Supplementary Materials for

Minimal functional driver gene heterogeneity among untreated metastases

Johannes G. Reiter, Alvin P. Makohon-Moore, Jeffrey M. Gerold, Alexander Heyde, Marc A. Attiyeh, Zachary A. Kohutek, Collin J. Tokheim, Alexia Brown, Rayne M. DeBlasio, Juliana Niyazov, Amanda Zucker, Rachel Karchin, Kenneth W. Kinzler, Christine A. Iacobuzio-Donahue, Bert Vogelstein, Martin A. Nowak

correspondence to: johannes.reiter@stanford.edu, martin_nowak@harvard.edu

This PDF file includes:

Materials and Methods

Patient selection

We performed a comprehensive literature search to find sequencing samples from patients where (i) at least two untreated samples from distinct lymphatic or distant metastases were available and (ii) either whole-exome or whole-genome sequencing was performed on these samples. We focused on systemic metastases (lymphatic and/or distant metastases) due to the possibly distinct underlying biology and evolutionary processes of locoregional metastasis occurring in the same organ or tissue as the primary tumor.

We excluded sequencing data of various studies from this reanalysis because clinical data could not be obtained retrospectively, or index sequencing was performed on some of the metastases, precluding a comprehensive characterization of genetic inter-metastatic heterogeneity. In total, we found only nineteen subjects in eight publications meeting the above criteria (*13–19*, *30*). We based the reanalysis of previously published data on the original somatic variant calls to ensure an objective and accurate interpretation of the original studies' sequencing data. The original authors of seven of these studies shared their mutation calls (incl. chromosomal position, alternate allele, variant allele frequency, and sequencing depth of each mutation) with us so that we could reanalyze these data in a uniform fashion across cancer types. We excluded one ovarian cancer subject because the original variant calls could not be recovered (*30*). After including data from two unpublished subjects (described below), we reanalyzed 115 sequencing samples of twenty subjects (eight cancer types), including 76 untreated metastases samples from diverse tissues (table S1, fig. S1).

Sequencing data generation and processing

Whole-exome sequencing (WES) data for additional subjects MSKA1 and MSKA2 were generated according to the warm autopsy program of the Memorial Sloan Kettering Cancer Center. Both patients provided informed consent. This program complies with the Health Insurance Portability and Accountability Act and received approval from the Memorial Sloan Kettering Cancer Center institutional review board. Standard autopsy techniques were used to open the body cavity. Normal tissues and each grossly identified metastasis were dissected and removed. All human tissues were immediately frozen using liquid nitrogen and subsequently stored at −80 °C. For each tissue, one-half of the sample was fixed using 10% formalin while the other matching half was frozen at −80 °C for subsequent genetic analysis. Each metastasis underwent macrodissection to remove normal or necrotic tissue. Subsequently, each frozen metastatic sample was embedded in Tissue-Tek OCT, and a Leica Cryostat was used for sectioning. A 10-μm section was created for staining with hematoxylin and eosin slide for microscopic review of neoplastic cellularity.

Sequencing data analysis

We are very grateful to the authors of the original publications for enabling the reanalysis by sharing their data, in particular to David Brown, Sotiriou Christos, William

Gibson, Erling Hoivik, Marek Cmero, Chris Hovens, Tae-Min Kim, Sug-Hyung Lee, Adam Bass, and Matthew Stachler. Raw sequencing reads of the investigated data sets were originally aligned to the human reference genome hg19. For all seven reanalyzed studies (*13–19*), the original mutation calls were used. Intergenic and intronic variants of whole-exome sequencing (WES) data were excluded. Similarly, variants present in more than 0.1% of the normal exomes in the ExAC data set were removed. For all remaining variants, we calculated the presence and absence posterior probabilities based on the number of supporting alternate reads and the sequencing depth utilizing the Bayesian inference model of *Treeomics* (*20*). PyEnsembl (*31*) and VarCode (*32*) were used to distinguish the mutation consequences (e.g., missense, nonsense, etc.) and thereby identify possibly functional variants.

Data of two new subjects. We performed WES of 14 total samples taken from a patient with endometrial cancer (MSKA1) and a patient with lung cancer (MSKA2) at autopsy (see Pam13 and Pam16 in ref. (*17*) for processing and sequencing details). Somatic variants were called using MuTect. All 14 samples passed our selection criteria (figs. S31-S32).

Reanalysis of ref. (*18*). Brown et al. performed WES of 51 samples across ten autopsied breast cancer patients and performed somatic variant calling using the Genome Analysis Toolkit (GATK). Two patients (1/69 and 2/57) were diagnosed with *de novo* metastatic disease and died without any therapy. Three metastases in each of these subjects passed the selection criteria (figs. S25-S26).

Reanalysis of ref. (*16*). Gibson et al. performed WES of 98 tumor biopsies from 52 endometrial cancer patients and used MuTect for somatic variant calling. Five cases (EC-008, EC-015, EC-025, EC-030) passed the selection criteria (figs. S19-S22). Subject EC-012 was excluded because the tumors were clinically difficult to classify as metastases or independent synchronous primary cancers according to the original authors.

Reanalysis of ref. (*14*). Hong et al. performed whole-genome sequencing (WGS) of 26 samples across four prostate cancer patients and utilized both MuTect and Somatic Sniper for variant calling. Subject 498 with two untreated metastases samples from the sacral (MetSac) and iliac crest (MetIlCr) passed the selection criteria (fig. S29).

Reanalysis of ref. (*15*). Kim et al. performed WES of 35 samples across five colorectal cancer patients and used MuTect for variant calling. Subjects CRC1 with three liver metastases, CRC3 with six metastases, and CRC4 with four liver metastases passed the selection criteria (figs. S23-S25).

Reanalysis of ref. (*17*). Makohon-Moore et al. employed WGS of four and WES of two pancreatic cancers. All 49 samples were untreated and passed the selection criteria (figs. S13-S18). Somatic variants were called using MuTect.

Reanalysis of ref. (*19*). Pectasides et al. performed WES of primary tumor and paired metastasis samples prior to systemic therapy in a cohort of eleven gastric adenocarcinoma.

MuTect was used for variant calling. For one (C1-11) of these eleven subjects, two distinct metastases (bone and ascites) passed the selection criteria (fig. S30).

Reanalysis of ref. (*13*). Sanborn et al. performed WES of 27 samples across eight melanoma cancer patients and used Somatic Sniper and MuTect for variant calling. Somatic variants were extracted from their published Dataset SD3. Since we excluded locoregional metastasis samples from this analysis, only subject F with a cervical lymph node (MetLNC) and a back metastasis (MetSkin) passed the selection criteria (fig. S28).

Driver gene mutation definition

We annotated all nonsynonymous and splice-site variants in genes that were in the TCGA consensus driver list (*10*). Utilizing the driver gene list provided in Table S1 of ref. (*33*) leads to similar results (data not shown). We intentionally used expansive lists of putative driver genes to minimize the risk of underestimating driver mutation heterogeneity. Additionally, we determined whether a variant is reported in COSMIC (v84) (*34*), is a known cancer hotspot (*35*), or is annotated in OncoKB (*36*) (fig. S3).

Phylogenetic analysis and variant classification

We utilized two distinct approaches to classify somatic variants into MetTrunk (CancerTrunk and MetOrigin), MetBranch (MetShared and MetPrivate), and resectable (Primary Tumor: Shared and Private; see fig. S2). For both approaches, variants in the CancerTrunk were present in all samples of a subject while variants in MetOrigin were present in all metastases samples but not in all primary tumor (PT) samples. Variants in MetShared were present in a subset of metastases samples and variants in MetPrivate were only present in one metastasis. Variants in Shared were present in multiple samples of the PT and variants in Private were present in one sample of the PT (fig. S2). In the first classification approach, the presence and absence of mutations was based on phylogenies inferred by *Treeomics* which helped to recover sequencing artifacts due to low coverage or low neoplastic cell content. In the second classification approach, information in each sample was assessed independently using a Bayesian inference model for the number of alternate and reference reads in order to quantify the probability that a variant is present or not (*20*). Generally, the classifications differed only for very few variants. All results in this study are based on the variant classifications of *Treeomics* and can be reproduced using those of the Bayesian inference model (data not shown). Inferred phylogenies for all twenty subjects are shown in figs. S13-S32. We did not find any evidence for polyphyletic metastases in 19 of the 20 patients. In breast cancer subject 2-57, metastasis M3 might be polyphyletic (seeded by distinct subclones) (*20*, *37*). Nevertheless, the identified subclones shared the same putative driver gene mutations (fig. S27).

Functional predictions of mutations in putative driver genes

We applied several prediction algorithms to assess the difference in functional impact of mutations. In addition to *VEP* (*38*), *FATHMM* (*21*), and *CHASMplus* (*22*, *39*), we also used *CanDrA* (*40*), *PolyPhen-2* (*41*), and *SIFT* (*42*) to predict the functional consequences of nonsynonymous variants (fig. S3). While we obtained similar results with *CanDrA*, the predictions of *PolyPhen-2* and *SIFT* did not reach statistical significance, perhaps because these methods aim to predict the damaging effects of germline and not somatic mutations.

Stochastic computer simulations

We utilize continuous-time, multi-type branching processes to model the evolutionary dynamics of primary tumors and their metastases (see section *Mathematical Modeling* for more details and analytical solutions) (*23*, *24*, *43–45*). We assume that the primary tumor grows from a single type $i = 0$ cell and seeds metastases until m metastases reach detection size. Type 0 cells divide with rate b_0 per day and die with rate d_0 per day, so that their net growth rate is $r_0 = b_0 - d_0$, and they disseminate to seed new metastases with rate q_0 per day (Fig. 4A). Whenever a cell divides, one of the daughter cells can acquire an additional driver gene mutation with rate $u = k \cdot \hat{u}$ where k is the number of distinct driver gene mutations across the exome and $\hat{u}$ is the point mutation rate per cell division. Cells of the i^{th} mutant clone divide at rate b_i, die at rate d_i, and seed a new metastasis at rate q_i. The selective growth advantage of the i^{th} clone is defined as $s_i = b_i/b_0 - 1$. The selective growth advantage may depend on the microenvironment and might only be conferred in the primary tumor or at a distant site. Generally, the mathematical framework allows us to investigate many potential effects of driver mutations (fig. S6-S11). For simplicity, we assumed that all subclones $i > 0$ grow with the same rates. We considered the first m metastases that reached detection size of 10^8 cells (~1 cm^3) to be heterogeneous if they contained some pair of metastases that were founded by cells of different clones (*46*). To reduce the runtime of individual simulations, we simulated the random occurrence of additional driver gene mutations and the random dissemination of cells through sampling from the corresponding distributions (*24*).

Various other measures of inter-metastatic heterogeneity as the probability that m detectable metastases are seeded by the same subclone are conceivable (*47*): for example, the Simpson Index (*48*) (probability that the founding cells of two random metastases share the same driver gene mutations), the Shannon Index (*49*) (driver gene mutation prediction uncertainty), the fraction of metastases that are seeded by driver subclones, or the number of distinct subclones that seeded any detectable metastasis. Qualitatively, all these measures show the same results (fig. S11).

Parameter selection

Cancer cells of various types (including colorectal and pancreatic ductal adenocarcinoma) have been estimated to divide roughly every ~4 days (*50*, *51*), leading to a division rate of 0.25 per day. To explore a wide range previously observed growth rates of primary tumors and metastases (*26*, *52*), we assumed a fixed death rate of $d = 0.2475$ and varied the birth rate. Following previous approaches (*25*, *51*, *53*), we explored death-birth rate ratios from 0.99 to 0.62 corresponding to growth rates of 0.25% and 15.25% per day, respectively. We considered a wide range of selective growth advantages s of driver gene mutations up to 5%; more than a magnitude higher than the previously estimated average selective advantage of 0.4% (*25*). The number of mutated positions in the exome that could confer such a growth advantage was estimated as $k = 34{,}000$. Assuming a point mutation rate (*50*, *54*) per cell division of $\hat{u} = 10^{-9}$, we obtain an effective driver gene mutation rate of $u = k \cdot \hat{u} = 3.4 \cdot 10^{-5}$ per cell division. A very similar estimate of $6.3 \cdot 10^{-5}$ has been inferred by Haeno *et al.* (*23*) for the mutation rate leading to a migratory (metastatic) phenotype. We use these estimates as a starting point but also explore lower mutation rates to resemble scenarios where multiple pathways have already been activated or higher mutation rates to resemble cancer types with a high mutational burden (e.g., lung cancer, melanoma). Much less is known about the migratory potential of individual cells of different cancer types. For pancreatic cancer, Haeno et al. (*23*) inferred a dissemination rate of $q = 6.3 \cdot 10^{-7}$ per cell per day. We examine dissemination rates from $q = 10^{-10}$ up to $q = 10^{-4}$ per cell per day.

Mathematical Modeling

We model the evolutionary dynamics of a primary tumor and its metastases as a continuous time, multi-type branching processes (*23, 24, 43–45*). We consider a primary tumor that grows from a single advanced cancer cell of type 0 and seeds secondary tumors (metastases) until M^* secondary tumors are detected. Type 0 cells divide with rate b_0 per day, die with rate d_0 per day, and disseminate to new secondary sites with rate q_0 per day. Whenever a type 0 cell divides, a daughter cell can acquire an additional driver gene mutation with rate $u = k \cdot \hat{u}$, where k is the number of estimated driver gene mutation positions across the exome and $\hat{u}$ is the point mutation rate per base-pair per cell division. In this event, the mutant daughter cell then becomes the first cell of type i, where $i = 1, 2, \ldots$ denotes the ordering of driver mutation appearance times. Cells of type i divide with rate b_i per day, die with rate d_i per day, and disseminate to a new secondary site with rate q_i per day. The selective growth advantage of the ith clone is defined as $s_i = \frac{b_i}{b_0} - 1$. We assume an infinite sites model such that the same mutation is not independently acquired twice (*55*), and we assume that these already advanced cancer cells can acquire at most one additional driver gene mutation. Since we focus on inter-metastatic heterogeneity (differences between the founding cells of metastases), we ignore additional driver mutations acquired within metastases during their growth phase (intra-metastatic heterogeneity) (*12*).

For times $t \geq 0$, let there be $N(t)$ distinct surviving mutant clones (subpopulations), each with $X_i(t)$ cells in the primary tumor and $M_i(t)$ secondary tumors founded, where $i = 0$ represents the original type and $i = 1, 2, \ldots, N(t)$ represent the mutant clones. We initialize the system at time $t = 0$ with a single type 0 cell in the primary tumor and with no metastases seeded, such that $N(0) = 0$, $X_0(0) = 1$, and $M_0(0) = 0$. Each clone i is assumed to grow with positive net rate $r_i = b_i - d_i > 0$ and hence have a nonzero survival probability of $\rho_i = r_i/b_i > 0$. As in the Luria-Delbruck model (*56*), we treat birth and death as deterministic, though conditioned on survival, but the generation of driver gene mutations and the seeding of metastases as stochastic. We consider the first M^* detected metastases to be heterogeneous if there is some pair of metastases that were founded by cells of different clones.

Mathematical analysis

The mean size of the original clone type $i = 0$ follows a simple exponential growth law $e^{r_0 t}$. The probability that this population has survived until time t is given by $P_0^+(t) = \rho_0/[1 - (1 - \rho_0)e^{-r_0 t}]$, and hence the long-term survival probability is $P_0^+(\infty) = \rho_0$ (*24*). Since we consider only patients with a primary tumor of nonzero size in the long run, we normalize the mean size of a tumor to be conditional on its long-term survival, obtaining

$$\bar{X}_0(t) = \frac{1}{\rho_0} e^{r_0 t} - \left(\frac{1}{\rho_0} - 1\right) e^{-r_0 t}, \tag{1}$$

as derived in Section *Deriving the conditional mean growth law*. This survival-conditioned mean growth law satisfies the desired initialization $\bar{X}_0(0) = 1$ and converges to the asymptotic scaling $\bar{X}_0(t) \to \frac{1}{\rho_0} e^{r_0 t}$ as the first term grows to dominate the second in magnitude.

This asymptotic scaling can serve as a reasonable approximation for the growth law, since the first term is always greater than the second. Moreover, the relative error of this approximation decays as $(1-\rho_0)e^{-2r_0 t}$ and rapidly becomes less than e^{-1} for times $t > \frac{1-\rho_0}{2r_0}$.

