IJC
International Journal of Cancer

Quantification of multicellular colonization in tumor metastasis using exome-sequencing data

Jo Nishino[1], Shuichi Watanabe[2], Fuyuki Miya[1,3], Takashi Kamatani[1,4], Toshitaka Sugawara[2], Keith A. Boroevich[3] and Tatsuhiko Tsunoda[1,3,4,5]

[1]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo, Japan
[2]Department of Hepatobiliary and Pancreatic Surgery, Tokyo Medical and Dental University (TMDU), Tokyo, Japan
[3]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan
[4]Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan
[5]CREST, JST, Tokyo, Japan

Metastasis is a major cause of cancer-related mortality, and it is essential to understand how metastasis occurs in order to overcome it. One relevant question is the origin of a metastatic tumor cell population. Although the hypothesis of a single-cell origin for metastasis from a primary tumor has long been prevalent, several recent studies using mouse models have supported a multicellular origin of metastasis. Human bulk whole-exome sequencing (WES) studies also have demonstrated a multiple "clonal" origin of metastasis, with different mutational compositions. Specifically, there has not yet been strong research to determine how many founder cells colonize a metastatic tumor. To address this question, under the metastatic model of "single bottleneck followed by rapid growth," we developed a method to quantify the "founder cell population size" in a metastasis using paired WES data from primary and metachronous metastatic tumors. Simulation studies demonstrated the proposed method gives unbiased results with sufficient accuracy in the range of realistic settings. Applying the proposed method to real WES data from four colorectal cancer patients, all samples supported a multicellular origin of metastasis and the founder size was quantified, ranging from 3 to 17 cells. Such a wide-range of founder sizes estimated by the proposed method suggests that there are large variations in genetic similarity between primary and metastatic tumors in the same subjects, which may explain the observed (dis)similarity of drug responses between tumors.

Introduction

Metastasis is the main cause of cancer-related death. Although it is essential to understand its mechanisms and the dynamics of distant site colonization in order to properly treat it, until recently little has been known. The founder cell population size of a metastatic tumor is one of the most important parameters for metastasis dynamics, which involves the change of mutational compositions from the primary to metastatic tumors (Fig. 1). The drastic genetic changes in the metastatic tumor from the primary one, brought by the limited cell migration, that is, "bottleneck effect," might result in a difference in drug response between both tumors in the same patient.

Although the hypothesis that a metastatic tumor originates from a single tumor cell has been long prevalent,[1–3] several recent studies using mouse models of cancer have demonstrated multicellular seeding.[4–6] In humans, bulk whole-exome sequencing (WES) studies of metastatic tumors, often including primary tumors from the same individuals, demonstrated metastases to have originated from multiple clones, where a "clone" was a cluster of tumor cells belonging to the same phylogenetic clade estimated by the variant allele frequency information.[7,8] While founder "cells," but not "clones," in the metastatic tumor have another clear meaning in understanding metastatic dynamics, the quantification of multicellular colonization has not been attempted so far in human metastatic tumors.

Here, we model metastatic colonization as "single bottleneck followed by rapid growth" for tumor cell populations and propose a method to quantify the founder cell population size of a metastatic tumor using a paired WES data from the primary and metachronous metastatic tumors. This method uses the outputs from commonly used mutation callers, that is, variant allele frequencies (mutant allele counts and sequence depths), and quickly estimates the founder size unbiasedly in a realistic range. We applied our proposed method to the high-depth WES data from a study of four colorectal cancer (CRC) patients.

Additional Supporting Information may be found in the online version of this article.

Key words: metastasis, multicellular colonization, founder population size, exome sequencing

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

DOI: 10.1002/ijc.32910
History: Received 14 Jun 2019; Accepted 14 Jan 2020; Online 5 Feb 2020

Correspondence to: Tatsuhiko Tsunoda, E-mail: tsunoda@bs.s.u-tokyo.ac.jp

What's new?

The founder cell population size of a metastatic tumor is one of the most important parameters for metastasis dynamics. However, multicellular colonization has not yet been quantified in human metastatic tumors. Using the 'single bottleneck followed by rapid growth' metastatic model and whole-exome sequencing data from primary and metastatic tumors in colorectal cancer patients, this quantification method supports the multi-cellular origin of metastasis, with founder population sizes ranging from 3 to 17 cells. The wide-ranging population sizes suggest large variations in genetic similarity between primary and metastatic tumors within individual patients, possibly explaining variations in drug responses between the tumors.

Methods

Overview for quantifying founder cell population size in metastasis

We use paired WES data of a primary and metachronous metastatic tumors together with the data from normal tissue (Fig. 1*a*). The input file is composed of sequence depths, D_1 and D_2, and the mutation read counts, m_1 and m_2 for each called mutation in the primary and metastatic tumors, respectively (Table 1 and Fig. 1*b*; See Supporting Information Appendix and Supporting Information Fig. S1 for more details of the input file). When the founder population size is large, the variant allele frequencies (VAFs) for called mutations in the metastatic tumor show high similarity to those in the primary tumor (Fig. 1*c*, right). Conversely, when the number of founder cells is small, the VAFs in the primary and metastatic tumors are not so correlated (Fig. 1*c*, left). In this case, due to the severe "bottleneck effect," many variants can become extinct or have significantly higher VAFs in the metastatic tumor.

