Supplementary Note

Makohon-Moore, A. P, Zhang, M., Reiter, J. G., Bozic, I., Allen, B., Kundu, D., Chatterjee, K., Wong, F., Jiao, Y., Kohutek, Z. A., Hong, J., Attiyeh, M., McMahon, B., Wood, L. D., Hruban, R. H., Nowak, M. A., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Iacobuzio-Donahue, C. A.

1 Summary

To quantify the genetic similarities and differences among metastases, we use the "Jaccard similarity coefficient" defined as the fraction of mutations shared by two given cells, out of all somatic mutations in either of them. The similarity coefficient considers only somatic mutations. Thus cells taken from identical twins have a similarity coefficient of 0 by definition (because they do not share any somatic mutations). At the opposite extreme, two cells that are identical at every nucleotide have a similarity coefficient of 1.

For reference points to the similarity coefficients among tumors, we calculate theoretical values of this similarity coefficient for two cells randomly sampled from the same organ in a single individual. In contrast to the observed homogeneity in tumors, we find that the expected similarity coefficient in healthy tissue is always below 0.2 (Fig. 2b). These values depend on whether or not the organ tissue is self-renewing. We consider three scenarios (Fig. A1):

Scenario 1: Non-self-renewing tissue. First we consider an organ that has grown to size N via a pure birth branching process. With each cell division, each daughter cell has a Poisson-distributed number of somatic mutations with mean u. In this case, the expected fraction of shared mutations is approximately $1/\log_2(N)$ for large N. Thus for $N = 10^{10}$, the expected similarity coefficient is about 0.03. Figure A2 shows the expected value of this coefficient as a function of the organ size N. The expected similarity

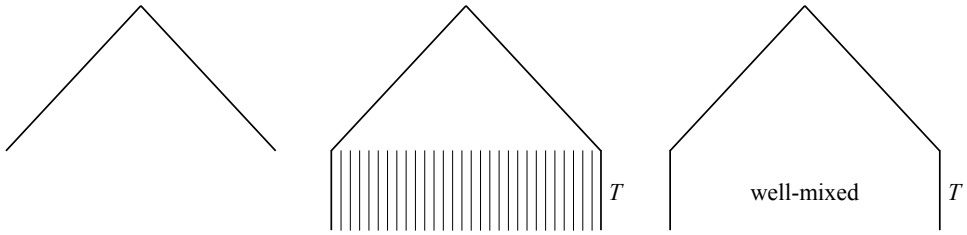

Figure A1: Schematic illustration of the three scenarios for somatic evolution in an organ. (a) In Scenario 1, an organ grows according to a pure birth process up to size N and no further cell division occurs. (b) In Scenario 2, a pure birth process leads to N_{crypt} stem cells, each of which founds a single crypt. Cells in different crypts do not replace each other. (c) In Scenario 3, a pure birth process leads to N_{stem} stem cells, which replace each other according to the Moran model of a well-mixed population.

coefficient is less than $\ln 4 - 1 \approx 0.39$ for any population size N. For relevant population sizes ($N > 100$), the expected similarity coefficient is always below 0.2 (Fig. A2).

Scenario 2: Self-renewing tissue with spatially segregated stem cells. Second we consider an organ such as the small intestine or colon, whose tissue is divided into crypts, with a small number of organ-specific stem cells per crypt. To model this situation we suppose that a pure birth process leads to some number of cells (N_{crypt}), each of which is the initial stem cell that founds a crypt. Genetic evolution then occurs separately in each crypt. We note that the similarity coefficient can only decrease from the time that the crypts are initiated, since they each acquire somatic mutations separately. Thus the expected similarity coefficient is bounded above by $1/\log_2(N_{\text{crypt}})$, as long as N_{crypt} is large. For example, if $N_{\text{crypt}} = 10^7$ then the expected similarity coefficient is less than 0.04.

Scenario 3: Self-renewing tissue with a well-mixed stem cell population. Finally we consider a cell population such as hematopoietic cells which is maintained by a subpopulation of N_{stem} stem cells. We assume the