The number of surviving tumors M_0 seeded by type 0 cells before time t follows an inhomogeneous Poisson process with time-varying mean $\bar{M}_0(t)$. In other words, the probability mass function for $M_0(t)$ is

$$P(M_0(t) = m_0) = \frac{1}{m_0!}\,\bar{M}_0(t)^{m_0}\,e^{-\bar{M}_0(t)}. \tag{2}$$

To calculate the mean number of seeded surviving secondary tumors $\bar{M}_0(t)$, we integrate the number of type 0 cells over time and multiply by the dissemination rate, q_0, and the survival probability of a newly founded clone, ρ_0. We obtain

$$\bar{M}_0(t) = \int_0^t q_0\,\rho_0\,\bar{X}_0(\tau)\,d\tau = \frac{q_0}{r_0}\Big(e^{r_0 t} - 1\Big)\Big(1 - (1-\rho_0)e^{-r_0 t}\Big) \to \frac{q_0}{r_0}\,e^{r_0 t}. \tag{3}$$

where the arrow denotes convergence to the dominant term over time. The time T_0^1 at which a type 0 cell seeds the first surviving secondary tumor (that is, the earliest time at which $M_0(t) = 1$) then follows an inhomogeneous exponential distribution with density function

$$f_{T_0^1}(t) = \bar{M}_0'(t)\,e^{-\bar{M}_0(t)} \to q_0 \exp\left\{r_0 t - \frac{q_0}{r_0}e^{r_0 t}\right\}. \tag{4}$$

The mean $\bar{T}_0^1$ of this distribution can be estimated by an expansion about small q_0 to obtain

$$\bar{T}_0^1 = \int_0^\infty t f_{T_0^1}(t)\,dt \approx \frac{\log\frac{r_0}{q_0} - \gamma}{r_0} \tag{5}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant (Fig. 4C, fig. S5). For example, for a dissemination rate of $q = 10^{-7}$ per cell per day and a net growth rate of $r_0 = 1.3026\%$ per day, we find by numerical integration that the first surviving metastasis of the original clone is seeded after 2.35567 (± 0.26959) years (mean ± one standard deviation) (fig. S5). Eq. (5) provides an excellent approximation with an estimate of 2.35563 years. Several examples of mean times are included in the table below.

	Appearance	**First metastasis**	**Second metastasis**
Clone	**time**	**seeding time**	**seeding time**
Original	0	2.36 (±0.27)	2.57 (±0.17)
Mutant 1	1.40 (±0.26)	3.60 (±0.36)	3.79 (±0.31)
Mutant 2	1.61 (±0.17)	3.80 (±0.30)	4.00 (±0.23)
Mutant 3	1.71 (±0.13)	3.91 (±0.28)	4.10 (±0.20)

Numerical examples for the mean appearance times of driver subclones and metastases seeding times of surviving lineages in years (± one standard deviation). Parameter values: death-birth-rate ratio $d/b_0 = 0.95$, cell death rate per day $d = 0.2475$, relative driver advantage $s = 0.4\%$, dissemination rate per cell per day $q = 10^{-7}$, driver mutation rate per cell division $u = 3.4 \cdot 10^{-5}$.

Similarly, the number of surviving clones $N(t)$ in the primary tumor follows an inhomogeneous Poisson process with mean $\bar{N}(t)$ and probability mass function

$$P(N(t) = n) = \frac{1}{n!}\,\bar{N}(t)^n\,e^{-\bar{N}(t)}. \tag{6}$$

To calculate the mean $\bar{N}(t)$, we integrate the number of type 0 cells over time and multiply by the division rate b_0, the driver mutation rate u, and the probability that a new lineage survives ρ_1, which we assume to be equal across all mutant lineages. This gives

$$\bar{N}(t) = \int_0^t u\,b_0\,\rho_1\,X_0(\tau)\,d\tau = u\,\frac{\rho_1}{{\rho_0}^2}\Big(e^{r_0 t} - 1\Big)\Big(1 - (1-\rho_0)e^{-r_0 t}\Big) \to u\,\frac{\rho_1}{{\rho_0}^2}\,e^{r_0 t}. \tag{7}$$

The time T_1 at which the first surviving mutant clone is originated has the density function

$$f_{T_1}(t) = \bar{N}'(t)\,e^{-\bar{N}(t)} \to u\,r_0\,\frac{\rho_1}{{\rho_0}^2}\,\exp\left\{r_0 t - u\,\frac{\rho_1}{{\rho_0}^2}\,e^{r_0 t}\right\}. \tag{8}$$

The mean $\bar{T}_1$ of this distribution can be estimated by an expansion about small u to obtain

$$\bar{T}_1 = \int_0^\infty t f_{T_1}(t)\,dt \approx \frac{\log\frac{{\rho_0}^2}{u\,\rho_1} - \gamma}{r_0}\,. \tag{9}$$

which is equivalent to the result derived by (*24*). Generalizing this to subsequent clones, the time T_i at which each clone $i \geq 1$ is originated has the probability density function

$$f_{T_i}(t) = \frac{\bar{N}(t)^{i-1}}{(i-1)!}\,e^{-\bar{N}(t)}\bar{N}'(t). \tag{10}$$

The mean number of cells $\bar{X}_i$ of type i in the primary tumor must be 0 until time T_i, after which it follows the same survival-conditioned mean growth law as the type 0 cells:

$$\left(\bar{X}_i(t)\middle|T_i\right) = \begin{cases} \frac{1}{\rho_i}\,e^{r_i(t-T_i)} - \Big(\frac{1}{\rho_i} - 1\Big)e^{-r_i(t-T_i)}, & t \geq T_i, \\ 0, & t < T_i, \end{cases} \tag{11}$$

which we can again approximate as $(\bar{X}_i(t)\big|T_i) \to \frac{1}{\rho_i}\,e^{r_i(t-T_i)}$ for the case $t \geq T_i$. The probability that $\bar{X}_i(t)$ is exactly zero is then $P_i^0(t) = P(t < T_i) = \int_t^\infty f_{T_i}(\tau)\,d\tau$. Otherwise, the probability density of $\bar{X}_i(t)$ is given by applying the change of variables $X_i(t) = \frac{1}{\rho_i}\,e^{r_i(t-T_i)}$ to the probability density function of T_i, to obtain

$$f_{\bar{X}_i(t)}(x_i) = \frac{1}{r_i x_i}\,f_{T_i}\Big(t - \frac{1}{r_i}\log(\rho_i x_i)\Big). \tag{12}$$

The number of surviving secondary tumors $M_i(t)$ seeded by each clone $i \geq 1$ follows an inhomogeneous Poisson process with mean $\bar{M}_i(t)$ and probability mass function

$$P(M_i(t) = m_i) = \frac{1}{m_i!}\,\bar{M}_i(t)^{m_i}\,e^{-\bar{M}_i(t)}. \tag{13}$$

To calculate $\bar{M}_i(t)$, we integrate the number of type i cells over time and scale by the dissemination rate, q_i, and the survival probability of a newly founded secondary tumor, ρ_i. We also note that $\bar{M}_i(t) = 0$ if $t < T_i$, so we have the piecewise relationship

$$\left(\bar{M}_i(t)|T_i\right) = \begin{cases} \int_{T_i}^{t} q_i\, \rho_i \left(X_i(\tau)\middle|T_i\right) d\tau \rightarrow \frac{q_i}{r_i}\left(e^{r_i(t-T_i)} - 1\right), & t \geq T_i, \\ 0, & t < T_i. \end{cases} \tag{14}$$

Hence, $\bar{M}_i(t)$ is exactly zero with probability $P_i^0(t)$ and otherwise has the density function

$$f_{\bar{M}_i(t)}(\lambda_i) = \frac{1}{q_i + r_i\lambda_i}\, f_{T_i}\left(t - \frac{1}{r_i}\log\left[1 + \frac{r_i\lambda_i}{q_i}\right]\right),\ \lambda_i > 0. \tag{15}$$

where λ_i, the argument of the density function, denotes a possible value of $\bar{M}_i$. The mean time $\bar{T}_1^1$ at which the first surviving mutant clone seeds its first surviving metastasis is obtained by summing the mean time $\bar{T}_1$ at which the first mutant clone is originated with the mean time interval until a cell of that clone metastasizes (fig. S5), which gives

$$\bar{T}_1^1 = \frac{\log\frac{{\rho_0}^2}{u\,\rho_1} - \gamma}{r_0} + \frac{\log\frac{r_1}{q_1} - \gamma}{r_1}. \tag{16}$$

For example, for the parameter values used in the table above, we find that the first surviving metastasis of the first surviving mutant clone is seeded after 3.595 (± 0.364) years (mean ± one standard deviation) (fig. S5). Eq. (16) provides an excellent approximation with an estimate of 3.593 years.

Combining the above results gives an integral expression for the marginal probability mass function for the number of surviving secondary tumors M_i seeded by clone $i \geq 1$,

$$\begin{aligned} P(M_i(t) = m_i) = &\int_0^\infty \frac{\lambda_i^{m_i}\, e^{-\lambda_i}}{m_i!}\, \frac{r_0\, \tilde{u}\, \gamma_i(t)^{i-1}\, e^{r_0 t - \gamma_i(t)}}{q_i (i-1)!} \left(1 + \frac{\lambda_i\, r_i}{q_i}\right)^{-\frac{r_0}{r_i} - 1} d\lambda_i \\ &+ \int_t^\infty \delta[m_i = 0]\, \frac{r_0\, \tilde{u}\, \bar{N}(\tau)^{i-1}\, e^{r_0 \tau - \bar{N}(\tau)}}{(i-1)!}\, d\tau, \\ &\text{where } \gamma_i(t) = \tilde{u}\left[\left(1 + \tfrac{\lambda_i\, r_i}{q_i}\right)^{-\frac{r_0}{r_i}} e^{r_0 t} - 1\right], \end{aligned} \tag{17}$$

and where $\tilde{u} = u \cdot \rho_1/{\rho_0}^2$ is a scaled mutation probability and δ is the Kronecker delta. We can approximate the integration via Laplace's method (Section *Integration by Laplace's method*), where the approximation rapidly converges to the exact result as time grows such that any transient finite effects become negligible, to obtain the probability mass function

$$\begin{aligned} P(M_i(t) = m_i) \approx &\left[\frac{\Lambda_i(t)^{m_i + \frac{1}{2}}}{m_i!} \sqrt{\frac{2\pi}{1 + \frac{r_0}{r_i}}} \exp\left\{(i-1)r_0 t - (1 + \tfrac{r_i}{r_0})\Lambda_i(t)\right\} \left(\tfrac{r_i}{q_i}\Lambda_i(t)\right)^{-\frac{r_0}{r_i}(i-1)} \right. \\ &\left. + \,\delta[m_1 = 0]\, e^{-\bar{N}(t)}\right] \frac{\tilde{u}^{i-1}}{(i-1)!}, \quad \text{where } \Lambda_i(t) = \frac{q_i}{r_i} \sqrt{\frac{\tilde{u}\, r_0}{q_i} e^{r_0 t}}^{\frac{\bar{r}_i}{r_0}}. \end{aligned} \tag{18}$$

Here $\Lambda_i(t)$ is the characteristic metastasis count that maximizes the first integrand, with $\bar{r}_i$ denoting the harmonic mean of the two net proliferation rates r_0 and r_i, and $\bar{N}(t) = \tilde{u}\, e^{r_0 t}$ is the approximate mean number of clones at long times. Using a superscript * to mark quantities evaluated at detection time $T^* = \frac{1}{r_0}\log(\frac{r_0}{q_0}M^*)$, defined approximately such that $\bar{M}_0^* = M^*$, we compute

$$\Lambda_i^* = \tfrac{q_i}{r_i}\sqrt{\tfrac{\tilde{u}\, r_0^2}{q_0\, q_i}M^*}^{\frac{\bar{r}_i}{r_0}} \text{ and } \bar{N}^* = \bar{N}(\hat{T}) = \tfrac{\tilde{u}\, r_0}{q_0}M^*. \tag{19}$$

To satisfy the validity conditions of our Laplace approximation, we require that $M^* \gg \frac{q_0\, q_i}{\tilde{u}\, r_0{}^2}$ such that $\Lambda_i^* \gg \frac{q_i}{r_i}$ and $\bar{N}^* \gg \frac{q_i}{r_0}$ at detection time. For sufficiently large M^*, the second term of Eq. (18) may also be neglected; then the probability P_{het} that the first clone seeds some nonzero number of metastases before detection can be simplified to give

$$P_{\text{het}} = P(M_1^* > 0) \approx 1 - e^{-(1+\frac{r_1}{r_0})\Lambda_1^*}\sqrt{\pi\, \Lambda_1^* \, \tfrac{\bar{r}_1}{r_0}}\,, \tag{20}$$

This quantity, a measure of the tendency of a cancer to lead to metastases of heterogeneous clonal origin, increases monotonically in the characteristic metastasis count Λ_i^* throughout the range of validity $r_1\Lambda_1^* > \frac{1}{4}\bar{r}_1 \gg q_1$. We explore how this quantity transitions between a regime of homogeneity to one of heterogeneity in Section *Critical selection advantage*.

Deriving the conditional mean growth law. To obtain Eq. (1), which gives the number of type 0 cells at time t conditional on the long-term survival of the type 0 subpopulation, we first define a related quantity. The probability that the type 0 subpopulation goes extinct before time t, conditional on its long-term extinction, is

$$P_0^0(t) = \frac{1 - P_0^+(t)}{1 - \rho_0}. \tag{21}$$

Then the probability that the population size X_0 is exactly x cells at time t is given by

$$P(X_0(t) = x) = \begin{cases} 1 - P_0^+(t), & x = 0, \\ P_0^+(t)\big[1 - P_0^0(t)\big]P_0^0(t)^{x-1}, & x \geq 1, \end{cases} \tag{22}$$

equivalent to the result of (*24*). Using Bayes' rule, we can then condition this probability mass function on the long-term survival of the subpopulation, $X_0(\infty) > 0$, to obtain

$$\begin{aligned} P(X_0(t) = x | X_0(\infty) > 0) &= \frac{P(X_0(\infty) > 0 | X_0(t) = x)}{P(X_0(\infty) > 0)}P(X_0(t) = x) && (23) \\ &= \frac{1 - (1 - \rho_0)^x}{\rho_0}P(X_0(t) = x), && (24) \end{aligned}$$

and this quantity is 0 when evaluated as $x = 0$, as is required. The mean population size conditioned on its long-term survival can then be calculated as

$$\bar{X}_0(t) = \sum_{x=1}^{\infty} x\, P(X_0(t) = x | X_0(\infty) > 0) = \frac{1}{\rho_0}e^{r_0 t} - \left(\frac{1}{\rho_0} - 1\right)e^{-r_0 t}, \tag{25}$$

which is precisely the same result as Eq (1). Equivalently, one can express this result more compactly as $\bar{X}_0(t) = e^{r_0 t}/P_0^+(2t)$.

Joint dynamics of all clonal populations. The model analysis can be modified to consider the dynamics of all clones jointly. The joint density function of origination times T_i for the first I clones ($i = 1, \ldots, I$) is

$$f_T(t_1, \ldots, t_I) = e^{-\bar{M}(t_I)} \prod_{i=1}^{I} \bar{M}'(t_i) = (\tilde{u} r_0)^I \exp\Big\{ - \tilde{u}(e^{r_0 t_I} - 1) + r_0 \sum_{i=1}^{I} t_i \Big\} \tag{26}$$

for $t_1 \leq \cdots \leq t_I$. Hence the joint density function of mutant clone population sizes is

$$f_X(x_1, \ldots, x_I) = f_T\Big(t - \frac{\log x_1}{r_1}, \ldots, t - \frac{\log x_I}{r_I}\Big) \prod_{i=1}^{I} \frac{1}{r_i x_i} \tag{27}$$

for $x_1 \geq \cdots \geq x_I$, and the joint density function of the vector Λ of means $\bar{M}_i$ is

$$f_\Lambda(\lambda_1, \ldots, \lambda_I) = f_T\Big(t - \frac{\log(1 + \frac{r_1 \lambda_1}{q_1})}{r_1}, \ldots, t - \frac{\log(1 + \frac{r_I \lambda_I}{q_I})}{r_I}\Big) \prod_{i=1}^{I} \frac{1}{q_i + r_i \lambda_i} \tag{28}$$

over the support $S_\Lambda = \{\Lambda : \lambda_1 \geq \cdots \geq \lambda_I > 0\}$. Integration over this support gives the joint mass function of the number of secondary tumors M_i seeded by each clone $i = 0, \ldots, I$:

$$P(M(t) = m) = \frac{[\frac{q_0}{r_0}(e^{r_0 t} - 1)]^{m_0}}{m_0!\, e^{\frac{q_0}{r_0}(e^{r_0 t} - 1)}} \int_{S_\Lambda} e^{-\gamma_I(t)} \prod_{i=1}^{I} \frac{\frac{u\, r_0}{q_i} e^{r_0 t - \lambda_i} \lambda_i^{m_i}}{m_i!\, (1 + \frac{r_i\, \lambda_i}{q_i})^{1 + r_0/r_i}}\, d\lambda_i. \tag{29}$$

This integral in general cannot be analytically evaluated over S_Λ due to the requirement that the rate parameters λ_i be ordered, but it nonetheless can be useful for numerical calculations.