Model and estimation methods

Consider a diploid tumor cell population in a primary tumor. One somatic variant in the population has the VAF, p_1, or the cancer cell fraction (CCF), $2p_1$ (see Table 1 for notations). The models assume no recurrence mutation at the same sites and therefore the VAF is at most 0.5, $p_1 \le 0.5$. The VAF follows some distribution, $p_1 \sim f(p_1)$, as is properly assumed in the present implementation assuming a 'neutral' evolution with a high cell birth rate for tumor population[9,10] (see Implementation section in Supporting Information Appendix; and see Results section for the robustness of the assumptions). In the bulk-WES of the primary tumor, the sampled mutation read count, m_1, at the variant site with sequence depth, D_1, follows a binomial distribution with parameters, D_1 and p_1,

$$m_1 \sim Bin\left(m_1|D_1,p_1\right).$$

Metastatic colonization is modeled as follows. A single bottleneck occurs during colonization and is followed by rapid growth, so that the VAF in the full-blown metastatic tumor is the same as that in the metastatic founder. We perform WES on samples from the full-blown metastatic tumor. Then, in the WES of the metastatic tumor, the sampled mutation read count, m_2, at the variant site with sequence depth, D_2, is generated by a composite process of metastatic colonization and exome sequencing as follows:

$$m_2 \sim \sum\nolimits_{M_b=0}^{N_b} \left\{ \mathrm{Bin}\left(M_b|N_b,\, 2p_1\right) \mathrm{Bin}\left(m_2|D_2,\, p_2\right) \right\},$$

where the N_b, M_b and p_2 are the number of founder cells (founder population size) in metastatic colonization, the number of mutant cells in the N_b founder cells, and the VAF in the metastatic tumor, respectively. In the above distribution for m_2, the N_b founder cells are assumed to be randomly selected from the primary tumor and colonize a metastatic site. Thus, the M_b mutant cells in the metastatic site follows a binomial distribution with parameters N_b and $2p_1$ (mutant cell fraction), where p_2 is given by $p_2 = \frac{M_b}{2N_b}$. In the bulk-WES of the metastatic tumor, the sampled mutation read count, m_2, follows a binomial distribution with parameter D_2 and p_2.

Taken together, the probability of observing m_1 and m_2 mutations in the primary and metastatic exome with depths D_1 and D_2, respectively, is given by

$$\int_{p_1=0}^{1} f\left(p_1\right) \mathrm{Bin}\left(m_1|D_1,p_1\right) \sum\nolimits_{M_b=0}^{N_b} \left\{ \mathrm{Bin}\left(M_b|N_b,\, 2p_1\right) \mathrm{Bin}\left(m_2|D_2,\, \frac{M_b}{2N_b}\right) \right\} dp_1.$$

For quality control, we use only the sites with $m_{1(\min)}$ (>0) or more mutant reads in the primary tumor. Note that, in the metastatic tumor, all mutations called in the primary tumor are tracked in order to use greater information on VAF change from the primary to the metastatic tumor. Finally, the probability of observing $m_1(\ge m_{1(\min)})$ and $m_2(\ge 0)$ mutation reads in the primary and metastatic tumors, respectively, is expressed as

$$\frac{\int_{p_1=0}^{1} f\left(p_1\right) \mathrm{Bin}\left(m_1|D_1,p_1\right) \sum_{M_b=0}^{N_b} \left\{ \mathrm{Bin}\left(M_b|N_b,\, 2p_1\right) \mathrm{Bin}\left(m_2|D_2,\, \frac{M_b}{2N_b}\right) \right\} dp_1}{\sum_{m_1'=m_{1(\min)}}^{D_{1i}} \int_{p_1=0}^{1} f\left(p_1\right) \mathrm{Bin}\left(m_1'|D_1,p_1\right) dp_1},$$

where m_1' is possible read counts in the primary tumor. Explicitly, let p_{1i}, D_{1i}, m_{1i}, D_{2i} and m_{2i} denote p_1, D_1, m_1, D_2

and m_2 for the specific ith variant site, respectively. Assuming independencies among all R variants, each with $m_{1i}(\geq m_{1(\min)})$ mutation reads in the primary tumor, the likelihood of the founder size, N_b, is given by

By maximizing the likelihood (1), we obtain the maximum likelihood estimate (MLE) of N_b (for implementation details, see Supporting Information Appendix). In reality, the independence assumption among variants does not hold since the

$$\text{Likelihood}(N_b) = \prod_{i \in R} \frac{\int_{p_{1i}=0}^{1} f(p_{1i}) \text{Bin}(m_{1i}|D_{1i}, p_{1i}) \sum_{M_b=0}^{N_b} \left\{ \text{Bin}(M_b|N_b, 2p_{1i}) \text{Bin}\left(m_{2i}|D_{2i}, \frac{M_b}{2N_b}\right) \right\} dp_{1i}}{\sum_{m'_{1i}=m_{1(\min)}}^{D_{1i}} \int_{p_{1i}=0}^{1} f(p_{1i}) \text{Bin}(m'_{1i}|D_{1i}, p_{1i}) dp_{1i}}. \quad (1)$$