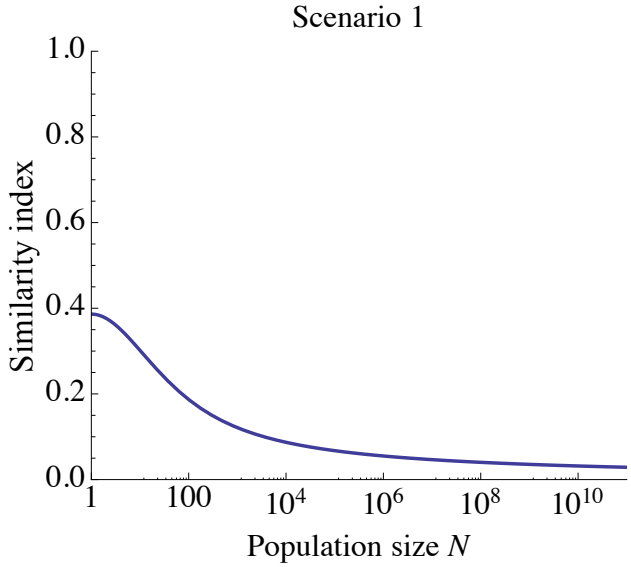

Figure A2: Similarity coefficient (expected fraction of shared mutations) for two cells sampled from a single organ in a single individual under Scenario 1. The population is exponentially growing, according to a pure-birth branching process of rate 1, up to size N. This fraction is bounded above by $\ln 4 - 1 \approx 0.39$, while for large N it is approximately $1/(\log_2 N)$. For relevant population sizes ($N > 100$), the expected similarity coefficient is always below 0.2.

stem cells replace each other according to the Moran model of a well-mixed population. We find that, for reasonable numbers of generations T, the similarity coefficient is less than what it would be for a pure birth process alone, which is approximately $1/\log_2(N_{\text{stem}})$.

Note that all three models quantify the homogeneity among stem-cell-like cells. Accounting for possibly additional mutations in short-lived terminally differentiated cells would further increase the heterogeneity within an organ.

Genetic distance. As a measurement for "genetic divergence", we also calculate the expected genetic distance for two cells randomly sampled from the same organ of an individual (Maley et al., 2006). The somatic genetic distance between two cells is defined as the total number of nonshared genetic mutations present in two cells. For this characterization we focus on dividing tissue (Scenario 2). We find that the expected genetic distance across the

exome of two random normal cells is approximately 140 (Section 6).

Last, we calculated confidence intervals for the observed similarity coefficients and genetic distances. We find that for reasonable parameter values the observed similarity coefficient often underestimates the true coefficient (Section 7).

2 Modeling heterogeneity in normal tissue

The Jaccard similarity coefficient (also known as Jaccard index) is defined as the fraction of mutations shared by two cells, out of the total mutations in either of them: $M_{\text{shared}}/(M_{\text{shared}} + M_{\text{nonshared}})$. To obtain theoretical values of this coefficient for cells sampled from non-cancerous tissue, we use the coalescent perspective. We consider the lineages of the two cells in question starting from conception. The lineages remain together for some amount of time τ_1, and then are separate for another amount of time τ_2 (see Fig. A3). τ_1 and τ_2 are random variables whose distributions depend on the scenario considered. Time is scaled so that cells divide at rate 1.

Overall, the lineages of the two cells have branch length τ_1 in common out of a total (shared and non-shared) branch length of $\tau_1 + 2\tau_2$. We assume that the number of mutations on a branch is Poisson distributed with rate proportional to branch length. With this assumption, and given specific values of τ_1 and τ_2, the expected fraction of shared mutations (similarity coefficient) is

$$\mathbb{E}\left[\frac{M_{\text{shared}}}{M_{\text{shared}} + M_{\text{nonshared}}}\right] = \frac{\tau_1}{\tau_1 + 2\tau_2} \tag{A1}$$

Note that this expected fraction does not depend on the mutation rate.

Additionally, if we also condition on a particular value m for the total number of mutations, i.e., $M_{\text{shared}} + M_{\text{nonshared}} = m$, then the number of shared mutations is binomially distributed:

$$M_{\text{shared}} \sim \text{Binom}\left[m, \frac{\tau_1}{\tau_1 + 2\tau_2}\right]. \tag{A2}$$

3 Scenario 1: Pure birth

Scenario 1 describes an organ that grows to full size, at which point cell division ceases. We assume this growth occurs via a pure birth branching

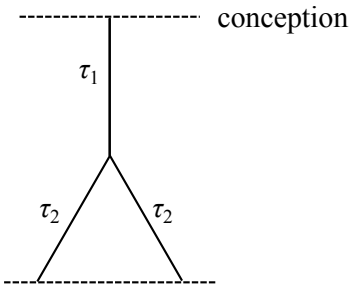

Figure A3: Theoretical values of the similarity coefficient for healthy tissue are obtained using coalescent theory. The lineages remain together for time τ_1, and then are separate for time τ_2, where τ_1 and τ_2 are random variables whose distributions depend on the process in question. The expected fraction of shared mutations is $\tau_1/(\tau_1 + 2\tau_2)$.

process that is terminated when the number of cells reaches N.