Integration by Laplace's method. To integrate Eq. (17), we use Laplace's property that in the limit of some large parameter $K \to \infty$, the following integral converges under generic assumptions:

$$\int_a^b g(\lambda) e^{-K f(\lambda)} d\lambda \to \sqrt{\frac{2\pi}{K |f''(\Lambda)|}}\, g(\Lambda)\, e^{-K f(\Lambda)}, \tag{30}$$

where Λ denotes the argument of the global minimum of $f(\lambda)$ over the domain $a < \lambda_0 < b$, where a, b are finite or infinite bounds of integration. To evaluate the first integral, we define

$$g(\lambda_i) = \frac{\lambda_i^{m_i}}{m_i!} \frac{\tilde{u}^i}{(i-1)!} \frac{r_0}{q_i} \left(1 + \frac{r_i}{q_i} \lambda_i\right)^{-i \frac{r_0}{r_i} - 1} \tag{31}$$

$$f(\lambda_i) = \Big(\lambda_i - i\, r_0\, t\Big) e^{-r_0 t} + \tilde{u} \left(1 + \frac{r_i}{q_i} \lambda_i\right)^{-\frac{r_0}{r_i}} \tag{32}$$

and we set $K = e^{r_0 t}$ so that the integral converges to the Laplace approximation exponentially in time. Specifically, if $K \gg \frac{q_i}{\tilde{u}\, r_0}$ (which occurs for $t \gg \frac{1}{r_0} \log \frac{q_i}{\tilde{u}\, r_0}$) then $\bar{N}'(t) = \tilde{u}\, r_0\, e^{r_0 t} \gg q_i$. In this regime, the global minimum of $f(\lambda_i)$ is obtained at a value Λ_i that satisfies $r_i \Lambda_i \gg q_i$:

$$\Lambda_i = \frac{q_i}{r_i} \left(\frac{\tilde{u}\, r_0}{q_i} e^{r_0 t} \right)^{\frac{r_i}{r_0 + r_i}} = \frac{q_i}{r_i} \sqrt[\frac{\bar{r}_i}{r_0}]{\frac{\tilde{u}\, r_0}{q_i} e^{r_0 t}} \tag{33}$$

At this value, f achieves the minimum value $f(\Lambda_i) = \left(1 + \frac{r_i}{r_0}\right) \Lambda_i e^{-r_0 t} - i r_0 t e^{-r_0 t}$, the first derivative $f'(\Lambda_i) = 0$, and the second derivative $f''(\Lambda_i) = \left(1 + \frac{r_0}{r_i}\right) \Lambda_i^{-1} e^{-r_0 t}$. We find that

$$g(\Lambda_i) = \frac{\Lambda_i^{m_i}}{m_i!} \frac{\tilde{u}^{i-1}}{(i-1)!}\, e^{-r_0 t} \left(\frac{r_i}{q_i} \Lambda_i \right)^{-\frac{r_0}{r_i}(i-1)} \quad . \tag{34}$$

Substituting these expressions into Laplace's approximation Eq. (30) gives the result Eq. (18).

Critical selective advantage. To approximate the boundaries dividing the regimes of inter-meta-stastic heterogeneity from inter-metastatic homogeneity, we analyze the probability that the first seeded $\hat{M}$ metastases were all seeded by type 0 cells or all by type 1 cells. For simplicity, we ignore here distinct driver subclones and do not account for the growth phase of metastases. We analyze the behavior of Eq. (20) under three cases: first, when a driver gene mutation confers an advantage in dissemination but not growth ($q_1 \geq q_0$, $b_1 = b_0$; Fig. 4F); second, when a driver gene mutation confers an advantage in growth but not dissemination ($b_1 \geq b_0$, $q_1 = q_0$; Fig. 4, B-E); and third, when a driver gene mutation confers a mixed advantage in both growth and dissemination ($b_1 \geq b_0$, $q_1 \geq q_0$).

Dissemination advantage. In the case of a dissemination advantage, we have $\frac{r_1}{r_0} = 1$ and $\frac{q_1}{q_0} = Q$ for some constant ratio $Q > 1$. Then $\bar{r}_1 = r_0$, and the scaled mutation probability is $\tilde{u} = \frac{u}{\rho_0}$. Substituting these relationships into Eq. (20) immediately gives that the probability P_{het} that some of the first M^* metastases were seeded by type 1 cells,

$$P_{\text{het}}^d \approx 1 - e^{-2\Lambda_1^*} \sqrt{\pi\, \Lambda_1^d}\ , \text{ where } \Lambda_1^d = \sqrt{\frac{Q\, u\, M^*}{\rho_0}}. \tag{35}$$

where the superscript d indicates the scenario of driver advantages in dissemination only. Solving $\frac{\partial P_{\text{het}}^d}{\partial Q} = 0$ for Q gives an estimate for the threshold seeding advantage $\hat{Q}$,

$$\hat{Q} \approx \frac{\rho_0}{16\, u\, M^*}, \quad \text{or equivalently,} \quad \hat{\Lambda}_1^d \approx \frac{1}{4}\ . \tag{36}$$

This simple threshold $\hat{Q}$ is plotted in Fig. 4F and fig. S4C and divides the two regimes in the case of driver mutations conferring an increased dissemination rate. The range $\frac{q_i}{r_i} \ll \Lambda_1^d < \frac{1}{4}$ corresponds to the regime of original type homogeneity $Q < \hat{Q}$, while the range $\Lambda_1^d > \frac{1}{4}$ corresponds to the regime of driver heterogeneity $Q > \hat{Q}$. Note that this regime is determined only by the product of the mutation rate u, the number of metastases at detection M^*, the reciprocal survival probability ρ_0^{-1}, and the dissemination advantage

Q, but not on the magnitude of the dissemination rates q_0, q_1.

Growth advantage. In the case of a growth advantage, we have $\frac{q_1}{q_0} = 1$ and $\frac{r_1}{r_0} = 1 + S$ for a positive selective advantage $S > 0$. Then $\frac{\bar{r}_1}{r_0} = \frac{1+S}{1+\frac{S}{2}} \approx 1 + \frac{S}{2}$, provided that S is small. Although the absolute mutation rate per division is independent of the growth advantage, the rate of generating surviving mutations is increased, and so the scaled mutation rate is $\tilde{u} = \frac{u\,\rho_1}{\rho_0^2} = \frac{u}{\rho_0} \cdot \frac{1+\rho_0 S}{1+S}$. Substituting these relationships into Eq. (20) gives the probability P_{het}^g that some of the first M^* metastases were seeded by type 1 cells,

$$P_{\text{het}}^g \approx 1 - e^{-2(1+S/2)\Lambda_1^g}\sqrt{\pi\,\Lambda_1^g\,(1+\tfrac{S}{2})}\,, \text{ where } \Lambda_1^g = \tfrac{q_0}{r_0(1+S)}\sqrt{\tfrac{u r_0^2 M^*}{\rho_0 q_0^2}\,\tfrac{1+\rho_0 S}{1+S}}^{\,(1+S/2)}. \quad (37)$$

where the superscript g indicates the scenario of growth driver advantages only. Absorbing the factor $(1 + S/2)$ into Λ_1^* and then expanding Λ_1^* about small S gives

$$P_{\text{het}}^g \approx 1 - e^{-2\Lambda_1^g}\sqrt{\pi\Lambda_1^g}\,, \text{ where } \Lambda_1^g \approx \sqrt{\frac{uM^*}{\rho_0}}\left(\frac{\sqrt{\frac{r_0}{q_0}\sqrt{\frac{uM^*}{\rho_0}}}}{e^{1-\frac{p_0}{2}}}\right)^{S}. \quad (38)$$

Solving $\frac{\partial P_{\text{het}}^g}{\partial S} = 0$ for S gives an estimate for the threshold growth advantage $\hat{S}$:

$$\hat{S} = \frac{\log 16 + \log(\frac{uM^*}{\rho_0})}{2 - \rho_0 - \log\frac{r_0}{q_0} - \frac{1}{2}\log(\frac{uM^*}{\rho_0})}, \quad \text{or equivalently,} \quad \hat{\Lambda}_1^g \approx \frac{1}{4}\,. \quad (39)$$

This threshold $\hat{S}$ divides the two regimes in the case of driver mutations conferring an increased growth rate. The range $\frac{q_i}{r_i} \ll \Lambda_1^g < \frac{1}{4}$ corresponds to the regime of original type homogeneity $S < \hat{S}$, while the range $\Lambda_1^g > \frac{1}{4}$ corresponds to the regime of driver heterogeneity $S > \hat{S}$.

If the growth advantage is instead measured as the ratio of birth rates $\frac{b_1}{b_0} = 1 + s$ for fixed death rates $\frac{d_1}{d_0} = 1$, then we convert the above results using the relationship $s = \rho_0 \cdot S$ where $\rho_0 = \frac{r_0}{b_0}$ is the survival probability of original type cells. This gives a threshold linear in the death-to-birth rate ratio: $\hat{s} \approx (1 - \frac{d_0}{b_0})\hat{S}$, where $\hat{S}$ is determined according to Eq. (39). Provided that $\rho_0 \ll 1$, this is almost equivalent to the expression

$$\hat{s} \approx \left(\frac{b_0}{d_0} - 1\right)\hat{S}, \quad (40)$$

which represents a linear boundary for $\hat{s}$ with respect to $\frac{b_0}{d_0}$, with a slope given by $\hat{S}$. This boundary, which divides the different regimes of inter-metastatic driver gene mutation heterogeneity as introduced in Fig. 1, is illustrated in Fig. 4, D and E and fig. S4, A and B. To compute this boundary numerically, we expand Eq. (37) to lowest-order, first about small q and then about small u. Evaluating this linearized expression for the first clone $i = 1$ using the values given in table above, we obtain the numerical estimate $\hat{S} = 0.411$, or equivalently, $\hat{s} = 0.411(\frac{b_0}{0.2475} - 1)$; a simple linear relationship between $\hat{s}$ and b_0. The estimate associated with the second clone $i = 2$ instead gives $\hat{S}^{\downarrow} = 0.985$, which provides a lower bound for b_0. To find an upper bound corresponding to this lower bound, we can find the value of S for which the homogeneity probability $P_H = P_0 + P_{\hat{M}}$ (plotted in fig. S4A) is equal along both the upper and lower bounds. The illustrated upper bound follows from solving the equation $P_H|_{\hat{S}^{\uparrow}} = P_H|_{\hat{S}^{\downarrow}}$, which provides the upper estimate $\hat{S}^{\uparrow} = 0.281$.

Mixed growth-dissemination advantage. In the case of a driver mutation that simultaneously confers a growth and dissemination advantage, we have $\frac{q_1}{q_0} = Q$ and $\frac{r_1}{r_0} = 1 + S$. Combining the factors of Eq. (35) and Eq. (37) as explained in the previous two sections gives a general expression for the probability P_{het} that some of the first M^* metastases were seeded by type 1 cells,

$$P^{gd}_{\text{het}} \approx 1 - e^{-2\Lambda^{gd}_1} \sqrt{\pi\, \Lambda^{gd}_1}\ , \text{ where } \Lambda^{gd}_1 = \tfrac{q_0 Q}{r_0} \tfrac{1+S/2}{1+S} \sqrt{\tfrac{u r_0^2 M^*}{\rho_0 q_0^2 Q} \tfrac{1+\rho_0 S}{1+S}}^{(1+S/2)} . \tag{41}$$

where the superscript gd indicates the scenario of mixed growth-dissemination advantage. This quantity is plotted in fig. S12a. As a specific example, if the growth and dissemination advantages of a particular variant type of interest are linearly correlated with some nonzero regression coefficient β, such that $Q = 1 + \beta S$, then Λ^{gd}_1 can be written independently of Q. After expanding Λ^{gd}_1 about small S as before, we obtain

$$\Lambda^{gd}_1 = \tfrac{q_0}{r_0} \tfrac{(1+\beta S)(1+S/2)}{1+S} \sqrt{\tfrac{u r_0^2 M^*}{\rho_0 q_0^2} \tfrac{1+\rho_0 S}{(1+\beta S)(1+S)}}^{(1+S/2)} \approx \sqrt{\tfrac{uM^*}{\rho_0}} \left(\frac{\sqrt{\frac{r_0}{q_0}\sqrt{\frac{uM^*}{\rho_0}}}}{e^{1-\frac{p_0+\beta}{2}}} \right)^S . \tag{42}$$

Once again, solving the condition $\frac{\partial P^{gd}_{\text{het}}}{\partial S} = 0$ for S gives an estimate $\hat{S}$ for growth component of the threshold advantage, where the value $\Lambda^{gd}_1 \approx \frac{1}{4}$ is achieved, from which we obtain a corresponding estimate $\hat{Q}$ for the dissemination component of the threshold advantage:

$$\hat{S}(\beta) = \frac{\log 16 + \log(\frac{uM^*}{\rho_0})}{2 - \rho_0 - \log\frac{r_0}{q_0} - \frac{1}{2}\log(\frac{uM^*}{\rho_0}) - \beta} \text{ and } \hat{Q}(\beta) = 1 + \beta \hat{S}(\beta). \tag{43}$$

To study how the relative contribution of these two components depends on the regression coefficient β, we note that their derivatives are opposite in sign:

$$\hat{S}'(\beta) < 0 \text{ and } \hat{Q}'(\beta) = \frac{\hat{S}(0)}{\left(1 - \beta\hat{S}(0)/\log(\frac{16uM^*}{\rho_0})\right)^2} > 0. \tag{44}$$

It follows that if the regression coefficient β is low, such that an advantage in growth is coupled only to a small advantage in dissemination, then the critical growth advantage will be large relative to the critical dissemination advantage; however, if the regression coefficient β is high, such that an advantage in growth is coupled only to a large advantage in dissemination, then the critical growth advantage will be small relative to the critical dissemination advantage. As a result, as β increases, the critical advantage increasingly favors dissemination over growth.

