z., Sun, R., Seoane, J.A., Scott Shaffer, J., Suarez, C.J., Berghoff, A.S., Creemolini, C., Falcone, A., Loupakis, F., Birner, P., Preussner, M., Lenz, H.J., Curtis, C., 2019. Quantitative evidence for early metastatic seeding in colorectal cancer. Nature Genetics 51, 1113–1122. URL: https://www.nature.com/articles/s41588-019-0423-x; doi:10.1038/s41588-019-0423-x.

Iwasaki, W.M., Innan, H., 2017. Simulation framework for generating intratumor heterogeneity patterns in a cancer cell population. PLOS ONE 12, e0184229. URL: https://dx.plos.org/10.1371/journal.pone.0184229; doi:10.1371/journal.pone.0184229.

Jolly, C., Van Loo, P., 2018. Timing stochastic events in the evolution of cancer. Genome Biology 19, 95. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1476-3.

Jones, S., Chen, W.D., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M.A., Siegel, C., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Willis, J., Markowitz, S.D., 2008. Comparative lesion sequencing provides insights into tumor evolution. Proceedings of the National Academy of Sciences of the United States of America 105, 4283–4288. URL: https://pubmed.ncbi.nlm.nih.gov/18337506/; doi:10.1073/pnas.0712465105.

Kather, J.N., Zöllner, F.G., Schad, I.L.R., Melchers, S.M., Sinn, H.P., Marx, A., Gaiser, T., Weis, C.A., 2017. Identification of a characteristic vascular belt zone in human colorectal cancer. PLOS ONE 12, e0171378. URL: https://dx.plos.org/10.1371/journal.pone.0171378; doi:10.1371/journal.pone.0171378.

Lambert, A.W., Pattabiraman, D.R., Weinberg, R.A., 2017. Emerging Biological Principles of Metastasis. Cell 168, 670–691. doi:10.1016/j.stem.2016.11.037.

Magri, A., Bardelli, A., 2019. Does early metastatic seeding occur in colorectal cancer? Nature Reviews Gastroenterology and Hepatology 16, 651–653. doi:10.1038/s41575-019-0200-4.

Marusyk, A., Tabassum, D.P., Altrock, P.M., Almendro, V., Michor, F., Polyak, K., 2014. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature 514, 54–58. URL: http://www.nature.com/doifinder/10.1038/nature13556; doi:10.1038/nature13556; arXiv:1408.00003.

Miller, C.A., White, B.S., Dees, N.D., Griffith, M., Welch, J.S., Griffith, O.L., Vij, R., Tomasson, M.H., Graubert, T.A., Walter, M.J., Ellis, M.J., Schierding, W., DiPersio, J.F., Ley, T.J., Mardis, E.R., Wilson, R.K., Ding, L., 2014. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. PLoS Computational Biology 10, e1003665. URL: https://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4125065&tool=pmcentrez&rendertype=abstract.

Nguyen, D.X., Massagué, J., 2007. Genetic determinants of cancer metastasis. Nature Reviews Genetics 8, 341–352. URL: http://www.nature.com/doifinder/10.1038/nrg2101; doi:10.1038/nrg2101.

Nik-Zainal, S., Van Loo, P., Wedge, D., Alexandrov, L., Greenman, C., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S., Hinton, J., Menzies, A., Stebbing, L., Leroy, C., Jia, M., Rance, R., Mudie, L., Gamble, S., Stephens, P., McLaren, S., Tarpey, P., Papaemmanuil, E., Davies, H., Varela, I., McBride, D., Bignell, G., Leung, K., Butler, A., Teague, J., Martin, S., Jonsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerod, A., Aparicio, S., Tutt, A., Siewerts, A., Borg, Å., Thomas, G., Salomon, A., Richardson, A., Bortesen-Dale, A.L., Futreal, A., Stratton, M., Campbell, P., 2012. The life history of 21 breast cancers. Cell 149, 994–1007. URL: https://www.biorxiv.org/content/10.1101/586735v1.abstract.

Nowell, P., 1976. The clonal evolution of tumor cell populations. Science 194, 23–28. URL: 10.1126/science.aat7171.