3.1 Distribution of splitting times

Using the coalescent perspective, we consider the population backwards in time as it shrinks from N cells to 1. At each step, a random pair of individuals is chosen to coalesce (meaning that they derive from the same parent in the previous step).

Consider two cells in the organ at its final size. Let $G(n)$ denote the probability that their lineages have not coalesced (are still apart) when the population size is n. Note that $G(n) - G(n-1)$ is the probability that the two lineages coalesce at the step when the population shrinks from n cells to $n-1$. This probability can be expressed as the probability $G(n)$ that coalescence has not already occurred, multiplied by the probability $\binom{n}{2}^{-1}$ that the correct pair is chosen to coalesce at this step. Thus

$$G(n) - G(n-1) = G(n)\binom{n}{2}^{-1} = \frac{2\ G(n)}{n(n-1)}. \tag{A3}$$

This recurrence relation has an exact solution:

$$G(n) = C\ \frac{n-1}{n+1}, \tag{A4}$$

valid for any constant C. Noting that $G(N) = 1$, since the two original cells are distinct, we have $C = (N+1)/(N-1)$, and thus

$$G(n) = \frac{n-1}{n+1}\frac{N+1}{N-1}. \tag{A5}$$

Interestingly, for large N, $G(n)$ is approximately independent of N:

$$G(n) \approx \frac{n-1}{n+1}. \tag{A6}$$

To get the probability of coalescence as a function of time, we approximate the total cell number by deterministic exponential growth, so that $n = e^t$ for all t. Then the probability that the lineages have split by time t is

$$F(t) \equiv \mathbb{P}[\tau_1 \leq t] \;=\; \frac{e^t - 1}{e^t + 1}\frac{N+1}{N-1}. \tag{A7}$$

This is the half-logistic distribution (Balakrishnan, 1985) conditioned on its value being less than $\ln N$. As $N \to \infty$ this converges in law to the (unconditioned) half-logistic distribution. Figure A4 shows the cumulative distribution function of τ_1 (i.e., the probability of splitting by time t) for $N = 10^7$.

The probability density of splitting at time t is

$$f(t) = F'(t) = \frac{2e^t}{(e^t+1)^2}\frac{N+1}{N-1}. \tag{A8}$$

The expected splitting time is

$$\begin{aligned}\mathbb{E}[\tau_1] &= \int_0^\infty t f(t)\, dt \\ &= \frac{N+1}{N-1}\ln 4 - \frac{2}{N-1}\big[(N+1)\ln(N+1) - N\ln N\big]\end{aligned} \tag{A9}$$

In the limit of large population size, the expected splitting time converges to $\ln 4$:

$$\lim_{N\to\infty} \mathbb{E}[\tau_1] = \ln 4. \tag{A10}$$

Thus the lineages of two randomly chosen cells in a large population are expected to split $\ln 4$ units of cell division time after the population is initialized. This quantity is nearly independent of final size of the organ. This is similar to a classic result obtained by Slatkin and Hudson (1991), except that these authors found that the expected splitting time is asymptotically equal to Euler's constant $\gamma \approx 0.577$. The difference arises because we use a pure birth process, whereas Slatkin and Hudson consider an exponentially growing variant of the Wright-Fisher process.

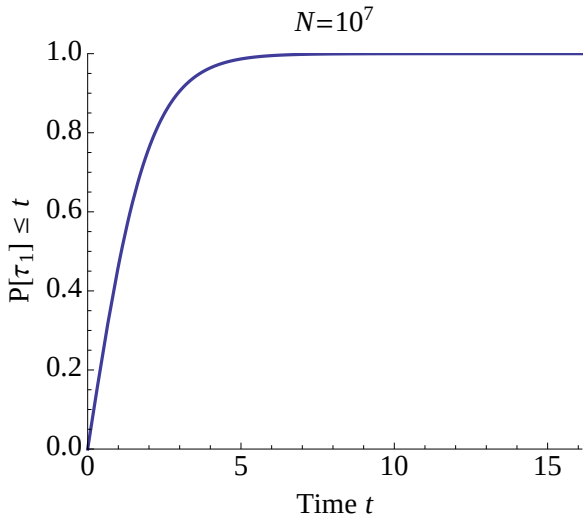

Figure A4: Cumulative distribution function $F(t)$ for the time τ_1 from conception to splitting in a pure birth process (Scenario 1), as given by Eq. (A7). This is the half-logistic distribution conditioned on its value being less than $\ln N$.