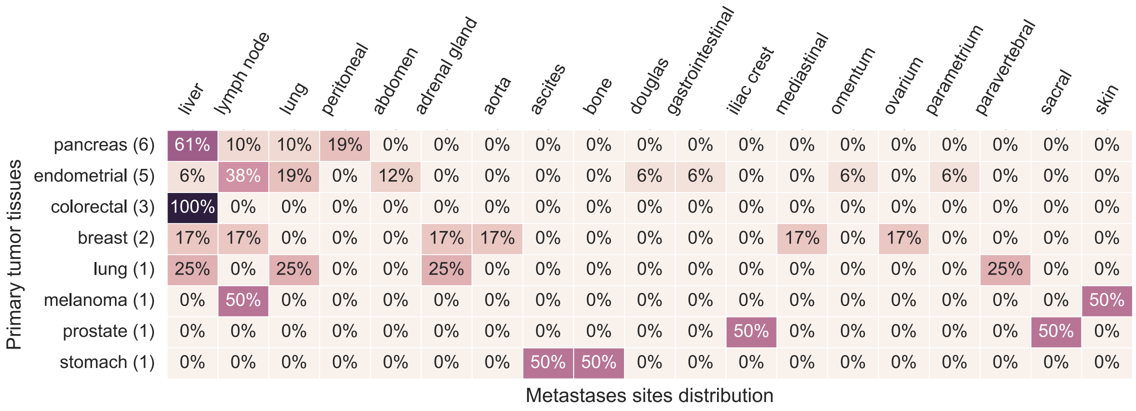
Combining the results of the three cases considered here, we can now evaluate for which driver advantages Q and S the possibility of mixed growth-dissemination driver mutations significantly alters the the probability of homogeneity in the founding cell of metastases, relative to the case where both growth and dissemination driver mutations arise independently. We compute this relative risk ratio, which we denote by R, to be

$$R = \frac{(1 - P_{\text{het}}^{gd})}{(1 - P_{\text{het}}^{g})(1 - P_{\text{het}}^{d})} \approx e^{-2(\Lambda_1^{gd} - \Lambda_1^{g} - \Lambda_1^{d})} \sqrt{\frac{\Lambda_1^{gd}}{\pi \Lambda_1^{g} \Lambda_1^{d}}} \tag{45}$$

where Λ_1^g and Λ_1^d are both calculated with half of the usual mutation probability, $u/2$, so that driver mutations are evenly grouped into growth and dissemination drivers. This quantity is plotted in fig. S12b. Substituting in the values of Λ_1^d, Λ_1^g, and Λ_1^{gd} from above and expanding about small S as before, we can obtain an approximate expression for the relative risk ratio,

$$R \approx \frac{1}{\alpha} \sqrt{\frac{e^{\frac{Q-1}{2}}}{\pi Q}} \exp\left\{ 2\alpha \left[\sqrt{Q} + \left(\frac{\sqrt{\frac{\alpha r_0}{q_0}}}{e^{1 - \frac{p_0}{2}}} \right)^S \left(1 - e^{\frac{Q-1}{2}} \right) \right] \right\} \tag{46}$$

where we have defined $\alpha = \sqrt{uM^*/\rho_0}$ to simplify the result. For a more exact result, numerical integration using Eq. 17 can give a more precise calculation of P_{het} for each of the three cases, and R can be computed according to Eq. 45 (fig. S12b). We find that, for the parameter values indicated in the table above, the relative risk ratio is always very close but slightly greater than $R = 1$. This indicates that simultaneous acquisition of the birth and dissemination driver advantages is not substantially more likely to result in homogeneous metastases than independent acquisition of both at equal rates. The mixed and independent cases are most similar when the driver advantages S and Q are small.



	liver	lymph node	lung	peritoneal	abdomen	adrenal gland	aorta	ascites	bone	douglas	gastrointestinal	iliac crest	mediastinal	omentum	ovarium	parametrium	paravertebral	sacral	skin
pancreas (6)	61%	10%	10%	19%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
endometrial (5)	6%	38%	19%	0%	12%	0%	0%	0%	0%	6%	6%	0%	0%	6%	0%	6%	0%	0%	0%
colorectal (3)	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
breast (2)	17%	17%	0%	0%	0%	17%	17%	0%	0%	0%	0%	0%	17%	0%	17%	0%	0%	0%	0%
lung (1)	25%	0%	25%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%
melanoma (1)	0%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%
prostate (1)	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	50%	0%	0%	0%	0%	0%	50%	0%
stomach (1)	0%	0%	0%	0%	0%	0%	0%	50%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Figure S1: Cohort overview: Cancer types and their metastases sites distributions. In total, 76 untreated metastases at 19 distinct sites from eight cancer types were analyzed. Mean of 3.8 metastases per subject. Median of 3 metastases per subject (range 2-8).

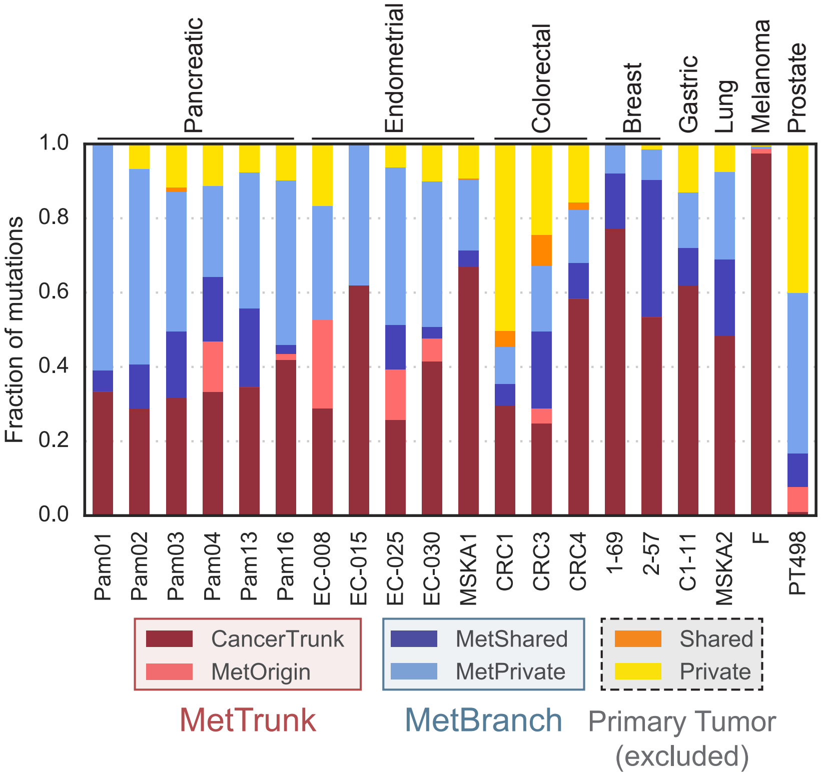


Figure S2: Mutations classified by tumor phylogenies. Variants were classified into six categories. *CancerTrunk*: present in all samples; *MetOrigin*: present in all metastases but absent in some samples of the primary tumor; *MetShared*: present in multiple but not all metastases; *MetPrivate*: present in a single metastasis; *Shared*: present in a subset of non-metastases samples; *Private*: present in a single non-metastasis sample. Most genetic heterogeneity stems from metastases (light and dark blue). *CancerTrunk* and *MetOrigin* form the ***MetTrunk***. *MetShared* and *MetPrivate* form the ***MetBranch***. Mutations that only occur in the primary tumor (*Shared*, *Private*) are excluded from the analysis since these variants are irrelevant for inter-metastatic heterogeneity.

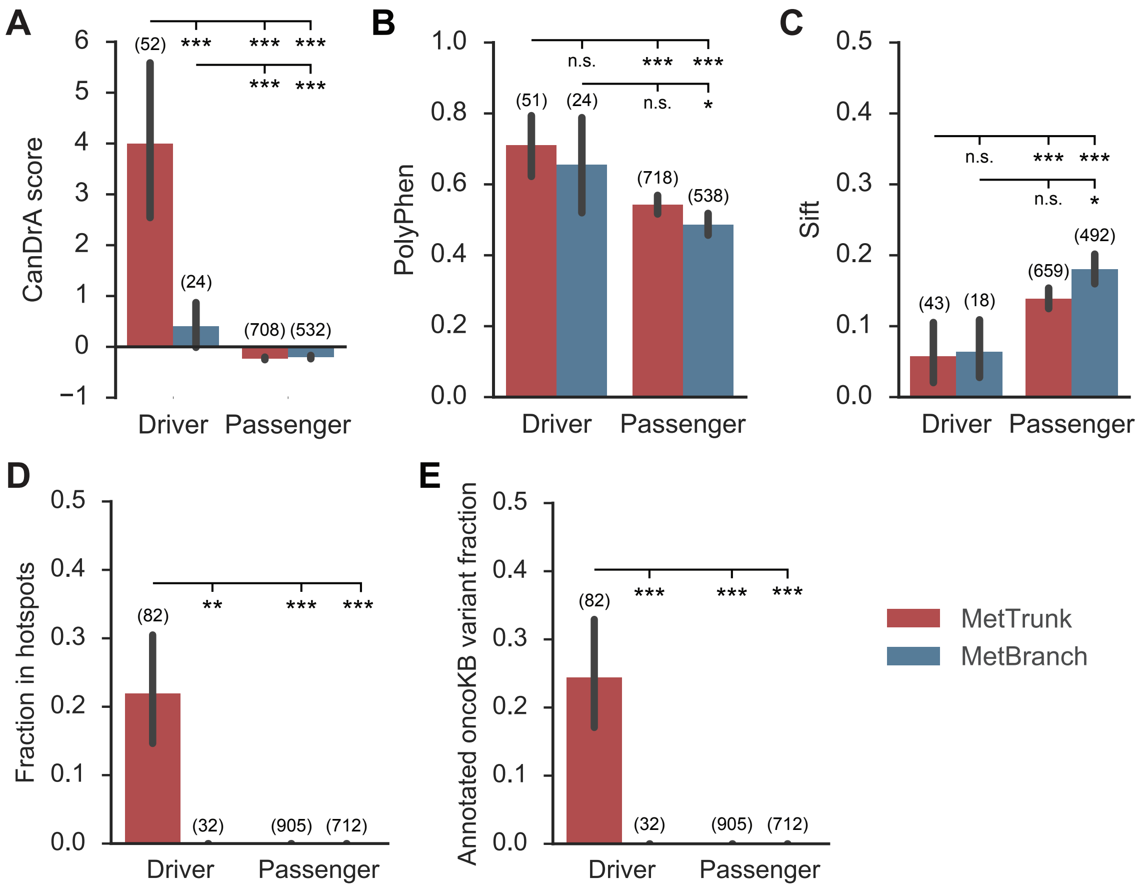


Figure S3: Additional analyses of nonsynonymous mutations in putative driver genes. (**A**) CanDrA predicted increased functional consequences for variants in putative driver genes in MetTrunk (*40*). (**B**-**C**) PolyPhen2 scores (high value indicates likely damaging consequence) show a trend but no significant difference between putative driver gene mutations on the MetTrunk (mean: 0.54) and MetBranch (mean: 0.49) was observed. SIFT prediction scores (below 0.05 predicted to affect protein function) show a trend but no significant difference between putative driver gene mutations on the MetTrunk (mean: 0.14) and MetBranch (mean: 0.18) was observed. Note that PolyPhen-2 and SIFT aim to find damaging germline variants and perhaps therefore cannot perfectly distinguish between somatic driver and passenger mutations (*41, 42*). (**D**) Mutations in previously reported cancer hotspots were only observed in MetTrunk (*35*). (**E**) Annotated variants in OncoKB were only observed in MetTrunk (*36*). Two-sided Wilcoxon rank-sum tests were used in panels (**A**)-(**C**). Two-sided Fishers exact tests were used in panels (**D**) and (**E**). Thick black bars denote 90% confidence interval. No other statistically significant differences between subgroups were observed. Numbers in brackets denote the number of evaluated variants observed in each group. * indicates $P < 0.05$, ** indicates $P < 0.01$, *** indicates $P < 0.001$, n.s. denotes no significant difference.

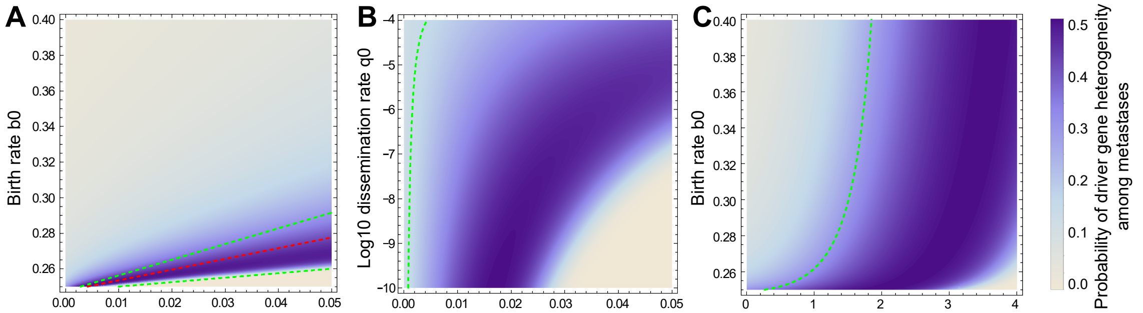


Figure S4: Probability estimates for driver gene mutation heterogeneity among four metastases. A primary tumor stochastically expands and seeds metastases starting from a single advanced cancer cell (section Mathematical Modeling). The presence of a driver gene mutation confers an advantage s in the birth rate (**A** and **B**) or in the dissemination rate q (**C**). Heat maps are colored according to the analytically estimated probability for heterogeneity of the *first* driver gene mutation among four metastases (effects of further drivers are explored via simulations; see Fig. 4). Low probability regions in the bottom-right of each plot correspond to the regime in which the first four metastases were seeded by the driver subclone. Green dashed lines depict bounds separating parameter regions of likely inter-metastatic driver homogeneity from heterogeneity, and the red dashed line depicts the critical growth advantage (Mathematical Modeling, Eq. (37)). Axes correspond to different parameter choices; unless otherwise indicated, the parameter values used were: growth rate $r_0 = 1.24\%$ per day, cell death rate $d = 0.2475$ per day, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$, driver gene mutation rate per cell division $u = k \cdot \hat{u}$, dissemination rate per cell per day of $q_0 = 10^{-7}$.

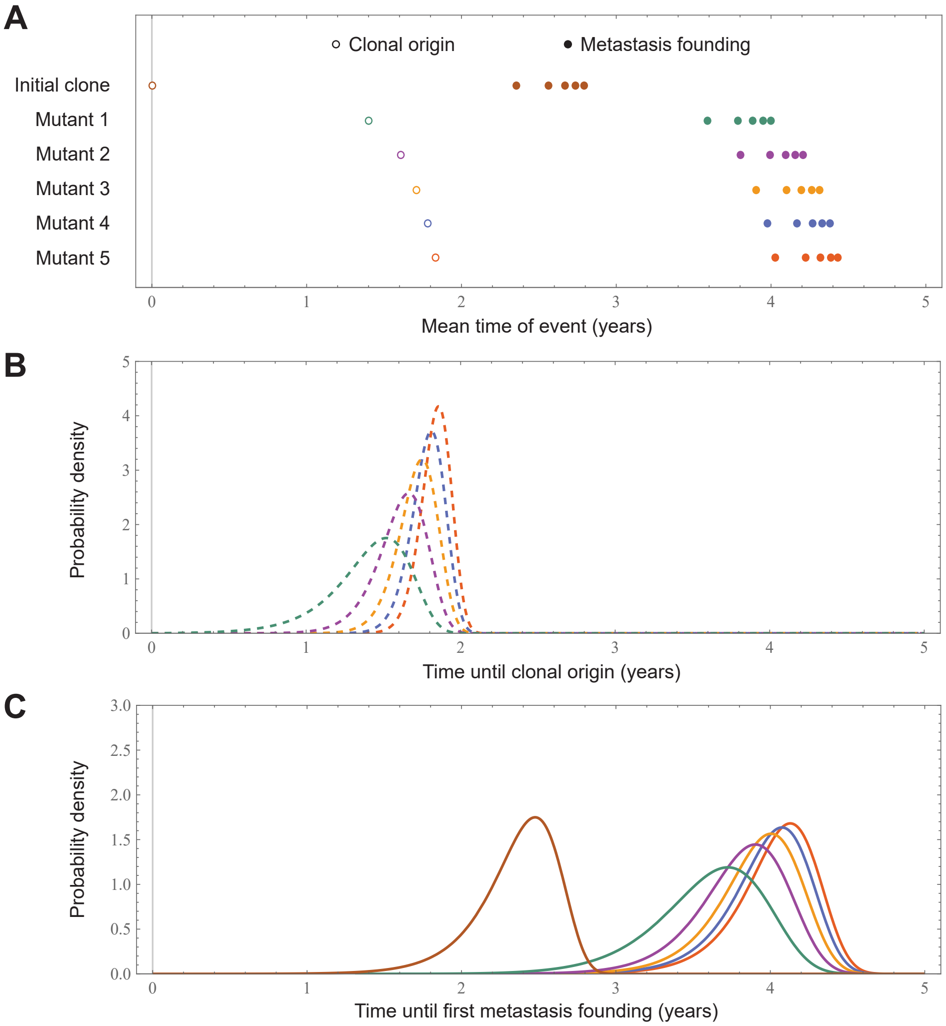


Figure S5: Analytical estimates for the timing of driver gene mutations and metastases founding events for a 95% death-to-birth rate ratio. (**A**) Mean time at which the first five surviving mutant clones are originated by new driver mutations (open circles) and the mean time at which each of these clones and the initial clone disseminates to found the first five surviving metastases (closed circles). (**B**) Probability density functions (PDFs) for the time until the first five surviving mutant clones are originated by new driver mutations. (**C**) PDFs for the time until the initial clone as well as each of the first five surviving mutant clones found their first surviving metastasis (solid lines). Parameter values: death rate $d = 0.2475$ per day, death-to-birth rate ratio $d/b_0 = 95\%$, relative driver advantage $s = 0.4\%$, dissemination rate per cell per day $q = 10^{-7}$, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$, driver gene mutation rate per cell division $u = k \cdot \hat{u}$ per cell division, number of metastases $m = 4$.