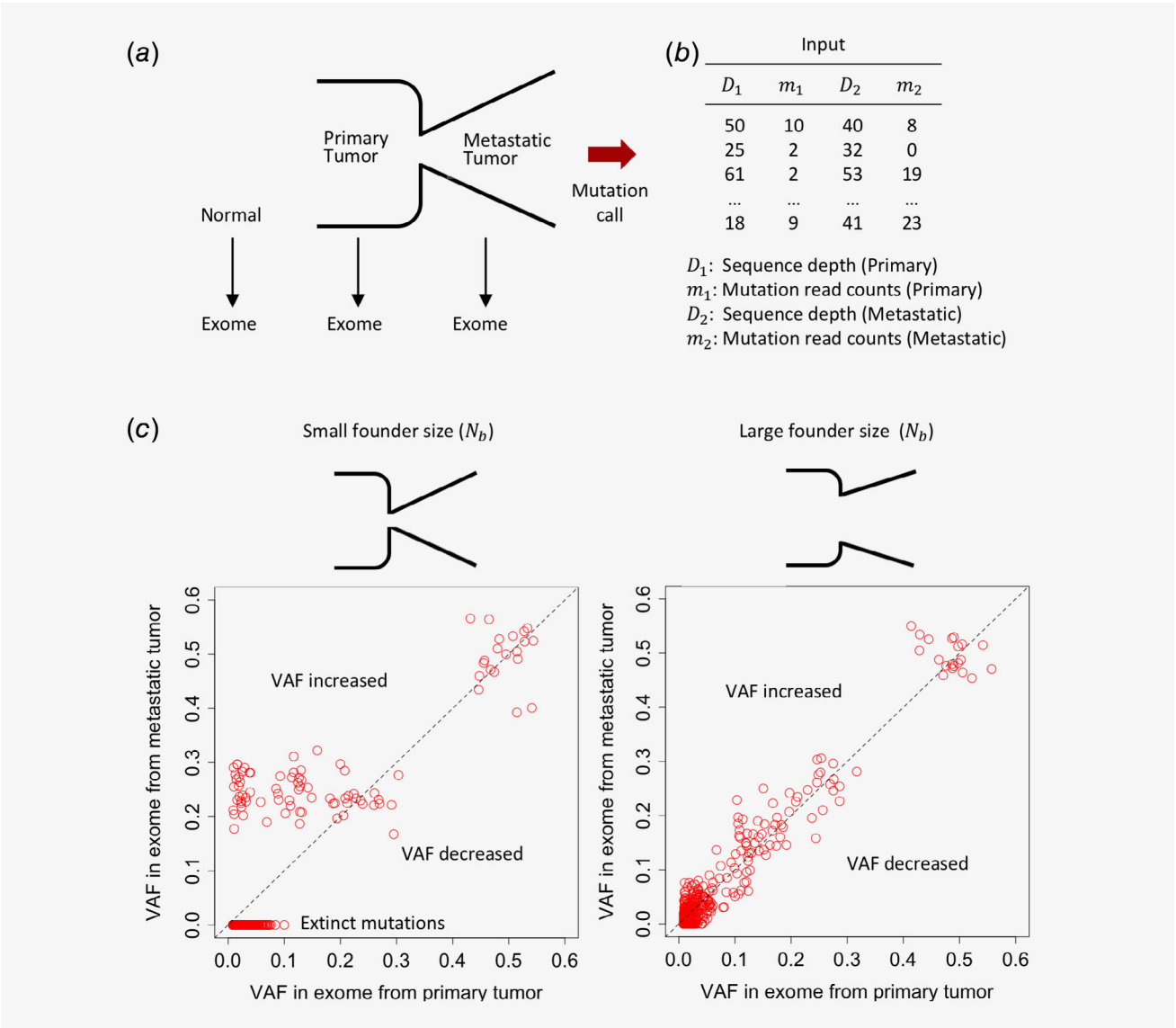

Figure 1. A schematic view of the proposed methodology. (*a*) Exome data from paired primary and metastatic tumors, and normal tissue. (*b*) Input of the method. (*c*) Illustration of basic premise for the estimation of founder sizes by computer simulations. Low correlation of observed VAFs in exome between the primary and the metastatic tumors in the small founder size, $N_b = 2$ (left). High correlation of observed VAFs between the primary and metastatic tumors in the large founder size, $N_b = 50$ (right).