3.2 Distribution of shared mutations given the total number

We have obtained the probability distribution for the amount of time τ_1 that the lineages remain together. Since the lineages of the two considered cells diverged at time τ_1 and we approximate the total population size by exponential growth, the total time $\tau_1 + \tau_2$ is assumed to equal to $\ln N$; thus $\tau_2 = \ln N - \tau_1$.

Now suppose we know that in total m mutations are present among the two cells: $M_{\text{shared}} + M_{\text{nonshared}} = m$. Then, conditional on this event and on a particular value of the splitting time τ_1, the number of shared and nonshared mutations are binomially distributed:

$$M_{\text{shared}} \sim \text{Binom}\left[m, \frac{\tau_1}{2\ln N - \tau_1}\right]. \tag{A11}$$

Now we allow τ_1 to vary, but maintain the total number of mutations as m. The probability distribution of M_{shared} can be obtained by integrating over all possible values of τ_1:

$$\begin{aligned}\mathbb{P}[M_{\text{shared}} = k \mid M_{\text{shared}} + M_{\text{nonshared}} = m] \\ = \frac{N+1}{N-1}\binom{m}{k}\int_0^{\ln N} \frac{t^k\left[2(\ln N - t)\right]^{m-k}}{(2\ln N - t)^m}\frac{2e^t}{(e^t+1)^2}\,dt.\end{aligned} \tag{A12}$$

This integral does not evaluate to closed form, but it can be approximated numerically. Note that this distribution does not depend on the mutation rate u (because the total number of mutations is fixed).

3.3 Fraction of shared mutations

To calculate the expected value of the similarity coefficient, we evaluate the expectation of Eq. (A1) over the probability density (A8) for τ_1:

$$\mathbb{E}\left[\frac{M_{\text{shared}}}{M_{\text{shared}} + M_{\text{nonshared}}}\right] = \frac{N+1}{N-1}\int_0^{\ln N} \frac{t}{2\ln N - t}\frac{2e^t}{(e^t+1)^2}\,dt. \tag{A13}$$

Again, this integral does not evaluate to closed form. Figure A2 plots this fraction as a function of T. Interestingly, as $N \to 1$, the fraction of shared mutations approaches $\ln 4 - 1$. This quantity is an upper bound for the expected fraction of shared mutations.

4 Scenario 2: Segregated crypts

Scenario 2 describes an organ consisting of N_{crypt} spatially segregated crypts, with each crypt founded by a single stem cell. These founder cells are assumed to arise via a pure birth process. We also assume that cells in one crypt cannot be replaced by cells in another. Once the crypts develop, they evolve separately for T units of cell division time. (We always rescale time so that cell divisions occur at rate 1. Thus the time scale for when the crypts remain stable and segregated may differ from the time scale of the pure birth process).

This coalescent process is modeled exactly as in Scenario 1, except that (a) the pure birth process is terminated when the cell population reaches size N_{crypt}, and (b) the lineages of two cells from different crypts must remain apart for time T until they can coalesce. Thus the distribution of τ_1 is given by Eq. (A7) with N replaced by N_{crypt}, and $\tau_2 = T + \ln N_{\text{crypt}} - \tau_1$. From Eq. (A1) it follows that the similarity coefficient for two cells from different crypts decreases monotonically with T.

5 Scenario 3: Well-mixed stem cell pool

Scenario 3 describes an organ whose cell population is maintained by a well-mixed subpopulation of stem cells. We model this as a pure birth process followed by a Moran process. The pure birth process represents the initial development of the stem cell pool, and is terminated when the stem cell population reaches a certain size N_{stem}. At this point, the stem cell population evolves as a Moran process, which lasts for T generations (where each generation consists of N_{stem} steps of the Moran process).

Again we consider a coalescent process starting with two cells. Two possibilities arise: the lineages of these cells may coalesce during the Moran phase, or they may remain separate during the Moran phase and coalesce during the pure birth process. Coalescent theory says that coalescence during the Moran phase occurs as a Poisson process with rate $N_{\text{stem}}/2$ (in units of cell division time). Thus coalescence occurs during the Moran phase with probability $1 - e^{-2T/N_{\text{stem}}}$, and during the growth phase with probability $e^{-2T/N_{\text{stem}}}$.