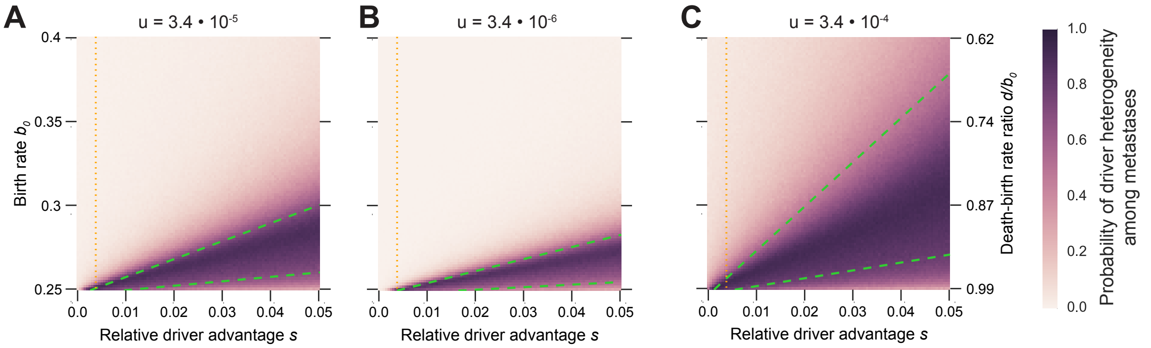


Figure S6: Probability for driver gene mutation heterogeneity among four metastases for varying mutation rates. A primary tumor stochastically expands and seeds metastases starting from a single advanced cancer cell (section Mathematical Modeling). An additional driver gene mutation confers a growth advantage to the cancer cells. The driver gene mutation rate per cell division, u, is the product of the number of driver gene mutations k across the exome and the point mutation rate $\hat{u}$. The birth rate of cancer cells with an additional driver gene mutation is given by $b_1 = b_0 \times (1 + s)$. High probability of driver gene mutation heterogeneity is shown in dark purple. Low probability for driver gene mutation heterogeneity is shown in light beige. Green dashed lines depict numerical bounds separating parameter regions of likely inter-metastatic driver mutation homogeneity from likely heterogeneity (section Mathematical Modeling). Orange line indicates the estimated average relative driver advantage s of 0.4% (*25*). (**A**) Typical driver gene mutation rate per cell division $u = 3.4 \cdot 10^{-5}$. (**B**) A 10-fold decreased driver gene mutation rate strongly decreases the probability for driver gene mutation heterogeneity. (**C**) An 10-fold increased driver gene mutation rate strongly increases the probability for driver gene mutation heterogeneity. Parameter values: death rate $d = 0.2475$, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$ per cell division, number of metastases $m = 4$, dissemination rate per cell per day $q = 10^{-7}$.

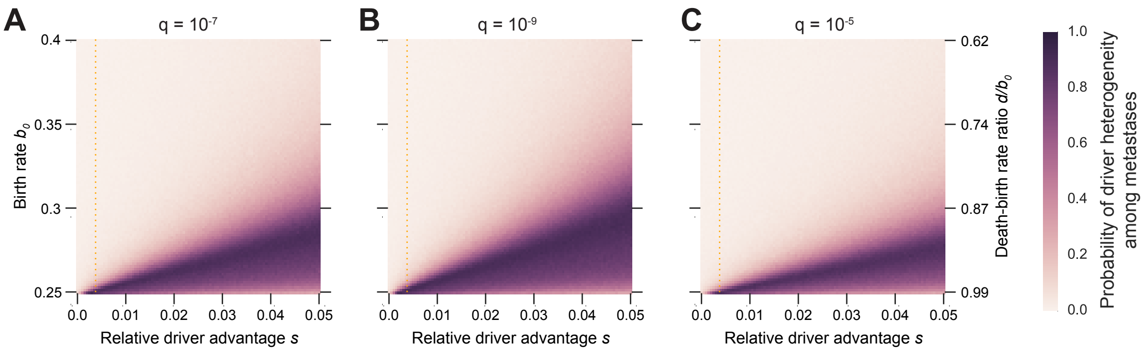


Figure S7: Probability for driver gene mutation heterogeneity among four metastases for varying dissemination rates. A primary tumor stochastically expands and seeds metastases starting from a single advanced cancer cell (section Mathematical Modeling). An additional driver gene mutation confers a growth advantage to the cancer cells. The driver gene mutation rate per cell division, u, is the product of the number of driver gene mutations k across the exome and the point mutation rate $\hat{u}$. The birth rate of cancer cells with an additional driver gene mutation is given by $b_1 = b_0 \times (1 + s)$. High probability of driver gene mutation heterogeneity is shown in dark purple. Low probability for driver gene mutation heterogeneity is shown in light beige. Orange line indicates the estimated average relative driver advantage s of 0.4% (*25*). (**A**) Dissemination rate per cell per day $q = 10^{-7}$. (**B**) A 100-fold decreased dissemination rate ($q = 10^{-9}$). (**C**) A 100-fold increased dissemination rate ($q = 10^{-5}$). Parameter values: death rate $d = 0.2475$ per day, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$ per cell division, number of metastases $m = 4$.

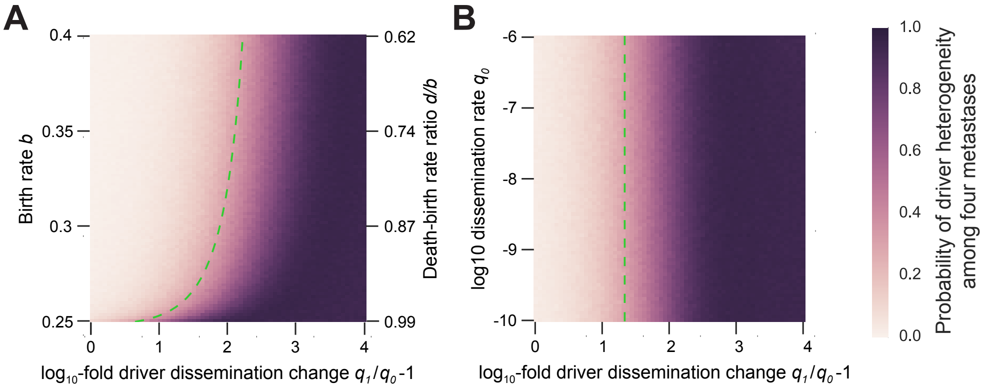


Figure S8: Probability for driver gene mutation heterogeneity among four metastases when driver gene mutations increase the dissemination rate. A primary tumor stochastically expands and seeds metastases starting from a single advanced cancer cell (section Mathematical Modeling). An additional driver gene mutation increase the dissemination rate q_1 of cancer cells (birth rates are equal $b_0 = b_1$). Such a scenario mimics the case where a particular location in the primary tumor (e.g., close proximity to a blood vessel) has an increased dissemination rate. The driver gene mutation rate per cell division, u, is the product of the number of driver gene mutations k across the exome and the point mutation rate $\hat{u}$. Green dashed lines depict numerical bounds separating parameter regions of likely inter-metastatic driver mutation homogeneity from likely heterogeneity (section Mathematical Modeling). (**A**) Fixed dissemination rate $q_0 = 10^{-7}$ of the original cells. (**B**) Fixed death-birth rate ratio $d/b_0 = 0.95$. Probability of heterogeneity is independent of the dissemination rate q_0 of the original cells (section Mathematical Modeling). Parameter values: death rate $d = 0.2475$ per day, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$, number of metastases $m = 4$.

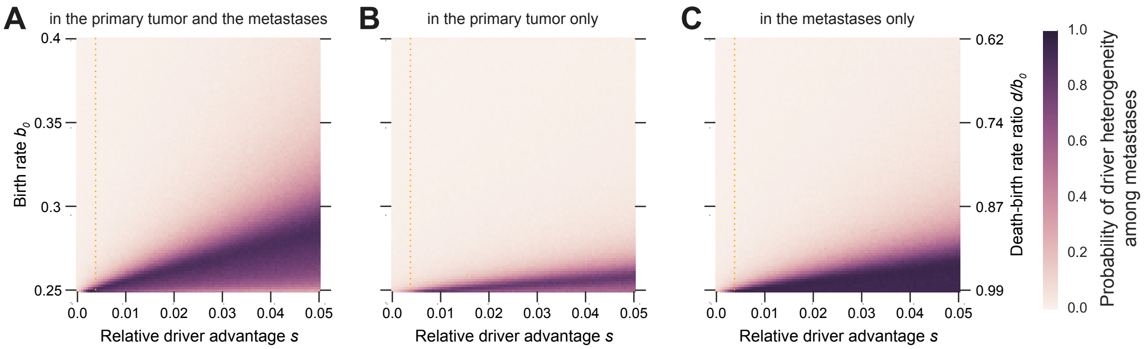


Figure S9: Probability for driver gene mutation heterogeneity among four metastases for site-dependent mutation effects. The probability for driver gene mutation heterogeneity among metastases strongly decreases if the driver advantage is not conferred both in the primary tumor and the metastases. A primary tumor stochastically expands and seeds metastases starting from a single advanced cancer cell (section Mathematical Modeling). The growth advantage conferred to the cancer cells by an additional driver gene mutation depends on the microenvironment. An additional driver gene mutation confers a growth advantage in both the primary tumor and the metastases (**A**), only in the primary tumor (**B**), or only in the metastases (**C**). Depending on the site, the birth rate of cancer cells with an additional driver gene mutation is either $b_1 = b_0$ or $b_1 = b_0 \times (1 + s)$. High probability of driver gene mutation heterogeneity is shown in dark purple. Low probability for driver gene mutation heterogeneity is shown in light beige. Orange line indicates the estimated average relative driver advantage s of 0.4% (*25*). Parameter values: death rate $d = 0.2475$ per day, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$, driver gene mutation rate per cell division $u = k \cdot \hat{u}$ per cell division, number of metastases $m = 4$.

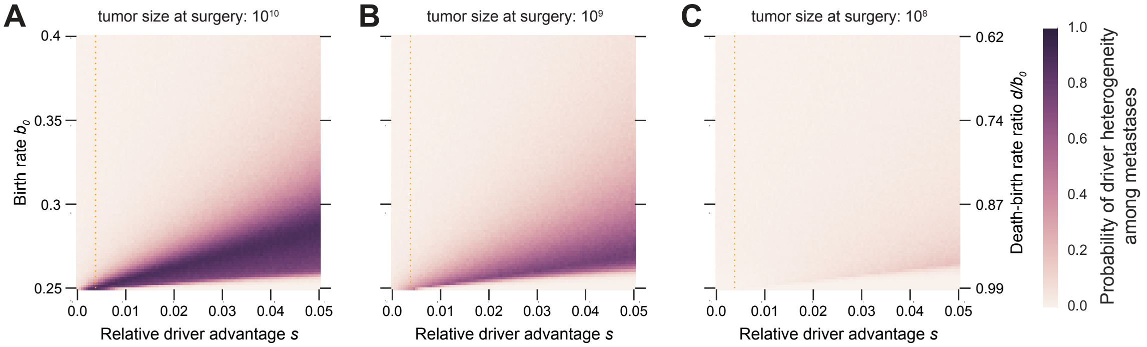


Figure S10: Probability for driver gene mutation heterogeneity among four metastases when the primary tumor is removed at a specific size. A primary tumor stochastically expands and seeds metastases starting from a single advanced cancer cell (section Mathematical Modeling). An additional driver gene mutation confers a growth advantage to the cancer cells. The birth rate of cancer cells with an additional driver gene mutation is given by $b_1 = b_0 \times (1 + s)$. When the primary tumor reaches a given size, it is removed and no more metastases can be seeded. All realizations where at least $m = 4$ surviving metastases were seeded before the removal of the primary tumor were evaluated. High probability of driver gene mutation heterogeneity is shown in dark purple. Low probability for driver gene mutation heterogeneity is shown in light beige. Orange line indicates the estimated average relative driver advantage s of 0.4% (*25*). Parameter values: death rate $d = 0.2475$ per day, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$, driver gene mutation rate per cell division $u = k \cdot \hat{u}$, dissemination rate per cell per day $q = 10^{-7}$, number of metastases $m = 4$.

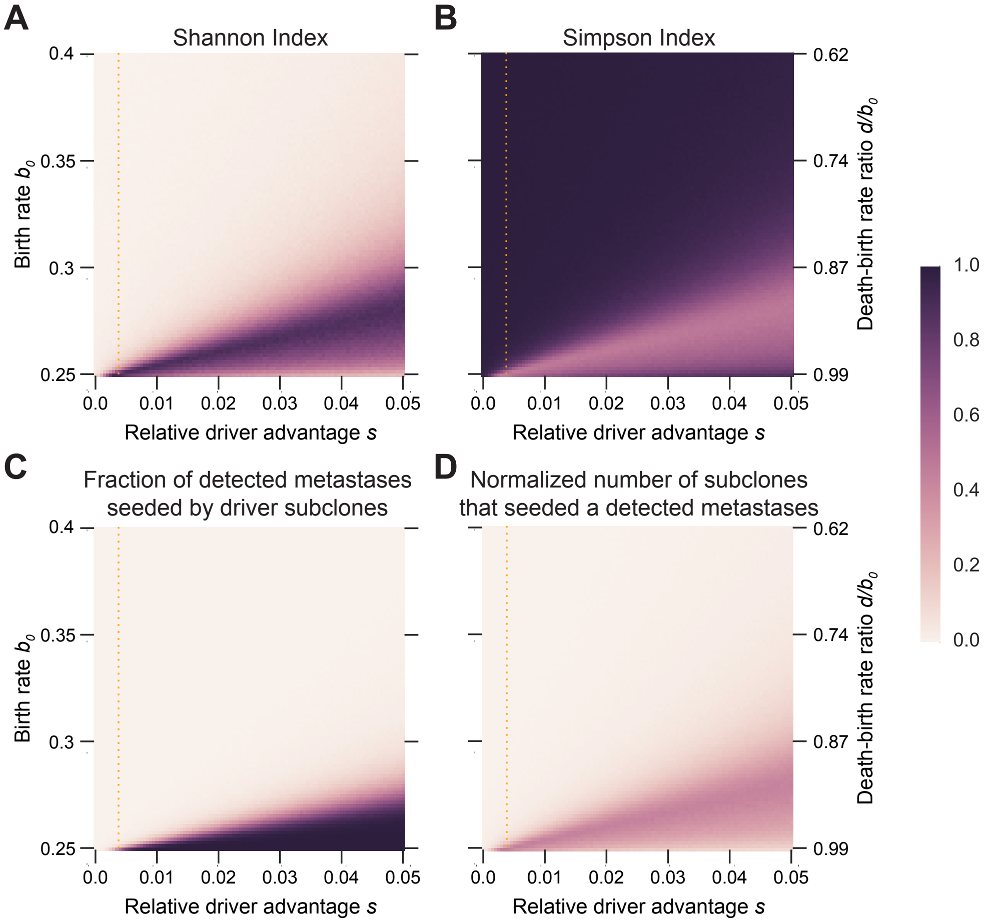


Figure S11: Different measures of driver gene mutation heterogeneity among four metastases. (**A**) Shannon Index is an information statistic that summarizes the diversity in a population. (**B**) Simpson Index denotes the probability that the founder cells of two randomly taken metastases share the same driver gene mutations. (**C**) Denotes the fraction of metastases that are seeded by any driver subclone. (**D**) Denotes the number of distinct subclones that seeded any detectable metastasis (normalized by the number of detected metastases m). Orange line indicates the estimated average relative driver advantage s of 0.4% (*25*). Parameter values: death rate $d = 0.2475$ per day, dissemination rate per cell per day $q = 10^{-7}$, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$, driver gene mutation rate per cell division $u = k \cdot \hat{u}$ per cell division, number of metastases $m = 4$.