Table 1. Notations in the model and the simulation study

Notation	Description
N_b	Founder cell population size, to be estimated.
R	Number of mutations used for estimation of N_b.
m_1, m_2	Mutation read counts for the primary (m_1) and metastatic tumors (m_2) at a site.
$m_{1(\min)}$	Minimum mutation read count in WES data from the primary tumor. For estimating N_b, we use only the sites with $m_{1(\min)}$ or more mutant reads.
D_1, D_2	Sequence depths for the primary (D_1) and metastatic tumors (D_2) at a site.
p_1, p_2	Population VAFs in the primary (p_1) and metastatic tumor (p_2) at a site.
$f(p_1)$	Probability distribution of p_1.
M_b	Number of mutant cells among N_b founders.
γ_1, γ_2	Tumor purity in the WES samples from the primary (γ_1) and metastatic tumors (γ_2).
Additional notations in the simulation study	
K	Number of clonal mutations inherited from the initial primary tumor.
μ	Mutation rate per tumor-cell division in the primary tumor.
N_1	Cell population size in the final primary tumor.
$\bar{D}$	Mean sequence depth in the primary and metastatic tumor.
γ	Tumor purity in the WES samples from the primary and metastatic tumors ($\gamma_1 = \gamma_2$).
Simulation for selection in the primary tumor	
b	Birth rate of cells in the primary tumor.
d	Death rate of cells in the primary tumor.
$N_{\text{sub. occ.}}$	Primary tumor size at which one advantageous mutation occurs.
a	Coefficient for birth rate. Birth rate of a cell with k non-neutral mutations is $(1 + a)^k$.
Simulation for selective colonization	
$p_{s_{\text{met}}}$	Proportion of mutations with advantage in metastatic colonization
s_{met}	Coefficient for ability of metastatic colonization ($s_{\text{met}} > 0$). Ability of metastatic colonization of a cell with l advantageous mutations is $(1 + s_{\text{met}})^l$.
Simulation for stochastic evolution of metastatic tumor	
b_{met}	Birth rate of cells in the metastatic tumor.
d_{met}	Death rate of cells in the metastatic tumor.
N_{met}	Cell population size in the final metastatic tumor.

unit of the tumor evolution is the cell, and mutations in the same cell evolve and colonize a metastatic site together. The effect of the independence assumption on the estimation of N_b is investigated below using simulations.

The tumor purities, the fraction of cancer cells, in the primary (γ_1) and metastatic tumor tissue samples (γ_2), are incorporated into the model simply by replacing p_{1i} in the term $\text{Bin}(m_{1i}|\, D_{1i}, p_{1i})$ and $\text{Bin}(m'_{1i}|\, D_{1i}, p_{1i})$ with $\gamma_1 p_{1i}$, and $\frac{M_b}{2N_b}$ in the term $\text{Bin}\left(m_{2i}|D_{2i}, \frac{M_b}{2N_b}\right)$ with $\gamma_2 \frac{M_b}{2N_b}$, respectively.

Data availability

The data that support the findings of our study were derived from Supporting Information Tables S5–S12 in the reference 8. The modified data will be made available upon reasonable request. The code to estimate the founder size used for the study is available from the github repository: https://github.com/jonishino/MetaCellNum.

Results

Validation of our proposed method: pure birth tumor evolution model

We assessed our proposed method using simulated data, generated by a "pure birth model" for tumor evolution (see Methods and Supporting Information Appendix for details; see also Table 1 for notations). Briefly, a single tumor cell with K mutations generates two daughter cells, each with average μ new mutations, and cell divisions repeat until the population has grown to the final primary tumor size, N_1. N_b cells are randomly sampled from the N_1 cells to make up a metastatic tumor. Note that the above K mutations result in clonal mutations in the primary and metastatic tumors. Exome samples in the primary and metastatic tumor have mean depth $\bar{D}$ and purity γ. Our proposed method was applied to sites with $\geq m_{1(\min)}$ mutant reads in the primary tumor. We ran 100 simulations for each parameter set. Mouse models have suggested that metastasis occurs *via* colonization of one circulating tumor cell (CTC) cluster rather than serial arrivals of CTC clusters (or single CTCs) and that the most CTC clusters contain between 2 and 20 tumor cells, with median of 6.[6] We mainly focused on this range of founder sizes in the simulations.

In Figures 2*a*–2*d*, all simulations were performed under the conditions of $N_1 = 100,000$, $\mu = 2.5$. First, the effect of varying mean depth, $\bar{D}$, on the estimation of N_b, was investigated under $K = 50$, $\gamma = 1$ and $m_{1(\min)} = 2$ (Fig. 2*a*). The number of variants generated in the exome samples in the simulations were realistic, ranging from 1 to around 500 (Supporting Information Fig. S2B). In cases of $N_b = 2$, 5, 10 and 20, when $\bar{D} \geq 50$, the medians of estimates were very close to the true values, that is, the estimator is median-unbiased, and the estimation accuracy is good. For example, when the depth was 50, the medians of the estimates (and interquartile ranges; IQRs) were 5.0 (4.0, 6.0), 10.0 (8.0, 13.0) and 20.0 (15.0, 28.25) for the true $N_b = 5$, 10 and 20, respectively. The unbiasedness with $\bar{D} \geq 50$ held for larger N_b (for $N_b = 1$–100, see Supporting Information Fig. S2*a*). The estimation accuracy increased as sequence depth increased. Even when the depth was $\bar{D} = 30$, the precision and accuracy were acceptable, and the medians of estimates (IQRs) were 5.0 (4.0, 6.0), 12.0 (8.0, 18.25) and 21.0 (14.0, 35.0) for the true $N_b = 5$, 10 and 20, respectively. Under $\bar{D} = 30$, and particularly for larger $N_b \geq 30$, N_b was biasedly estimated and a reliable estimation was difficult to obtain (for $N_b = 1$–100, see Supporting