5.1 Case 1: Coalescence during the Moran process

Given that coalescence occurs during the Moran process, the distribution for the coalescence time τ_2 is given by the exponential distribution with rate $N_{\text{stem}}/2$, conditioned on its value being less than T. Since the total time is $\ln N_{\text{stem}} + T$, we have $\tau_1 = \ln N_{\text{stem}} + T - \tau_2$

5.2 Case 2: Coalescence during the pure birth process

In this case the two lineages remain apart for time T, after which their coalescence follows the process described in Scenario 1. The time τ_1 that their lineages are together has probability distribution given by Eq. (A7) with N replaced by N_{stem}. Since the total time is $\ln N_{\text{stem}} + T$, we have $\tau_2 = \ln N_{\text{stem}} + T - \tau_1$.

5.3 Overall fraction of shared mutations

Combining our analyses of the two cases with Eq. (A1), we obtain the expected similarity coefficient for this scenario:

$$\begin{aligned}\mathbb{E}\left[\frac{M_{\text{shared}}}{M_{\text{shared}} + M_{\text{nonshared}}}\right] \\ = \frac{2}{N_{\text{stem}}}\int_0^T \frac{\ln N_{\text{stem}} + T - t}{\ln N_{\text{stem}} + T + t} e^{-2t/N_{\text{stem}}}\, dt \\ + e^{-2T/N_{\text{stem}}}\frac{N_{\text{stem}} + 1}{N_{\text{stem}} - 1}\int_0^{\ln N_{\text{stem}}} \frac{t}{2\ln N_{\text{stem}} + 2T - t}\frac{2e^t}{(e^t+1)^2}\, dt. \quad \text{(A14)}\end{aligned}$$

Interestingly, for fixed N, the similarity coefficient is nonmonotonic in T (Fig. A5). For $T \ll N/2$, Case 2 dominates in probability. In this case, the lineages likely split during the birth phase, with an expected splitting time of $\ln 4$. Thus increasing T only increases the amount of time that the lineages are apart in this case. For $T \gg N/2$, Case 1 dominates in probability. In this case the expected coalescence time is $N/2$, and increasing T increases the time that the lineages spend together. Thus overall the similarity coefficient first decreases and then increases in T.

In Fig. A5 we assumed $N_{\text{stem}} = 10^4$ (approximating the human hematopoietic stem cell population). Relevant values of T are below $1,000$ generations (as stem cells in the hematopoietic system divide approximately once a

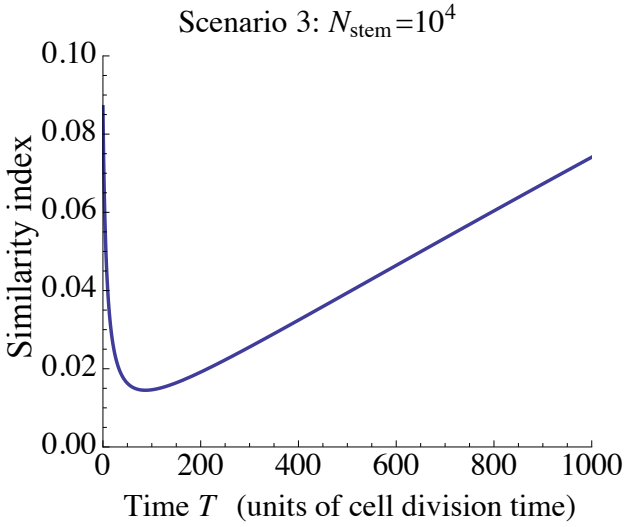

Figure A5: Similarity coefficient in Scenario 3, as a function of the number T of generations of self-replacement after all stem cells have been produced. Number of stem cells $N_{\text{stem}} = 10^4$. Relevant values of T are below $1,000$ generations, for which case the similarity coefficient is less than it would be for a pure birth process alone.

month). For such values, the similarity coefficient is less than in a pure-birth process alone. For larger stem cell populations the similarity coefficient is even smaller.

6 Scenario 2: Numbers of shared and non-shared mutations

To calculate the expected number of shared and nonshared mutations, we use the probability distributions for the amount of time τ_1 and τ_2. Both times are defined as numbers of cell divisions and hence by multiplying with a mutation rate u we directly obtain the number of acquired mutations in this period of time. The expected numbers of shared and nonshared mutations,