Opacic, L., Zhou, D., Werner, B., Dingli, D., Traulsen, A., 2019. How many samples are needed to infer truly clonal mutations from heterogeneous tumors? bioRxiv, 600653URL: https://www.biorxiv.org/content/10.1101/600653v1; doi:10.1101/600653.

Pich, O., Muñoz, F., Lolkema, M.P., Steeghs, N., Gonzalez-Perez, A., Lopez-Bigas, N., 2019. The mutational footprints of cancer therapies. Nature Genetics 51, 1732–1740. URL: https://www.nature.com/articles/s41588-019-0525-5; doi:10.1038/s41588-019-0525-5.

Ramaswamy, S., Ross, K.N., Lander, E.S., Golub, T.R., 2003. A molecular signature of metastasis in primary solid tumors. Nature Genetics 33, 49–54. URL: https://www.nature.com/articles/ng1060; doi:10.1038/ng1060.

Reiter, J.G., Hung, W.T., Lee, I.H., Nagpal, S., Giunta, P., Degner, S., Liu, G., Wassenaar, E.C.E., Heck, W.R., Taylor, M.S., Farahani, A.A., Marble, H.D., Knott, S., Krausenberg, O., Lennerz, J.K., Naxerova, K., 2020. Lymph node metastases develop through a wider evolutionary bottleneck than distant metastases. Nature Genetics 1–9; doi:10.1038/s41588-020-0633-2.

Reiter, J.G., Makohon-Moore, A.P., Gerold, J.M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C.A., Vogelstein, B., Nowak, M.A., 2017. Reconstructing metastatic seeding patterns of human cancers. Nature Communications 8, 14114. URL: http://www.nature.com/doifinder/10.1038/ncomms14114; doi:10.1038/ncomms14114.

Reiter, J.G., Makohon-Moore, A.P., Gerold, J.M., Heyde, A., Attiyeh, M.A., Kohutek, Z.A., Tokheim, C.J., Brown, A., DeBlasio, R.M., Niyazov, J., Zucker, A., Karchin, R., Kinzler, K.W., Iacobuzio-Donahue, C.A., Vogelstein, B., Nowak, M.A., 2018. Minimal functional driver gene heterogeneity among untreated metastases. Science (New York, N.Y.) 361, 1033–1037. URL: http://www.ncbi.nlm.nih.gov/pubmed/30190408; doi:10.1126/science.aat7171.

RS, K., C., W., B., K., A., L., M., B., 1988. Clonal dominance of primary tumours by metastatic cells: genetic analysis and biological implications. Cancer Surveys 7, 597–629.
Modeling Metastatic Primary Genomic Divergence

Sansregret, L., Swanton, C., 2017. The Role of Aneuploidy in Cancer Evolution. Cold Spring Harbor Perspectives in Medicine 7, a028373. URL: http://www.ncbi.nlm.nih.gov/pubmed/28049656http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a028373, doi:10.1101/cshperspect.a028373.

Savas, P., Teo, Z.L., Lefevre, C., Flensburg, C., Caramia, F., Alsop, K., Mansour, M., Francis, P.A., Thorne, H.A., Silva, M.J., Kanu, N., Dietzen, M., Rowan, A., Kschisch, M., Fox, S., Bowtell, D.D., Dawson, S.J., Speed, T.P., Swanton, C., Loi, S., 2016. The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program “CASCADE”. PLoS Medicine 13, e1002204. URL: http://dx.plos.org/10.1371/journal.pmed.1002204, doi:10.1371/journal.pmed.1002204.

Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., Curtis, C., 2015. A Big Bang model of human colorectal tumor growth. Nature Genetics 47, 209–216. URL: http://www.nature.com/doifinder/10.1038/ng.3214, doi:10.1038/ng.3214.

Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C., Tavaré, S., 2013. Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. Cancer Research 73, 41–49. doi:10.1158/0008-5472.CAN-12-2273.

Sun, R., Hu, Z., Curtis, C., 2017a. Big Bang Tumor Growth and Clonal Evolution. Cold Spring Harbor Perspectives in Medicine, a028381URL: http://www.ncbi.nlm.nih.gov/pubmed/28710260http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a028381, doi:10.1101/cshperspect.a028381.