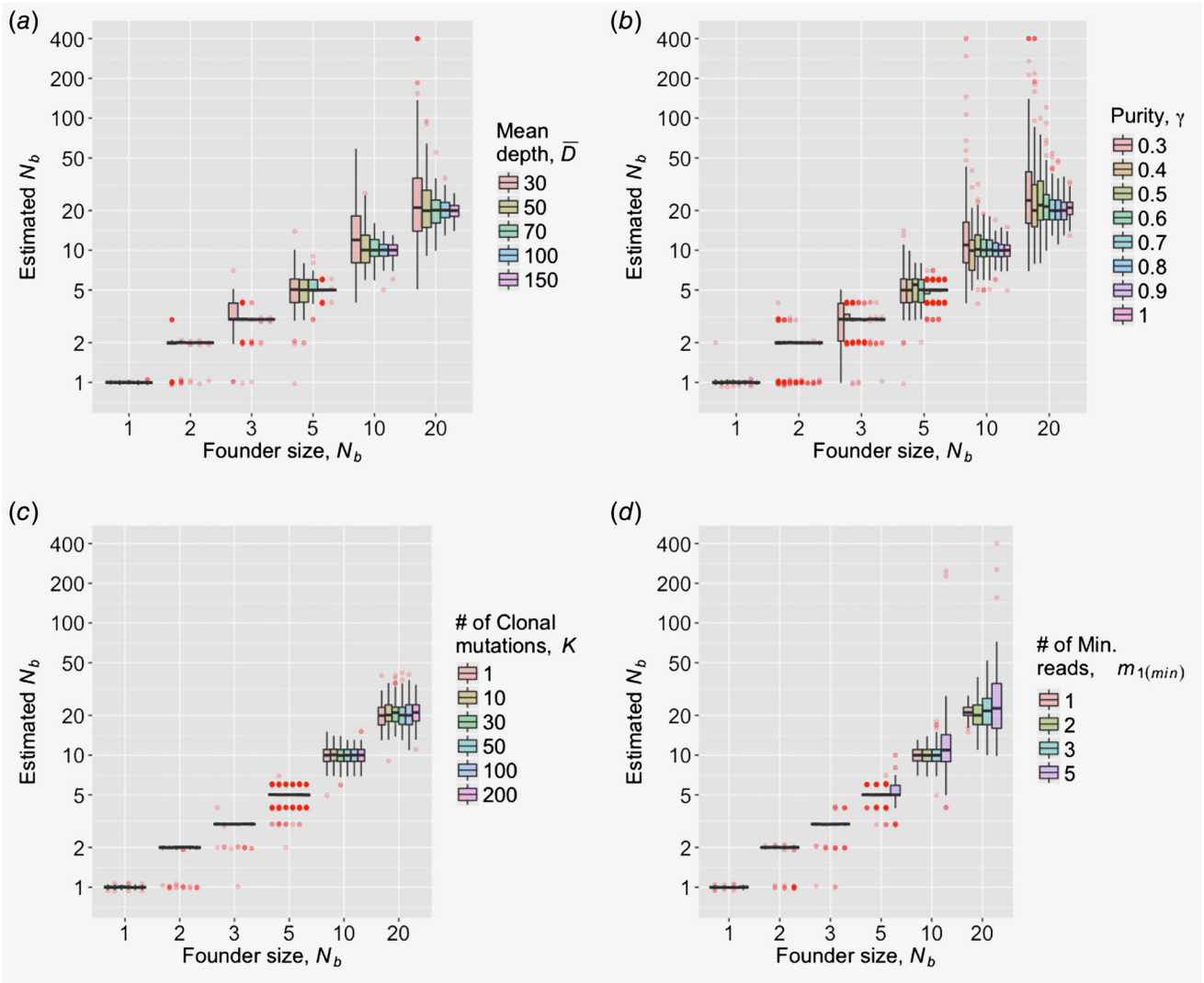

Figure 2. Valid quantification of founder size, N_b, confirmed by simulations. All simulations used "pure birth model" with the primary tumor population size, N_1 = 100,000, and mutation rate per cell division per exome, μ = 2.5. For each parameter set, number of simulations is 100. The lower and upper hinges correspond to the first and third quartiles. Boxplots show medians, 25th and 75th percentiles (hinges). The upper/lower whiskers extend to the largest/smallest value at most 1.5 times of IQR from the upper/lower hinges. (*a*) Varying mean sequencing depth, $\bar{D}$ for $K = 50$, $\gamma = 1$ and $m_{1(\min)} = 2$. (*b*) Varying tumor purity, γ, for $K = 50$, $\bar{D} = 100$ and $m_{1(\min)} = 2$. (*c*) Varying number of clonal mutations, K, for $\bar{D} = 100$, $\gamma = 1$ and $m_{1(\min)} = 2$. (*d*) Varying minimum number of mutation reads, $m_{1(\min)}$, for $K = 50$, $\bar{D} = 100$ and $\gamma = 1$. (Variants with $m_{1(\min)}$ or more mutation reads were used.)

Information Fig. S2*a*). Note that, for all depth settings, the relative estimation errors were better for smaller N_b, as you can see from the smaller log-scaled boxplots of the estimated N_b in Figure 2*a* (see also Supporting Information Fig. S2*a*).

Next, the effects of the tumor purity, γ, on the estimation were investigated under $K = 50$, $\bar{D} = 100$ and $m_{1(\min)} = 2$ (Fig. 2*b*). When $\gamma \geq 50\%$, the estimation was median-unbiased and the accuracy was acceptable. For example, when $\gamma = 50\%$, the medians of the estimates (IQRs) were 5.0 (4.0, 6.0), 10.0 (9.0, 13.0) and 22.0 (17.0, 33.5) for the true $N_b = 5$, 10 and 20, respectively. In conjunction with the result of Figure 2*a*, defining the "effective sequence depth" as the depth multiplied by tumor purity, the proposed method gave unbiased results with acceptable accuracy when the effective sequence depth was 50. In the case of less purity, and large founder size, for example, $\gamma \leq 40\%$ and $N_b \geq 30$, a reliable estimation was difficult obtain (for $N_b = 1$–100, see Supporting Information Fig. S3).