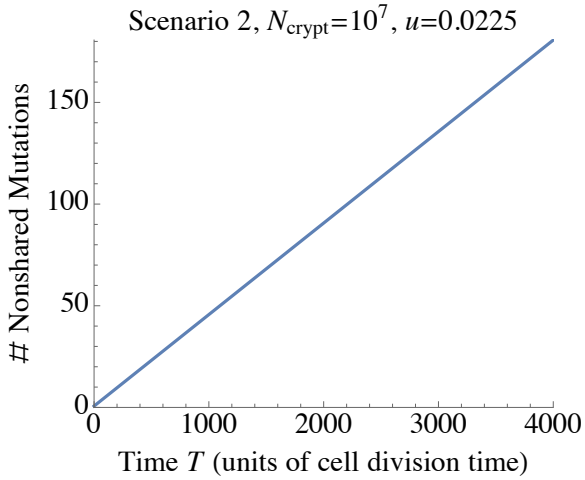

Figure A6: Expected number of nonshared mutations in Scenario 2, according to Eq. (A15). In this scenario, a pure-birth process gives rise to N_{crypt} crypts, each of which follow a self-renewal process for T generations. Here we have set $u = 0.0225$ and $N = 10^7$. For these parameters, the vast majority of mutations occur after all crypts have developed, and therefore $\mathbb{E}[M_{\text{nonshared}}] \approx 2uT$.

respectively, are given by

$$\begin{aligned} \mathbb{E}[M_{\text{shared}}] &= u\mathbb{E}[\tau_1] = u \ln 4 \\ \mathbb{E}[M_{\text{nonshared}}] &= 2u\mathbb{E}[\tau_2] = 2u(T + \ln N_{\text{crypt}} - \ln 4). \end{aligned} \tag{A15}$$

The number of nonshared mutations is shown as a function of T in Fig. A6.

To obtain comparison values for the measured genetic distance of coding mutations among the metastases, we assume a mutation rate of $u = 0.0225$ (expected number of acquired mutations across the exome per cell division assuming a point mutation rate of $5 \cdot 10^{-10}$ and 45 megabases covered by the sequencing machine) and $T = 60 \cdot 52$ (cell divides once per week for 60 years). This calculation yields an expected genetic distance of 141.1 for two random cells sampled from the same organ.

7 Confidence intervals for the similarity coefficient and genetic distance

Here we consider how the sensitivity and specificity of bulk sequencing errors may affect observed values of the similarity coefficient and genetic distance. Because we used a very conservative scheme for when a mutation is counted as observed, we consider only type II errors (false negatives), in which a mutation that is present is not observed. False positives, in which a mutation is observed but is not actually present, are sufficiently rare under our scheme so as not to significantly affect the similarity coefficient or genetic distance (see Online Methods for the calculated false positive rate).

7.1 Notation

We consider two tissue samples taken from different metastases in the same patient. We define

- A is the true set of mutations in the first sample
- B the true set of mutations in the second sample
- $m_{\text{shared}} = |A \cap B|$ is the true number of shared mutations
- $m_{\text{nonshared}} = |A \cap B^C| + |A^C \cap B|$ is the true number of nonshared mutations (superscript C denotes set complement).

Then the true similarity coefficient is

$$J = \frac{|A \cap B|}{|A \cup B|} = \frac{m_{\text{shared}}}{m_{\text{shared}} + m_{\text{nonshared}}}.$$

The true genetic distance is

$$D = |A \cap B^C| + |A^C \cap B| = m_{\text{nonshared}}.$$

We let p denote the false negative rate, so that a true mutation is observed with probability $1-p$ and otherwise (with probability p) is missed. The sets of observed mutations are denoted $\hat{A}$ and $\hat{B}$. The observed similarity coefficient is

$$\hat{J} = \frac{|\hat{A} \cap \hat{B}|}{|\hat{A} \cup \hat{B}|}.$$

The observed genetic distance is

$$\hat{D} = |\hat{A} \cap \hat{B}^C| + |\hat{A}^C \cap \hat{B}|.$$

7.2 Analysis of probabilities

We partition the total set of true mutations $A \cup B$ into three disjoint subsets: $A \cap B^C$, $A^C \cap B$, and $A \cap B$, Whether or not the mutations in these sets are observed can be considered independent events.

- For each mutation in $A \cap B^C$,
  - With probability $1-p$ it is observed and is therefore in $\left(A \cap B^C\right)$,
  - With probability p it is not observed and is not in $\hat{A} \cup \hat{B}$.
- For each mutation in $A^C \cap B$,
  - With probability $1-p$ it is observed and is therefore in $\hat{A} \cap \hat{B}^C$,
  - With probability p it is not observed and is therefore not in $\hat{A} \cup \hat{B}$.
- For each mutation in $A \cap B$,
  - With probability $(1-p)^2$ it is observed in both samples and is therefore in $\hat{A} \cap \hat{B}$
  - With probability $p(1-p)$ it is observed in the first sample but not the second, and is therefore in $\hat{A} \cap \hat{B}^C$,
  - With probability $p(1-p)$ it is observed in the second sample but not the first, and is therefore in $\hat{A}^C \cap \hat{B}$,
  - With probability p^2 it is not observed in either sample, and is therefore not in $\hat{A} \cup \hat{B}$.