Sun, R., Hu, Z., Sottoriva, A., Graham, T.A., Harpak, A., Ma, Z., Fischer, J.M., Shibata, D., Curtis, C., 2017b. Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nature Genetics 49, 1015–1024. URL: http://www.nature.com/doifinder/10.1038/ng.3891, doi:10.1038/ng.3891.

Turajlic, S., Sottoriva, A., Graham, T., Swanton, C., 2019. Resolving genetic heterogeneity in cancer. Nature Reviews Genetics. 1 URL: http://www.nature.com/articles/s41576-019-0114-6, doi:10.1038/s41576-019-0114-6.

Turajlic, S., Swanton, C., 2016. Metastasis as an evolutionary process. Science 352, 169–175. doi:10.1126/science.aaf2784.

Uchi, R., Takahashi, Y., Niida, A., Shimamura, T., Hirata, H., Sugimachi, K., Sawada, G., Iwaya, T., Kurashige, J., Shinden, Y., Iguchi, T., Eguchi, H., Chiba, K., Shiraishi, Y., Nagae, G., Yoshida, K., Nagata, Y., Haeno, H., Yamamoto, H., Ishii, H., Doki, Y., Inuma, H., Sasaki, S., Nagayama, S., Yamada, K., Yachida, S., Kato, M., Shibata, T., Oki, E., Saeki, H., Shirabe, K., Oda, Y., Machara, Y., Komune, S., Mori, M., Suzuki, Y., Yamamoto, K., Aburatani, H., Ogawa, S., Miyano, S., Mimori, K., 2016. Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. PLoS Genetics 12, e1005778. URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005778, doi:10.1371/journal.pgen.1005778.

Waclaw, B., Bozic, I., Pitman, M.E., Hruban, R.H., Vogelstein, B., Nowak, M.A., 2015. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature 525, 261–264. URL: http://www.nature.com/articles/nature14971, doi:10.1038/nature14971.

Wakano, J.Y., Lehmann, L., 2014. Evolutionary branching in deme-structured populations. Journal of Theoretical Biology 351, 83–95. doi:10.1016/j.jtbi.2014.02.036.

Werner, B., Traulsen, A., Sottoriva, A., Dingli, D., 2017. Detecting truly clonal alterations from multi-region profiling of tumours. Scientific Reports 7, 44991. URL: http://www.nature.com/articles/srep44991, doi:10.1038/srep44991.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Iacobuzio-Donahue, C.A., 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467, 1114–1117. doi:10.1038/nature09515.

Yang, L., Wang, S., Lee, J.J.K., Lee, S., Lee, E., Shinbrot, E., Wheeler, D.A., Kucherlapati, R., Park, P.J., 2019. An enhanced genetic model of colorectal cancer progression history. Genome biology 20, 168. URL: http://www.ncbi.nlm.nih.gov/pubmed/31416464http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6694562, doi:10.1186/s13059-019-1782-4.

Yates, L.R., Knappskog, S., Wedge, D., Farmery, J.H., Gonzalez, S., Martincorena, I., Alexandrov, L.B., Van Loo, P., Haugland, H.K., Lilleg, P.K., Gundem, G., Gerstung, M., Pappaemmanuili, E., Gazinska, P., Bhosle, S.G., Jones, D., Raine, K., Mudie, L., Latimer, C., Sawyer, E., Desmedt, C., Sotiriou, C., Stratton, M.R., Sieuwerts, A.M., Lynch, A.G., Martens, J.W., Richardson, A.L., Tutt, A., Lønning, P.E., Campbell, P.J., 2017. Genomic Evolution of Breast Cancer Metastasis and Relapse. Cancer Cell 32, 169–184.e7. doi:10.1016/j.ccell.2017.07.005.
Graphical Abstract

Elements and Evolutionary Determinants of Genomic Divergence Between Paired Primary and Metastatic Tumors

Ruping Sun, Athanasios N. Nikolakopoulos
Highlights

Elements and Evolutionary Determinants of Genomic Divergence Between Paired Primary and Metastatic Tumors

Ruping Sun, Athanasios N. Nikolakopoulos

- Depth of most recent detectable variant characterizes Metastatic-Primary divergence
- Decay in variant detectability determines the expected M-P divergence
- Dissemination from late detectable subclone leads to an abrupt drop in M-P divergence
- Spatial model verifies growth mode governs M-P divergence dependency on seeding time