In the algorithm for N_b estimation, the proportion of clonal mutations in the primary tumors is fixed at 10% (Implementation section in Supporting Information Appendix). Practically, however, clonal mutations vary among tumors. Thus, the impact of the number of clonal mutations was investigated under $\bar{D} = 100$, $\gamma = 1$ and $m_{1(\min)} = 2$ (Fig. 2*c*). The number of clonal mutations in the population, K, had no effect on both the unbiasedness and the accuracy of estimation of N_b. The same

is true for larger N_b with various number of variants in WES samples (for $N_b = 1$–100, see Supporting Information Fig. S4).

For the input of the proposed method, we use variants with $m_{1(\min)}$ or more mutation reads in the primary tumor. Then, the effects of various values of $m_{1(\min)}$ on the estimation of N_b were investigated under $K = 50$, $\bar{D} = 100$ and $\gamma = 1$ (Fig. 2*d*). The estimation results for up to $N_b = 100$ were also assessed (Supporting Information Fig. S5). In the case of $m_{1(\min)} = 5$, the estimation accuracy was worse than those for $m_{1(\min)} < 5$. The decreased accuracy was not due to lower numbers of variants used for input (for the case of larger number of variants, see Supporting Information Fig. S6 replacing $\mu = 2.5$ with $\mu = 12.5$). For the case of including singletons in the input ($m_{1(\min)} = 1$), a small upward bias can occur (for more clear bias in the large N_b, see Supporting Information Figure S5). Thus, we recommend the criteria of "at least 2 or 3 mutation read counts," $m_{1(\min)} = 2$ or 3, for the input of the proposed method.

The simulations above were performed mainly under the conditions of the primary tumor size, $N_1 = 100,000$ and mutation rate, $\mu = 2.5$. When values of N_1 ranging from 1,000 to 300,000 were used under $\bar{D} = 100$, $\mu = 2.5$, $K = 50$, $\gamma = 1$, and $m_{1(\min)} = 2$, the behavior of estimates were generally the same as that under $N_1 = 100,000$ (Supporting Information Fig. S7). When values of μ ranging from 0.5 to 10 were used under $\bar{D} = 100$, $N_1 = 100,000$, $K = 50$, $\gamma = 1$ and $m_{1(\min)} = 2$, the behavior of estimates were generally the same as that under $\mu = 2.5$ although the estimation accuracy was a little lower as the mutation rate is small (Supporting Information Fig. S8).

Robustness for cell death and selection in the primary tumor evolution

So far, in the development of primary tumor, it was assumed there was no cell death and no difference in cell division rates. The violation of the assumptions might make estimation of N_b difficult, since VAF distribution, $f(p_1)$, can potentially deviate from the postulated distribution under "neutral" evolution with high cell birth rate of tumor population. Here, we investigated the consequences of this violation, keeping $\bar{D} = 100$ and the all other settings as in Figures 2*a*, that is, $\mu = 2.5$, $K = 50$, $\gamma = 1$ and $m_{1(\min)} = 2$, $N_1 = 100,000$. We ran 100 simulations for each parameter set.

First, to investigate the effect of "cell death," a death rate, d and a birth rate, b, per unit time were introduced. Limiting the case to $d < b$, which means growth of the tumor population, various values, $d = 0.01$, 0.1, 0.2, 0.5, 0.7, 0.9 and 0.99 against unit birth rate, $b = 1$, were assumed (the ratio of d to b define the evolutional system). High death rates, that is, $d > 0.7$, might be more realistic.[11,12] Exceptionally, $N_1 = 10,000$ was used for $d = 0.99$ due to computer capacity limitations. For all death rates, the estimator for the founder size, N_b, is median-unbiased and the estimation accuracy is sufficient, as with the case of no death ($d = 0$; Supporting Information Fig. S9*a*). This is due to the fact that VAF distribution for $d \neq 0$ does not vary greatly from that of the no death case (Supporting Information Figs. S9*b*–S9*d*).

Second, we considered the case that one positively selective subclone in the primary tumor appears in the WES samples.[13] One starting primary tumor cell with $b = 1$, $d = 0.1$ are assumed to evolve and at the time when the population reaches a particular population size, $N_{\text{sub. occ.}}$, one selectively advantageous mutation occurs. The subclone with the advantageous mutation will have a larger birth rate, $b = 2$, 5 or 10. The values of $N_{\text{sub. occ.}}$ are determined so that corresponding frequencies of the selective mutation are low (~2%), middle (~16%) and high (~30%) at the WES sampling point. Although distributions of VAF were shifted to the frequency of the selective mutation (Supporting Information Figs. S10–S12*b*,S12*c*, S12*d*), the estimator of the founder size, N_b, is median-unbiased and the estimation accuracy is sufficient, as in the case of no selective subclone (Supporting Information Figs. S10–S12*a*).