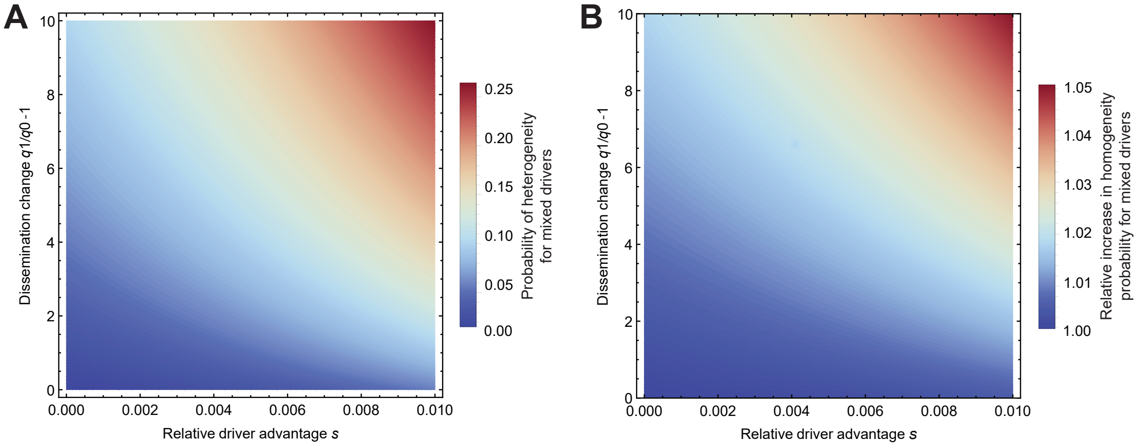
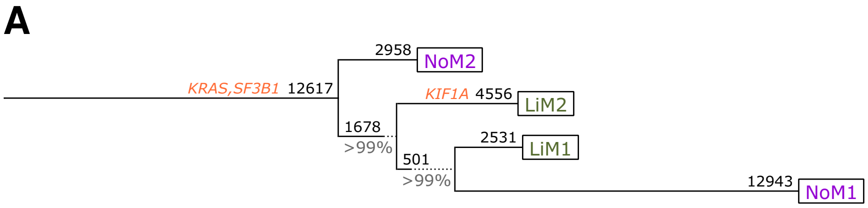


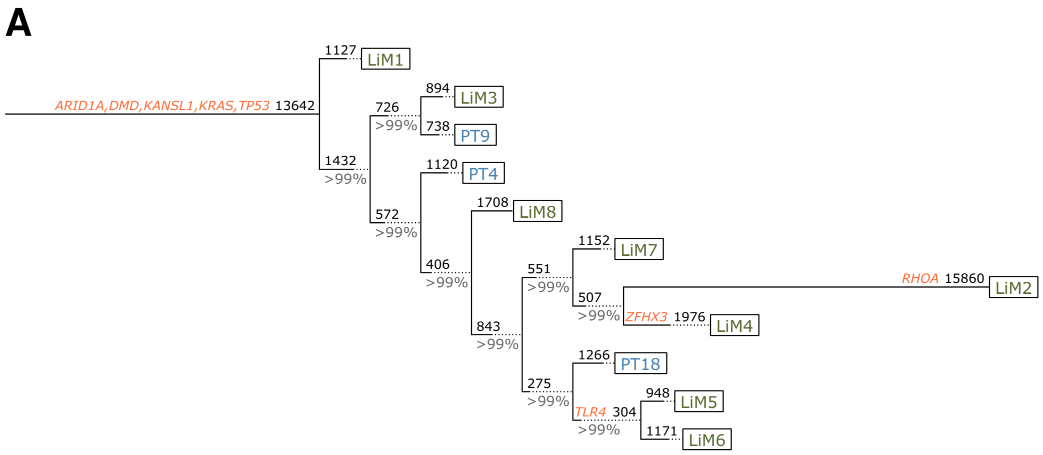
Figure S12: Probability estimates for mixed growth-dissemination drivers. A primary tumor stochastically expands and seeds metastases starting from a single advanced cancer cell (section Mathematical Modeling). The presence of a driver gene mutation simultaneously confers an advantage $s = b_1/b_0 - 1$ in the birth rate and $q_1/q_0 - 1$ in the dissemination rate. Heat maps are colored according to the analytically estimated probability for heterogeneity of the *first* driver gene mutation among four metastases (effects of further drivers are explored via simulations; see Fig. 4). (**A**) The colorbar indicates the probability that the first four seeded metastases are heterogeneous when drivers confer a simultaneous growth and dissemination advantage. (**B**) The colorbar indicates the relative risk ratio R (Eq. (45)), defined as the probability that the first four seeded metastases are homogeneous when drivers confer a simultaneous growth and dissemination advantage, divided by the probability that the first four seeded metastases are homogeneous when each driver instead confers either a growth or dissemination advantage with equal probability. Because the values are all near $R \approx 1.00$, these two cases behave roughly similarly. Parameter values: death rate $d = 0.2475$ per day, death-birth ratio $d/b_0 = 95\%$, dissemination rate per cell per day $q = 10^{-7}$, number of driver gene mutations $k = 34,000$, point mutation rate $\hat{u} = 10^{-9}$, driver gene mutation rate per cell division $u = k \cdot \hat{u}$ per cell division, number of metastases $m = 4$.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*KRAS*	missense	1319	< 0.001	-2.32	10.76	moderate	yes	yes
MT	*SF3B1*	intronic splice	0	n/a	n/a	n/a	modifier	no	no
MB	*KIF1A*	missense	0	1.0	3.06	0.06	moderate	no	no

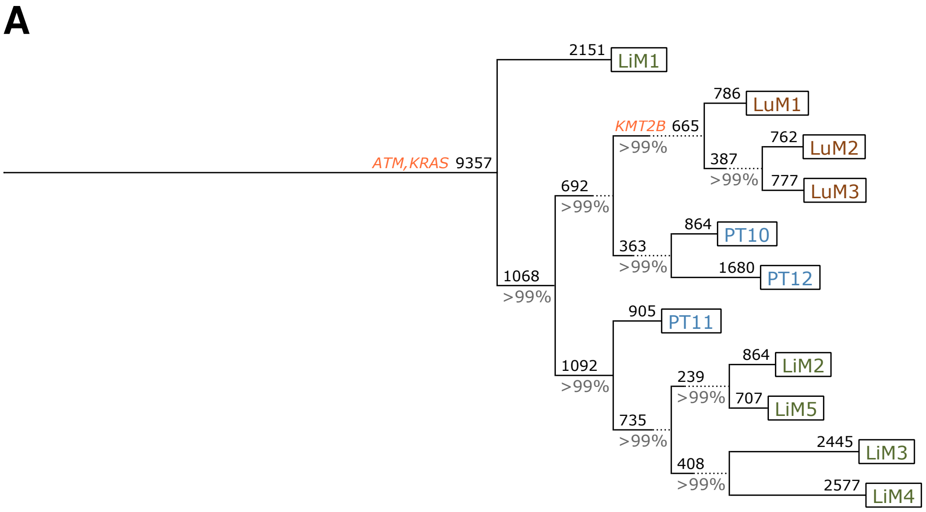
Figure S13: Two variants in the putative driver genes *KRAS* and *SF3B1* were universally present in pancreatic cancer patient Pam01 of Makohon-Moore et al. (*17*). A variant in *KIF1A* was only present in LiM2. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: LiM1 left liver, LiM2 right liver, NoM1 pelvic lymph node, NoM2 portal lymph node. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*ARID1A*	nonsense	0	n/a	n/a	n/a	high	no	no
MT	*DMD*	missense	0	1.0	1.47	1.46	moderate	no	no
MT	*KANSL1*	nonsense	0	n/a	n/a	n/a	high	no	no
MT	*KRAS*	missense	1319	< 0.001	-2.32	10.76	moderate	yes	yes
MT	*TP53*	missense	4	< 0.001	-6.77	30.72	moderate	no	yes
MB	*RHOA*	missense	0	1.0	n/a	n/a	modifier	no	no
MB	*TLR4*	missense	0	1.0	4.20	-0.75	moderate	no	no
MB	*ZFHX3*	missense	0	1.0	2.26	0.03	moderate	no	no

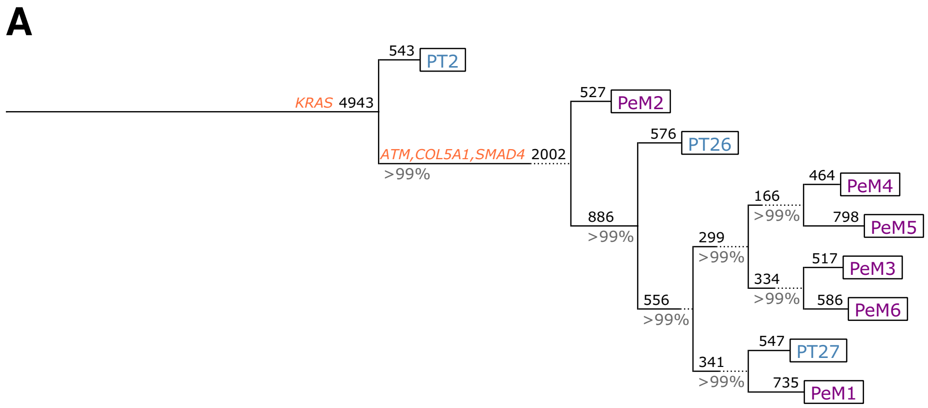
Figure S14: Five variants in the putative driver genes *ARID1A*, *DMD*, *KANSL1*, *KRAS*, and *TP53* were universally present in pancreatic cancer patient Pam02 of Makohon-Moore et al. (*17*). Branched variants were found in the putative driver genes *RHOA*, *ZFHX3*, and *TLR4*. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and agreed well with the originally reported phylogeny which was based on targeted sequencing data despite only partially overlapping sample sets. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: pancreas: PT4, PT9, PT18; liver: LiM1-LiM8. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch. Functional analysis predicts no effect for the heterogeneous variants.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*ATM*	splice donor	0	n/a	n/a	n/a	high	no	no
MT	*KRAS*	missense	0	< 0.001	-2.93	10.45	moderate	yes	yes
MB	*KMT2B*	missense	0	1.0	0.90	0.14	moderate	no	no

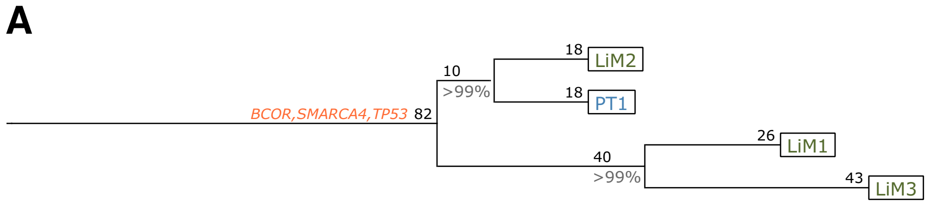
Figure S15: Two variants in the putative driver genes *ATM* and *KRAS* were universally present in pancreatic cancer patient Pam03 of Makohon-Moore et al. (*17*). A variant in *KMT2B* was present in all lung metastases but predicted to have no functional effect. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and agreed well with the originally reported one. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: pancreas: PT10-PT12; liver: LiM1-LiM5, lung: LuM1-3. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*ATM*	missense	0	1.0	0.87	4.39	moderate	no	no
MT	*COL5A1*	missense	0	1.0	0.27	-0.46	moderate	no	no
MT	*KRAS*	missense	1079	< 0.001	-2.32	10.68	moderate	yes	yes
MT	*SMAD4*	missense	5	< 0.001	-5.74	1.64	moderate	yes	no

Figure S16: Four variants in the putative driver genes *ATM*, *COL5A1*, *KRAS*, and *SMAD4* were present in all metastases of pancreatic cancer patient Pam04 of Makohon-Moore et al. (*17*). Sample PT2 had very low purity (16.2%) and hence many variants were undetected although they were validated to be present by targeted sequencing (incl. *ATM*, *COL5A1*, *SMAD4*). (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: pancreas: PT2, PT26, PT27; peritoneal: PeM1-PeM6. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases.

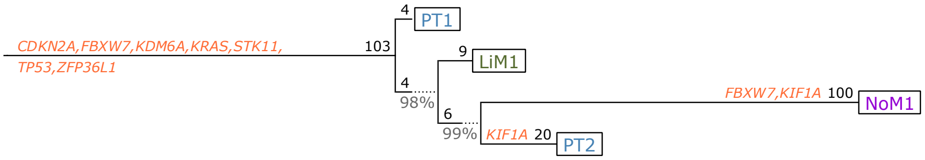


B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*BCOR*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*SMARCA4*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*TP53*	splice donor	44	n/a	n/a	n/a	high	no	no

Figure S17: Three variants in the putative driver genes *BCOR*, *SMARCA4*, and *TP53* were universally present in pancreatic cancer patient Pam13 of Makohon-Moore et al. (*17*). (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: pancreas: PT1; liver: LiM1-LiM3. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases.

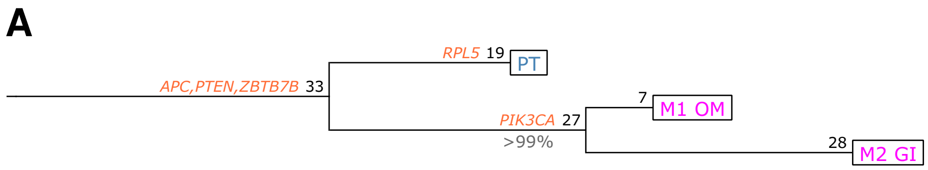
A



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*CDKN2A*	nonsense	0	0.44	n/a	n/a	high	yes	no
MT	*FBXW7*	missense	0	0.45	-3.70	4.05	moderate	no	no
MT	*KDM6A*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*KRAS*	missense	1079	< 0.001	-2.32	10.68	moderate	yes	yes
MT	*STK11*	missense	0	n/a	-2.97	4.73	moderate	no	no
MT	*TP53*	deletion	6	n/a	n/a	n/a	moderate	no	no
MT	*ZFP36L1*	deletion	0	n/a	n/a	n/a	moderate	no	no
MB	*FBXW7*	missense	0	1.0	0.35	3.96	moderate	no	no
MB	*KIF1A*	missense	0	1.0	2.83	0.23	moderate	no	no
PT	*KIF1A*	intronic splice	0	n/a	n/a	n/a	low	no	no

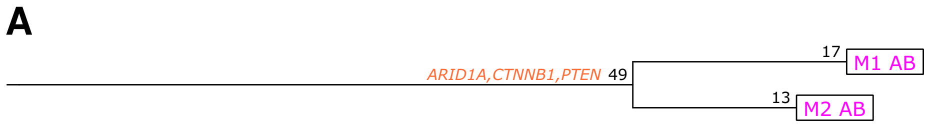
Figure S18: Unlikely functional driver gene heterogeneity among two liver metastases of pancreatic cancer subject Pam16 of Makohon-Moore et al. (*17*). Seven variants in putative driver genes *CDKN2A*, *FBXW7*, *KDM6A*, *KRAS*, *STK11*, *TP53*, and *ZFP36L1* were universally present. Heterogeneous variants in putative driver gene *FBXW7*, *KIF1A*, and *KIF1A* were predicted to have no functional effect. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: pancreas: PT1, PT2; liver: LiM1, lymph node: NoM1. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch, PT denotes primary tumor samples.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*APC*	unknown	0	n/a	n/a	n/a	high	no	no
MT	*PIK3CA*	missense	819	$<$ 0.001	-4.42	2.52	moderate	yes	yes
MT	*PTEN*	splice donor	2	n/a	n/a	n/a	high	no	no
MT	*ZBTB7B*	missense	0	1.0	2.80	0.07	moderate	no	no
PT	*RPL5*	missense	0	1.0	0.95	-0.04	moderate	no	no

Figure S19: Four variants in *APC*, *PTEN*, *PIK3CA*, *ZBTB7B* are ubiquitously present in two metastases of endometrial cancer subject EC-008 of Gibson et al. (*16*). No support for the originally reported variant in *ARID1A* was found in the WES data (previously identified by targeted deep sequencing). (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and is equivalent to the originally reported phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: PT endometrioid grade 2, M1 omentum, M2 gastrointestinal. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, PT denotes primary tumor samples.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*ARID1A*	missense	0	0.010	1.51	1.02	moderate	no	no
MT	*CTNNB1*	missense	59	$<$ 0.001	-7.09	1.04	moderate	yes	yes
MT	*PTEN*	missense	9	$<$ 0.001	-6.42	3.06	moderate	no	yes

Figure S20: Three variants in *ARID1A*, *CTNNB1*, *PTEN* are ubiquitously present in two metastases of endometrial cancer subject EC-015 of Gibson et al. (*16*). (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: abdomen: M1 and M2. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases.