We define the following random variables:

- $\hat{M}_1 = \left|\left(A \cap B^C\right) \cap \left(\hat{A} \cap \hat{B}^C\right)\right| + \left|\left(A^C \cap B\right) \cap \left(\hat{A}^C \cap \hat{B}\right)\right|$ is the number of mutations that are correctly observed to be in only one sample. $\hat{M}_1$ is distributed as $\text{Binom}[m_{\text{nonshared}}, 1-p]$.

- $\hat{M}_2 = \left|(A \cap B) \cap \left(\hat{A} \cup \hat{B}\right)\right|$ is the number of mutations in both samples that are observed at all. $\hat{M}_2$ is independent of $\hat{M}_1$ and is distributed as $\text{Binom}[m_{\text{shared}}, 1-p^2]$.
- $\hat{M}_3 = \left|(A \cap B) \cap \left(\hat{A} \cap \hat{B}\right)^C\right|$ is the number of mutations that are truly in both samples, but observed in only one sample. $\hat{M}_3$ is dependent on $\hat{M}_2$. Conditioned on $\hat{M}_2 = m$, M_3 has distribution $\text{Binom}[m, 2p(1-p)/(1-p^2)]$.
- $\hat{M}_{\text{shared}} = \left|\hat{A} \cap \hat{B}\right|$ is the number of mutations that are observed to be in both samples (correctly so, since there are no false positives). $\hat{M}_{\text{shared}}$ is dependent on $\hat{M}_2$. Conditioned on $\hat{M}_2 = m$, M_{shared} has distribution $\text{Binom}[m, (1-p)^2/(1-p^2)]$. Clearly we also have $\hat{M}_3 + \hat{M}_{\text{shared}} = \hat{M}_2$.

7.3 Probability distribution for observed similarity coefficient

The observed similarity coefficient $\hat{J}$ can be written as

$$\hat{J} = \frac{\hat{M}_{\text{shared}}}{\hat{M}_1 + \hat{M}_2} \tag{A16}$$

The cumulative distribution function (CDF) for the observed similarity coefficient is then

$$\begin{aligned}
\mathbb{P}\left[\hat{J} \le j\right] &= \sum_{m=0}^{m_{\text{shared}}} \mathbb{P}\left[\hat{M}_2 = m\right] \mathbb{P}\left[\hat{J} \le j \middle| \hat{M}_2 = m\right] \\
&= \sum_{m=0}^{m_{\text{shared}}} \binom{m_{\text{shared}}}{m} (1-p^2)^m p^{2(m_{\text{shared}}-m)} \\
&\qquad \times \mathbb{P}\left[\hat{M}_{\text{shared}} \le j(\hat{M}_1 + m) \middle| \hat{M}_2 = m\right].
\end{aligned} \tag{A17}$$

Figure A7 illustrates the CDF for two different false negative rates and number of shared and nonshared mutations of $m_{\text{shared}} = m_{\text{nonshared}} = 50$. These values for the number of mutations roughly correspond to the numbers measured in the data (Supplemental Figs. 1-4). The number of truly