Finally, we considered the case that many mutations with small effects are accumulated in the developmental process of the primary tumor. Neutral mutations and non-neutral mutations occur with the probabilities of 0.3 and 0.7, respectively, which mimics synonymous and nonsynonymous mutation rates in exon regions.

The birth rate of a cell with k non-neutral mutations is given by $(1 + a)^k$, where a is a coefficient for birth rate and set as $a = \pm 0.01$, ± 0.05, ± 0.1, ± 0.15 and ± 0.2. A positive and negative value of a denotes advantageous and deleterious mutations, respectively. The death rate is always set to be one-tenth of population mean of birth rates. Advantageous mutations, particularly when $a \geq 0.1$, shift VAF distribution toward intermediate frequency (Supporting Information Figs. S13*b*–S13*k*). For deleterious or advantageous mutations with $a \leq 0.1$, the estimator for the founder size, N_b, is median-unbiased and the estimation accuracy is sufficient, as in the case of no selection, $a = 0$ (Supporting Information Fig. S13*a*). When strong selection is observed ($a \geq 0.15$), the estimator is biased upwards and the accuracy is low. However, it is unrealistic that 70% of all mutations would have effects as strong as $a \geq 0.15$.

Robustness for selective colonization in the metastasis tumor

The proposed method assumes that the founders of metastatic tumor are randomly sampled from the primary tumor population during the process of metastatic colonization. Practically, however, some mutations might be preferentially selected. Here, we investigated the consequences of selective colonization (Supporting Information Fig. S14). The proportion $p_{s_{\text{met}}}$ of all mutations in the primary tumor populations are advantageous for metastatic colonization and increase metastatic ability of cells multiplicatively by $(1 + s_{\text{met}})$ (see Supporting Information Appendix for more details). The other parameters were set as in Figure 2*a*, that is, $\mu = 2.5$, $K = 50$, $\gamma = 1$ and $m_{1(\min)} = 2$, $\bar{D} = 100$, $N_1 = 100,000$. We ran 100 simulations for each parameter set.

When $p_{s_{met}} = 0.1\%$, which corresponds to around 530 selective mutations in the primary tumor population, the estimator for N_b is almost median-unbiased and the estimation accuracy is sufficient as is the case of neutral ones, even for very strong selection, s_{met} = 100 (Supporting Information Fig. S14*a*). When $p_{s_{met}} = 1\%$, which corresponds to around 5,300 selective mutations in the primary tumor populations, and $s_{met} \leq 5$, the estimator for N_b is robust and comparable to the neutral case. However, when selection is strong ($s_{met} > 5$), the estimator for N_b clearly underestimates the true founder size (Supporting Information Fig. S14*d*). When substantial fraction of mutations is advantageous, $p_{s_{met}} = 10\%$, which corresponds to around 53,000 selective mutations in the primary tumor populations, the estimator for N_b is robust and comparable to the neutral case for $s_{met} \leq 1$. However, when $s_{met} > 1$, the estimator for N_b again clearly underestimates the true founder size (Supporting Information Fig. S14*g*). In summary, except for the cases of numerous advantageous mutations and/or a strong selection coefficient, the proposed method gave robust estimates.

Behavior of the proposed estimator in the stochastic evolution of the metastatic tumor

The proposed model attributes all genetic drift in metastatic tumor evolution to a "single bottleneck," and the estimate of "the founder size" reflects those drifts. Nevertheless, genetic drift due to stochastic evolution of the metastatic tumor should occur during an early stage of development of the metastatic tumor. Here, we investigated the behavior of the proposed estimator in the birth and death processes for metastatic development (Supporting Information Fig. S15). As with the primary tumor, a metastatic founder population consisting of N_b cells are assumed to develop according to the birth and death process, with birth rate, b_{met} and, death rate, d_{met}, per unit time. Cell divisions repeat until the metastatic tumor has grown to the final primary tumor size, N_{met}, at which exome sequencing is done. Limiting the case to $d_{met} < b_{met}$, which means growth of the tumor population, various values, d_{met} = 0.0–0.9 against unit birth rate, b_{met} = 1 and N_{met} = 10,000 were assumed (the ratio of d_{met} to b_{met} defines the evolutional system). The other parameters were set as in Figure 2*a*, that is, μ = 2.5, K = 50, γ = 1 and $m_{1(min)}$ = 2, $\bar{D} = 100$, N_1 = 100,000. We ran 100 simulations for each parameter set.

In the case of d_{met} = 0 (i.e., the birth only process), although the estimation accuracy gets worse compared to the case of no stochastic metastatic evolution (Fig. 2*a*), the estimator for the founder size, N_b, remains nearly median-unbiased (Supporting Information Fig. S15). In addition to cell birth, if slight to moderate cell death is introduced ($d_{met} \leq 0.1$), N_b remains nearly median-unbiasedly estimated. For example, when d_{met} = 0.1, the medians of the estimates (IQRs) were 6.0 (5.0, 7.0), 9.0 (8.0, 11.0) and 16.0 (14.0, 18.0) for the true N_b = 5, 10 and 20, respectively. However, in the case of substantial cell death ($d_{met} > 0.1$), the estimator for N_b clearly underestimates the true founder size. For example, when d_{met} = 0.5, the medians of the estimates (IQRs) were 3.0 (2.0, 5.0), 7.0 (5.0, 8.0) and 11.0 (9.0, 12.25) for the true N_b = 5, 10 and 20, respectively. This is expected as substantial genetic drift due to cell death occur during early stage of development of metastatic tumor. Actually, when the death rate is large (e.g., $d_{met} > 0.1$), substantial founder cells drop out of the metastatic population in the early stage (~100-cell stage; Supporting Information Fig. S16). Note that a small amount of founder cells drop out in sufficiently large populations and the proposed method gives the same results irrespective of the final primary tumor size for $N_{met} \geq 1{,}000$ (data not shown).