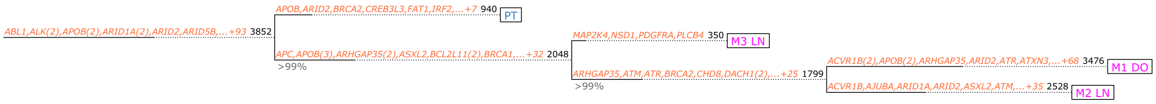
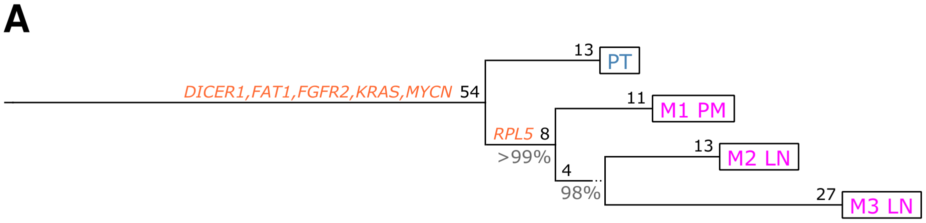

ABL1,ALK(2),APOB(2),ARID1A(2),ARID2,ARID5B,...+93 3852
APOB,ARID2,BRCA2,CREB3L3,FAT1,IRF2,...+7 940
PT
APC,APOB(3),ARHGAP35(2),ASXL2,BCL2L11(2),BRCA1,...+32 2048
>99%
MAP2K4,NSD1,PDGFRA,PLCB4 350
M3 LN
ARHGAP35,ATM,ATR,BRCA2,CHD8,DACH1(2),...+25 1799
>99%
ACVR1B(2),APOB(2),ARHGAP35,ARID2,ATR,ATXN3,...+68 3476
M1 DO
ACVR1B,AJUBA,ARID1A,ARID2,ASXL2,ATM,...+35 2528
M2 LN

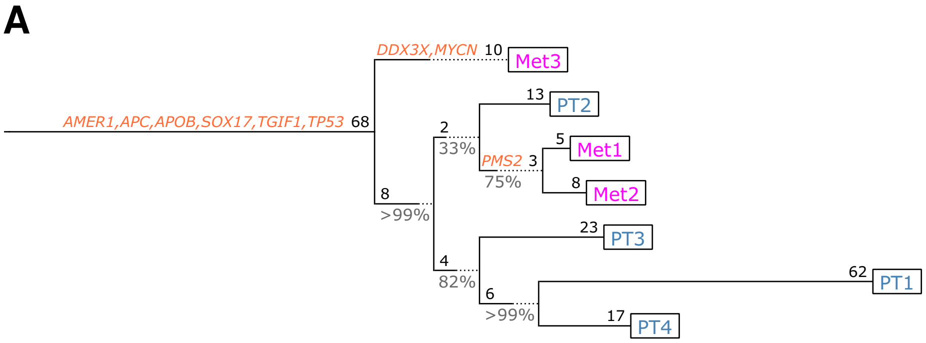
Figure S21: Identified 310 variants in 148 putative driver genes while the original publication highlighted five variants in *FGFR2*, *FLT3*, *NF1*, *PIK3CA*, and *PTEN* ubiquitously present among three metastases of endometrial cancer subject EC-025 of Gibson et al. (*16*). This cancer has been classified as POLE subtype as a mutation in *POLE* has been acquired leading to ultrahigh mutation rates. Due to the high mutation numbers (9065 missense mutations; Fig. 2A), we were unable to identify functional driver gene mutations. Cancer phylogeny was inferred by *Treeomics* (*20*) and is equivalent to the originally reported phylogeny, assuming that M3 corresponds to Lymph node 1 in the original publication). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: PT endometrioid grade 3, M1 douglas, M2 lymph node, M3 lymph node.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*DICER1*	missense	0	0.030	-1.88	-0.15	moderate	no	yes
MT	*FAT1*	missense	0	1.0	3.38	0.37	moderate	no	no
MT	*FGFR2*	missense	0	0.010	-1.20	1.13	moderate	no	yes
MT	*KRAS*	missense	2	< 0.001	-2.66	0.46	moderate	yes	yes
MT	*MYCN*	missense	26	0.32	-0.57	-0.26	moderate	yes	yes
MT	*RPL5*	missense	0	1.0	0.95	0.10	moderate	no	no

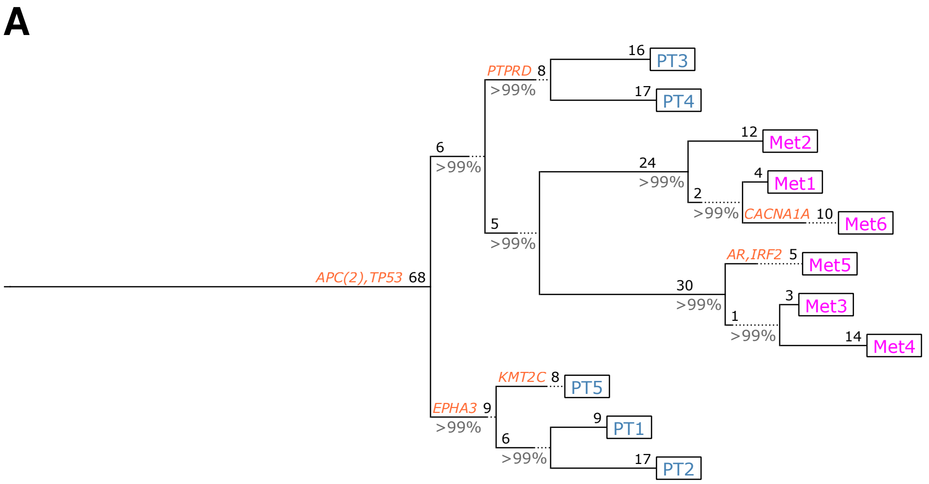
Figure S22: Six variants in *DICER1*, *FAT1*, *FGFR2*, *KRAS*, *MYCN*, and *RPL5* were ubiquitously present in all metastases of endometrial cancer subject EC-030 of Gibson et al. (*16*). (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and is equivalent to the originally reported phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: PT non-endometrioid grade 3; parametrium: M1; lymph nodes: M2 and M3. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*AMER1*	missense	0	1.0	-0.63	0.25	moderate	no	no
MT	*APC*	frame shift	2	n/a	n/a	n/a	high	no	no
MT	*APOB*	missense	0	1.0	1.97	-0.42	moderate	no	no
MT	*SOX17*	missense	0	1.0	1.26	-0.31	moderate	no	no
MT	*TGIF1*	missense	0	1.0	-0.24	-0.28	moderate	no	no
MT	*TP53*	missense	89	< 0.001	-12.33	8.60	moderate	yes	no
MB	*DDX3X*	missense	0	n/a	1.38	0.21	moderate	no	no
MB	*MYCN*	frame shift	0	n/a	n/a	n/a	high	no	no
MB	*PMS2*	missense	0	1.0	-0.23	1.02	moderate	no	no

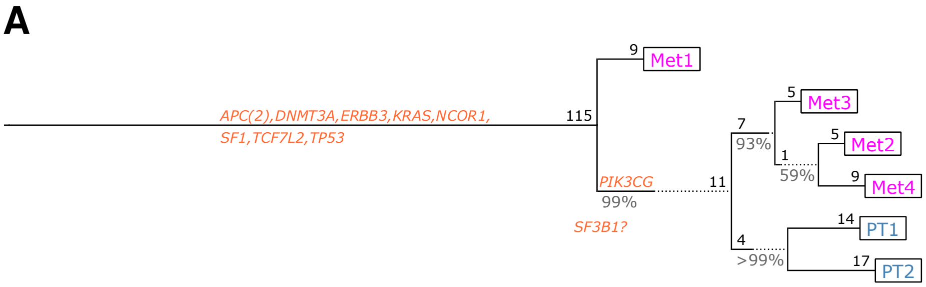
Figure S23: Met3 of colon cancer subject CRC1 of Kim et al. (*15*) may have accumulated three additional subclonal variants in putative driver genes. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and did not agree well with the originally reported phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: sigmoid: PT1-4; liver: Met1-3. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch. Low confidence variants due to very low VAFs are shaded gray.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*APC*	nonsense	9	n/a	n/a	n/a	high	no	no
MT	*APC*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*TP53*	frame shift	123	n/a	n/a	n/a	high	no	no
MB	*AR*	deletion	0	n/a	n/a	n/a	moderate	no	no
MB	*CACNA1A*	deletion	3	n/a	n/a	n/a	moderate	no	no
MB	*IRF2*	missense	0	n/a	0.31	0.07	moderate	no	no
PT	*EPHA3*	nonsense	0	n/a	n/a	n/a	high	no	no
PT	*KMT2C*	splice acceptor	0	n/a	n/a	n/a	high	no	no
PT	*PTPRD*	missense	0	1.0	n/a	0.27	modifier	no	no

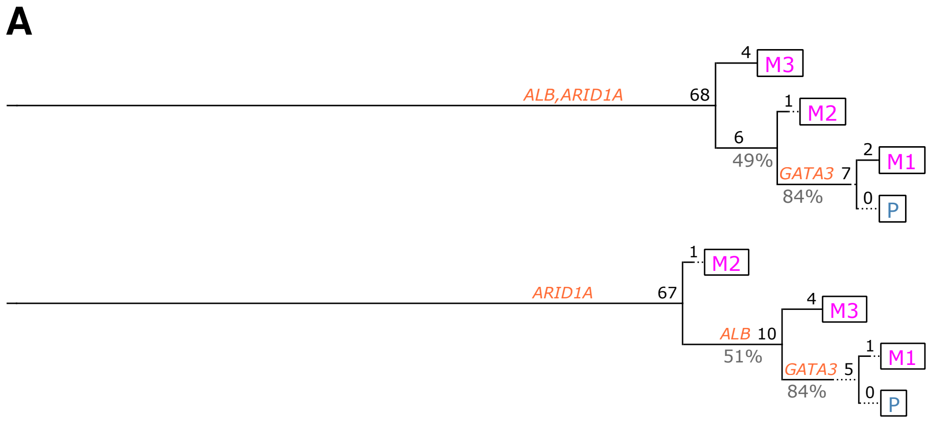
Figure S24: Met5 and Met6 of rectum cancer subject CRC3 of Kim et al. (*15*) may have acquired additional mutations in the putative driver genes *AR*, *CACNA1A*, and *IRF2*. Variants in *APC* (one frameshift and one stop gain), and *TP53* were present universally. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and agreed well with the originally reported phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: rectum: PT1-5; liver: Met1-6. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch, PT denotes primary tumor samples. Low confidence variants due to very low coverage (median below 10) are shaded gray.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*APC*	nonsense	10	n/a	n/a	n/a	high	no	no
MT	*APC*	nonsense	24	n/a	n/a	n/a	high	no	no
MT	*DNMT3A*	missense	0	1.0	0.80	0.72	moderate	no	no
MT	*ERBB3*	missense	0	1.0	-1.46	-0.00	moderate	no	no
MT	*KRAS*	missense	332	< 0.001	-2.14	3.64	moderate	yes	yes
MT	*NCOR1*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*SF1*	missense	0	1.0	n/a	-0.04	modifier	no	no
MT	*TCF7L2*	missense	4	< 0.001	-0.81	0.88	moderate	no	no
MT	*TP53*	missense	0	< 0.001	-9.71	8.69	moderate	no	yes
MB	*PIK3CG*	missense	1	1.0	-4.28	-0.55	moderate	no	no
MB	*SF3B1*	missense	0	1.0	-1.17	-0.10	moderate	no	no

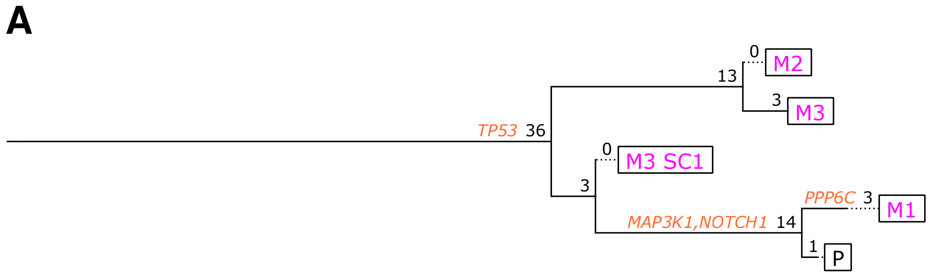
Figure S25: Nine variants in eight putative driver genes were universally present across four metastases of colon cancer subject CRC4 of Kim et al. (*15*). The variant in *SF3B1* is likely universally present but coverage in Met1 was too low to be conclusive (VAF 3.6%, 1/28 reads; estimated neoplastic cell content in Met1 34%; shaded gray). The original authors inferred the variant in *SF3B1* as ubiquitously present. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and did not agree well with the originally reported phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: sigmoid: PT1-2; liver: Met1-4. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch, PT denotes primary tumor samples.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*ALB*	missense	0	n/a	2.03	-0.10	moderate	no	no
MT	*ARID1A*	frame shift	0	n/a	n/a	n/a	high	no	no
MB	*GATA3*	frame shift	2	n/a	n/a	n/a	high	no	no

Figure S26: Variants in putative driver genes *ALB* and *ARID1A* were likely universally present in all samples of breast cancer subject 1-69 of Brown et al. (*18*). The variant in *ALB* with a VAF 5.3% supported by 1/18 reads may or may not be present in M2. The support for the absence of an additional variant in *GATA3* in samples M2 (VAF 6.2%, 1/15 reads) and M3 (VAF 4.2%, 1/23 reads) was relatively low. (**A**) Cancer phylogenies were inferred by *Treeomics* (*20*) (upper phylogeny has a higher confidence value according to Treeomics) and did only partially agree with the originally reported phylogeny perhaps because M2 was excluded from the phylogenetic reconstructed due to low neoplastic cell content. The signal to reconstruct phylogenies was rather low because only nonsynonymous mutations were available. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: breast: P (primary); mediastinal soft tissue: M1; hilar lymph node: M2; aorta M3. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch.

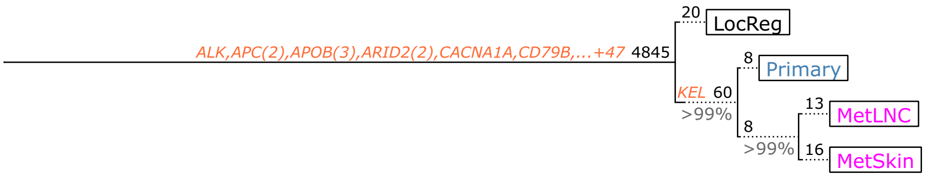


B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*TP53*	missense	32	< 0.001	-9.93	19.30	moderate	yes	yes
MB	*MAP3K1*	missense	0	< 0.001	1.11	5.38	moderate	no	no
MB	*NOTCH1*	missense	0	1.0	4.06	0.40	moderate	no	no
MB	*PPP6C*	missense	0	1.0	0.50	-0.27	moderate	no	no

Figure S27: Variant in putative driver gene *TP53* was universially present in all samples of breast cancer subject 2-57 of Brown et al. (*18*). Additional variants in *MAP3K1* (supported by 2/3 reads), *NOTCH1*, and *PPP6C* (supported by 3/19 reads) were only present in a subset of metastases samples. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and agreed well with the originally reported phylogeny. However, although *Treeomics* also detected horizontal cross-seeding, it reported a migrating subclone in M3 and not in M2. M2 seems unlikely to be a decent of an ancestor of P and M1 as there are fourteen variants (including two in putative driver genes *MAP3K1* and *NOTCH1*) present in P and M1 (with a mean VAF of 41%) but absent in M2. More likely an ancestor of M2 and an ancestor of P and M1 independently gave rise to M3. Importantly, the identified subclones shared the same putative driver gene mutations. These inferred scenarios are based on missense and frame-shift mutations only as the publicly available data was prefiltered. The estimated neoplastic cell content in M2 (33.6%) was less than half than in the other three samples and could therefore influence the inferred phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: breast: P (primary); liver: M1; adrenal gland: M2; ovarium: M3. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch. Low confidence variants are shaded gray.

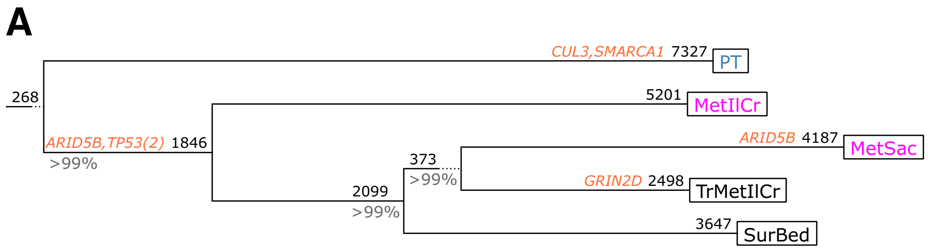
A



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*KEL*	missense	0	1.0	0.76	-0.33	moderate	no	no

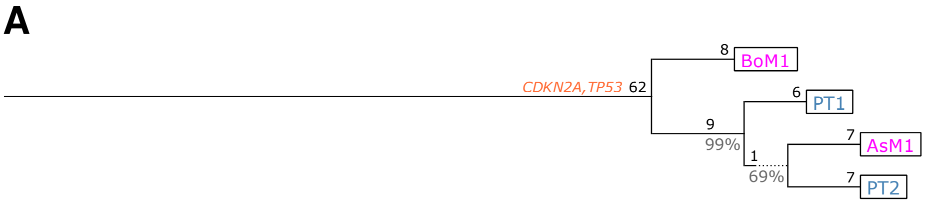
Figure S28: No driver gene mutation heterogeneity between a lymph node (MetLNC) and a back metastasis (MetSkin) of melanoma cancer subject F of Sanborn et al. (*13*). (**A**) 58 point mutations in 46 putative driver genes were concordant among the two metastases samples. Cancer phylogeny was inferred by *Treeomics* (*20*) and was highly similar to the originally reported phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: ear: Primary, LocReg (locoregional); cervical lymph node: MetLNC; back skin: MetSkin. (**B**) Predicted functional effects of variants in putative driver genes that were not present in all samples. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all lymphatic and distant metastases. Functional analysis of the remaining 57 variants is not depicted as those were present in every sample.