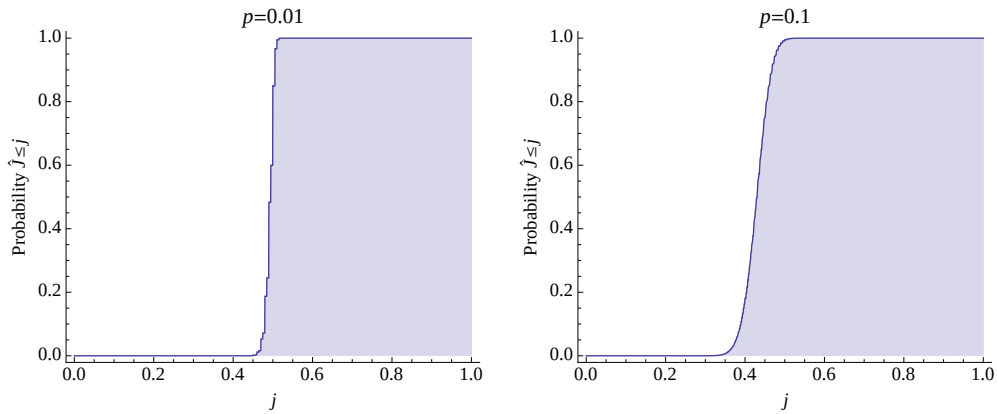

Figure A7: Cumulative distribution functions for the observed similarity coefficient $\hat{J}$ when the true mutation numbers are $m_{\text{shared}} = m_{\text{nonshared}} = 50$. The false negative rates are $p = 0.01$ for the left panel and $p = 0.1$ for the right. Note that the true similarity coefficient is 0.5 for these examples. With these parameters and $p = 0.01$, the 95% confidence interval is $0.47 \leq \hat{J} \leq 0.51$. For $p = 0.1$, the 95% confidence interval is $0.37 \leq \hat{J} \leq 0.49$. This shows that, with a false negative rate of 10%, the observed similarity coefficient is almost certainly an underestimate of its true value. Calculations were done using the Probability function in Mathematica.

nonshared mutations might be higher because bulk sequencing can not detect mutations at very low frequencies. However, since we are interested in the mutations present in the founding cells of the metastases, the observed number of mutations should be an upper bound on what was present when the metastasis was seeded. We note that the observed values of the similarity coefficient tend to be smaller than the true value (Fig. A7), because false negatives allow shared mutations to be classified as nonshared.

7.4 Probability distribution for observed genetic distance

The observed genetic distance $\hat{G}$ can be written as

$$\hat{G} = \hat{M}_1 + \hat{M}_3 \tag{A18}$$

The cumulative distribution function (CDF) for the observed similarity coefficient is then

$$\begin{aligned} \mathbb{P}\left[\hat{G} \leq g\right] &= \mathbb{P}\left[\hat{M}_1 + \hat{M}_3 \leq g\right] \\ &= \sum_{m=0}^{m_{\text{shared}}} \binom{m_{\text{shared}}}{m} (1-p^2)^m p^{2(m_{\text{shared}}-m)} \\ &\quad \times \mathbb{P}\left[\hat{M}_1 + \hat{M}_3 \leq g \middle| \hat{M}_2 = m\right]. \end{aligned} \tag{A19}$$

Figure A8 illustrates the CDF for two different false negative rates.

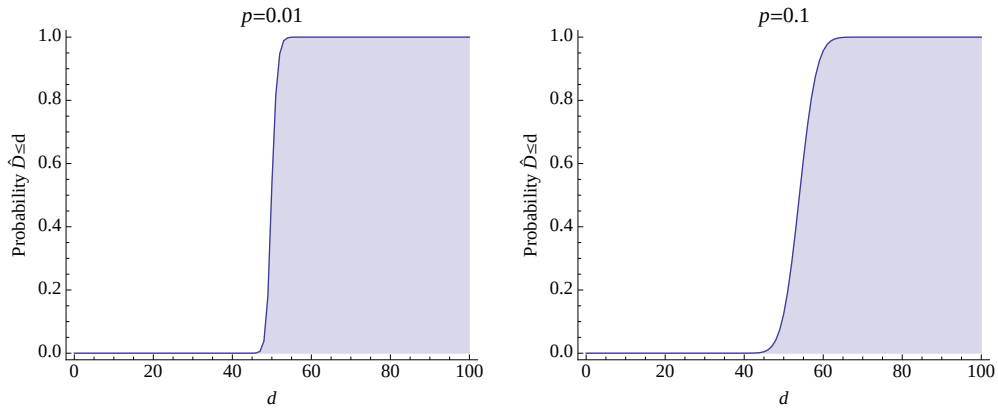

Figure A8: Cumulative distribution functions for the observed genetic distance $\hat{G}$ when the true mutation numbers are $m_{\text{shared}} = m_{\text{nonshared}} = 50$. The false negative rates are $p = 0.01$ for the left panel and $p = 0.1$ for the right. Note that the true genetic distance is 50 for these examples. With these parameters and $p = 0.01$, the 95% confidence interval is $47 \leq \hat{G} \leq 53$. For $p = 0.1$, the 95% confidence interval is $47 \leq \hat{G} \leq 62$. Calculations were done using the Probability function in Mathematica.

References

Balakrishnan, N. (1985). Order statistics from the half logistic distribution. *Journal of Statistical Computation and Simulation 20*(4), 287–309.

Maley, C. C., P. C. Galipeau, J. C. Finley, V. J. Wongsurawat, X. Li, C. A. Sanchez, T. G. Paulson, P. L. Blount, R.-A. Risques, P. S. Rabinovitch, and B. J. Reid (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics 38*(4), 468–473.

Slatkin, M. and R. R. Hudson (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics 129*(2), 555–562.