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*ARID5B*	missense	0	1.0	0.81	0.12	moderate	no	no
MT	*TP53*	missense	174	< 0.001	-9.17	4.97	moderate	yes	yes
MT	*TP53*	missense	0	0.009	-9.43	4.92	moderate	no	no
MB	*ARID5B*	exonic splice	0	1.0	0.74	-0.15	moderate	no	no
PT	*CUL3*	missense	0	0.56	2.29	-0.27	moderate	no	no
PT	*SMARCA1*	nonsense	0	n/a	n/a	n/a	high	no	no
Tr	*GRIN2D*	missense	0	n/a	2.42	-0.25	moderate	no	no

Figure S29: Possible driver gene heterogeneity between two metastases of prostate cancer subject 498 of Hong et al. (*14*). Given that the primary tumor sample does not share a single driver gene mutations with the other samples, we agree with the original authors' hypothesis that the metastases were seeded from an independent cancer. The low fraction of variants along the trunk is also indicative of this hypothesis. Three variants in the putative driver genes *ARID5B* and *TP53* (2 distinct missense mutations) were present in all metastases samples. An additional exonic splice-site variants was acquired in *ARID5B*. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*) and is equivalent to the originally reported phylogeny. Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: prostate: PT (primary tumor), SurBed (surgical bed); sacral: MetSac (untreated); iliac crest: MetIlCr (untreated), TrMetIlCr (treated). (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch, PT denotes primary tumor samples, Tr denotes treated samples.

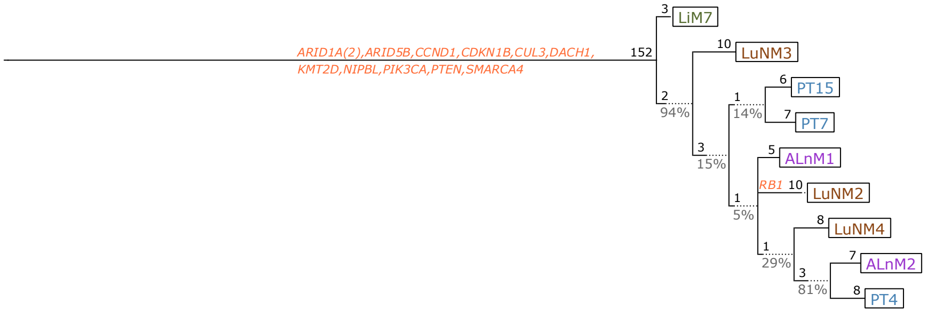


B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*CDKN2A*	missense	0	0.005	n/a	6.97	moderate	yes	yes
MT	*TP53*	missense	0	$<$ 0.001	n/a	30.59	moderate	no	no

Figure S30: No driver gene mutation heterogeneity between a bone (BoM1) and a ascites metastasis (AsM1) of gastric adenocarcinoma cancer subject C1-11 of Pectasides et al. (*19*). A missense mutation in *CDKN2A* and a missense mutation in *TP53* were universally present. The original authors also reported no driver gene mutation heterogeneity. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: stomach: PT1, PT2; bone: BoM1; ascites: AsM1. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases.

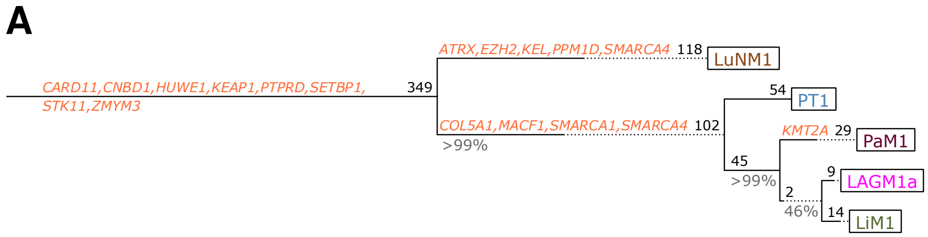
A



B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*ARID1A*	missense	0	0.36	4.87	1.08	moderate	no	no
MT	*ARID1A*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*ARID5B*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*CCND1*	deletion	2	n/a	n/a	n/a	moderate	no	no
MT	*CDKN1B*	nonsense	0	n/a	n/a	n/a	high	no	no
MT	*CUL3*	frame shift	0	n/a	n/a	n/a	high	no	no
MT	*DACH1*	splice donor	0	n/a	n/a	n/a	high	no	no
MT	*KMT2D*	missense	0	1.0	2.04	0.16	moderate	no	no
MT	*NIPBL*	splice acceptor	0	n/a	n/a	n/a	high	no	no
MT	*PIK3CA*	missense	95	$<$ 0.001	-4.35	2.67	moderate	yes	yes
MT	*PTEN*	missense	36	$<$ 0.001	-5.84	3.21	moderate	yes	yes
MT	*SMARCA4*	missense	0	1.0	-0.27	0.05	moderate	no	no
MB	*RB1*	intronic splice	0	n/a	n/a	n/a	low	no	no

Figure S31: Very unlikely functional driver gene heterogeneity among three lung nodules, a liver metastasis, and two aortocaval lymph node metastases of endometrial cancer subject MSKA1. Twelve variants in putative driver genes were universally present. An intronic splice site variant in putative driver gene *RB1* was present at a VAF of 5.8% in one sample but also at VAFs between 1.9% and 4.1% in all other metastases samples. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: uterine: PT4, PT7, PT15; lung: LuNM2, LuNM3, LuNM4 (all nodules); liver: LiM7; aortocaval lymph nodes: ALnM1, ALnM2. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch.


A
CARD11,CNBD1,HUWE1,KEAP1,PTPRD,SETBP1,
STK11,ZMYM3
349
ATRX,EZH2,KEL,PPM1D,SMARCA4
118
LuNM1
COL5A1,MACF1,SMARCA1,SMARCA4
102
>99%
54
PT1
KMT2A
29
PaM1
45
>99%
9
LAGM1a
2
46%
14
LiM1

B

	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	CanDrA score	VEP	Cancer hotspot	OncoKB
MT	*CARD11*	missense	0	1.0	4.31	-0.10	moderate	no	no
MT	*CNBD1*	missense	0	1.0	1.20	0.65	moderate	no	no
MT	*HUWE1*	nonsense	0	n/a	n/a	n/a	high	no	no
MT	*KEAP1*	missense	1	0.06	-3.69	1.15	moderate	no	no
MT	*PTPRD*	missense	12	1.0	-3.16	-0.09	moderate	no	no
MT	*SETBP1*	missense	1	1.0	2.31	-0.12	moderate	no	no
MT	*STK11*	missense	2	0.002	-0.68	1.25	moderate	no	no
MT	*ZMYM3*	intronic splice	0	n/a	n/a	n/a	low	no	no
MB	*ATRX*	missense	0	1.0	0.31	-0.18	moderate	no	no
MB	*COL5A1*	missense	0	1.0	2.93	0.19	moderate	no	no
MB	*EZH2*	missense	0	1.0	1.96	0.10	moderate	no	no
MB	*KEL*	missense	0	1.0	2.38	-0.21	moderate	no	no
MB	*KMT2A*	deletion	0	n/a	n/a	n/a	moderate	no	no
MB	*MACF1*	intronic splice	0	n/a	n/a	n/a	low	no	no
MB	*PPM1D*	missense	1	1.0	0.66	0.17	moderate	no	no
MB	*SMARCA1*	missense	0	1.0	-0.38	-0.31	moderate	no	no
MB	*SMARCA4*	missense	6	0.004	-1.86	0.10	moderate	no	no
MB	*SMARCA4*	missense	0	0.044	0.21	0.33	moderate	no	no

Figure S32: Elevated mutation rate in lung cancer patient MSKA2 might have lead to functional driver gene heterogeneity among four untreated metastases. Eight out of eighteen variants in putative driver genes were universally present. Two distinct missense mutations in *SMARCA4* were acquired on different branches. (**A**) Cancer phylogeny was inferred by *Treeomics* (*20*). Nonsynonymous mutations in putative driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Sample origin: right lung: PT1; lung nodule: LuNM1; paravertebral: PaM1; liver: LiM1; left adrenal gland mass: LAGM1a. (**B**) Predicted functional effects of variants in putative driver genes. Orange shading corresponds to likely functional consequences, green to unlikely functional consequences. MT (MetTrunk) denotes that the variant was acquired on the trunk of all metastases, MB (MetBranch) denotes a metastases branch.

Table S1. Patient cohort and sample overview.

Sequencing data of 20 subjects with samples from 76 distinct and untreated metastases meeting the selection criteria.

Publication	Cancer type	Passed #subjects	#samples	#untreated metastases samples
Brown et al. (*18*)	breast	2	8	6
Gibson et al. (*16*)	endometrial	4	13	10
Hong et al. (*14*)	prostate	1	5	2
Kim et al. (*15*)	colorectal	3	24	13
Makohon-Moore et al. (*17*)	pancreatic	6	43	31
Pectasides et al. (*19*)	gastric	1	4	2
Sanborn et al. (*13*)	melanoma	1	4	2
new data	endometrial	1	9	6
new data	lung	1	5	4
	total	**20**	**115**	**76**

References and Notes

1. P. C. Nowell, The clonal evolution of tumor cell populations. *Science*. **194**, 23–28 (1976).
2. E. R. Fearon, B. Vogelstein, A genetic model for colorectal tumorigenesis. *Cell*. **61**, 759–767 (1990).
3. R. H. Hruban, M. Goggins, J. Parsons, S. E. Kern, Progression model for pancreatic cancer. *Clin. Cancer Res.* **6**, 2969–2972 (2000).
4. M. Greaves, C. C. Maley, Clonal evolution in cancer. *Nature*. **481**, 306–313 (2012).
5. M. Jamal-Hanjani *et al.*, Tracking the evolution of non--small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
6. M. Gerlinger *et al.*, Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
7. A. Sottoriva *et al.*, A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
8. S. Turajlic, C. Swanton, Metastasis as an evolutionary process. *Science*. **352**, 169–175 (2016).
9. J. Massagué, A. C. Obenauf, Metastatic colonization by circulating tumour cells. *Nature*. **529**, 298–306 (2016).
10. M. H. Bailey *et al.*, Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. **173**, 371–385.e18 (2018).
11. F. Blokzijl *et al.*, Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*. **538**, 260–264 (2016).
12. B. Vogelstein *et al.*, Cancer Genome Landscapes. *Science*. **339**, 1546–1558 (2013).
13. J. Z. Sanborn *et al.*, Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc Natl Acad Sci USA*. **112**, 10995–11000 (2015).
14. M. K. H. Hong *et al.*, Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun*. **6**, 6605 (2015).
15. T.-M. Kim *et al.*, Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clin. Cancer Res.* **21**, 4461–4472 (2015).
16. W. J. Gibson *et al.*, The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat Genet*. **48**, 848–855 (2016).
17. A. P. Makohon-Moore *et al.*, Limited heterogeneity of known driver gene mutations among the metastases of individual pancreatic cancer patients. *Nat. Genet.* **49**, 358–366 (2017).
18. D. Brown *et al.*, Phylogenetic analysis of metastatic progression in breast cancer

using somatic mutations and copy number aberrations. *Nat. Commun.* **8**, 14944 (2017).

19. E. Pectasides *et al.*, Genomic heterogeneity as a barrier to precision medicine in gastroesophageal adenocarcinoma. *Cancer Discov.* **8**, 37–48 (2018).

20. J. G. Reiter *et al.*, Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* **8**, 14114 (2017).

21. H. A. Shihab, J. Gough, D. N. Cooper, I. N. M. Day, T. R. Gaunt, Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*. **29**, 1504–1510 (2013).

22. C. Tokheim, R. Karchin, Enhanced context reveals the scope of somatic missense mutations driving human cancers. *bioRxiv*, 313296 (2018).

23. H. Haeno *et al.*, Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies. *Cell*. **148**, 362–375 (2012).

24. R. Durrett, *Branching process models of cancer* (Springer, 2015).

25. I. Bozic *et al.*, Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci.* **107**, 18545 (2010).

26. H. Furukawa, R. Iwata, N. Moriyama, Growth rate of pancreatic adenocarcinoma: initial clinical experience. *Pancreas*. **22**, 366–369 (2001).

27. I. Bozic, J. M. Gerold, M. A. Nowak, Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLoS Comput. Biol.* **12**, e1004731 (2016).

28. M. J. Williams *et al.*, Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).

29. B. H. Good, M. J. McDonald, J. E. Barrick, R. E. Lenski, M. M. Desai, The dynamics of molecular evolution over 60,000 generations. *Nature*. **551**, 45–50 (2017).

30. J.-Y. Lee *et al.*, Tumor evolution and intratumor heterogeneity of an epithelial ovarian cancer investigated using next-generation sequencing. *BMC Cancer*. **15**, 85 (2015).

31. A. Rubinsteyn *et al.*, hammerlab/pyensembl: Version 1.0.1 (2016), , doi:10.5281/zenodo.154747.

32. A. Rubinsteyn *et al.*, varcode v0.4.15 (2016), , doi:10.5281/zenodo.58031.

33. C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, R. Karchin, Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci USA*. **113**, 14330–-14335 (2016).

34. S. Forbes *et al.*, COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr. Protoc. Hum. Genet.*, 1–37 (2016).

35. M. T. Chang *et al.*, Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov.* **8**, 174–183 (2018).

36. D. Chakravarty *et al.*, OncoKB: a precision oncology knowledge base. *JCO*

Precis. Oncol. **1**, 1–16 (2017).

37. A. McPherson *et al.*, Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* (2016), doi:10.1038/ng.3573.

38. W. McLaren *et al.*, The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

39. D. L. Masica *et al.*, CRAVAT 4: cancer-related analysis of variants toolkit. *Cancer Res.* **77**, e35--e38 (2017).

40. Y. Mao *et al.*, CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS One*. **8**, e77945 (2013).

41. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods*. **7**, 248–249 (2010).

42. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).

43. D. Wodarz, N. L. Komarova, *Computational Biology of Cancer: Lecture Notes and Mathematical Modeling* (World Scientific Pub Co Inc, 2005).

44. N. Beerenwinkel, R. F. Schwarz, M. Gerstung, F. Markowetz, Cancer evolution: mathematical models and computational inference. *Syst. Biol.* **64**, e1--e25 (2015).

45. P. M. Altrock, L. L. Liu, F. Michor, The mathematics of cancer: integrating quantitative models. *Nat. Rev. Cancer*. **15**, 730–745 (2015).

46. U. Del Monte, Does the cell number 10^9 still really fit one gram of tumor tissue? *Cell Cycle*. **8**, 505–506 (2009).

47. C. C. Maley *et al.*, Classifying the evolutionary and ecological features of neoplasms. *Nat Rev Cancer*. **17**, 605--619 (2017).

48. R. Durrett, J. Foo, K. Leder, J. Mayberry, F. Michor, Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*. **188**, 461–477 (2011).

49. V. Almendro *et al.*, Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep.* **6**, 514–527 (2014).

50. S. Jones *et al.*, Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci USA*. **105**, 4283–4288 (2008).

51. S. Yachida *et al.*, Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*. **467**, 1114–1117 (2010).

52. K. Amikura, M. Kobari, S. Matsuno, The time of occurrence of liver metastasis in carcinoma of the pancreas. *Int. J. Pancreatol.* **17**, 139–146 (1995).

53. R. Sun *et al.*, Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet*. **49**, 1015–1024 (2017).

54. C. Tomasetti, B. Vogelstein, G. Parmigiani, Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl*

Acad Sci USA. **110**, 1999–2004 (2013).

55. J. Ma *et al.*, The infinite sites model of genome evolution. *Proc Natl Acad Sci USA*. **105**, 14254–14261 (2008).

56. Q. Zheng, Progress of a half century in the study of the Luria--Delbrück distribution. *Math. Biosci.* **162**, 1–32 (1999).