Real data analysis for CRC patients

We used high-depth WES data from a study of four CRC patients, which included at least one primary and metachronous metastatic tumor sample per patient.[8] For each patient, the metastatic tumor(s) were sampled 1–3 years after the removal of the primary tumor(s). Information for called mutations of each tumor were derived from the article.[8] As follows, we estimated the founder population size of metastatic (or lymph node) tumors using all pairs of primary and metastatic or lymph node tumors in each patient.

We applied quality-controls to each tumor exome data. Only called mutations with a sequencing depth of 300 or less and no copy number aberrations were considered. The second criterion ensured diploid tumor sequences, which is assumed in the current model. Copy number aberrations were retrieved from the article.[8] For mutation data meeting the criteria, we estimated tumor purities using PurBayes.[14] Purity estimates ranged from 0.147 to 0.821 (Supporting Information Table S1). Next, we conducted quality-controls on the exome data of each primary and metastatic (or lymph node) tumor pair. Mutation sites with at least two mutation reads in the primary tumor, that is, $m_{1(min)}$ = 2 and no mutation read in the normal sample were considered for further analysis. After quality-control, the number of mutations ranged from 70 to 220 and an average sequence depth of 75.61–127.64 and 90.96–144.37 in the primary and metastatic tumor exomes, respectively (Supporting Information Table S2). The observed VAFs were somewhat correlated between the primary and metastatic tumors in each patient (Supporting Information Fig. S17, left).

For exome pairs with sufficient purity estimates (averaged purity estimate ≥0.3) that passed quality control, we estimated the founder population size of the metastatic (or lymph node) tumors using the proposed method (Fig. 3). Founder population sizes were estimated to be from 3 to 17 as MLEs, supporting the "multi-cellular origin" of metastatic tumors. Although founder sizes varied from sample pair to sample pair, similar estimates were obtained for each patient, which supported our implicit assumption of well-mixed tumors to

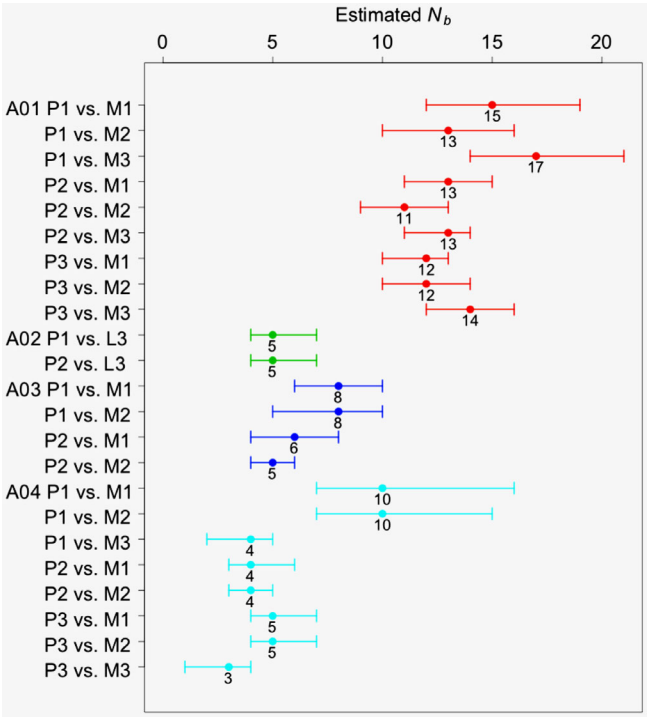

Figure 3. Estimated founder sizes (N_b) for the four colorectal cancer reported by Wei *et al.* (2017). Results using only diploid regions (excluding copy number aberrations) are shown. "P", "M" and "L" means primary, metastatic and lymph node tumors. Circles with bars indicate maximum likelihood estimates of N_b and these 90% confidence intervals, based on 1,000 nonparametric bootstrap samples.

some extent. All pairs of exomes from patient A01 gave consistently large founder sizes, ranging from 11 to 17. Exceptionally, for A04 the pairs "P1 and M1" and "P1 and M2" also had relatively larger founder estimates of 10, while other pairs from the patient ranged from 3 to 5. Although estimates of founder sizes for exome pairs with low purity estimates (averaged purity estimate <0.3) were also obtained (Supporting Information Fig. S18 for estimates using all exome pairs; for VAFs, Supporting Information Figure S17, right), those were not considered reliable because of the low purities.

Sensitivity and validation analysis for CRC data

We conducted three types of sensitivity analysis and one validation analysis to examine the stability/validity of our main results (Fig. 3). First, to see the impact of copy number aberrations, we estimated the founder population size using mutations in the all WES data (summarized in Supporting Information Table S3) without limiting diploid region. The number of all mutations averaged ~20% higher than those limited to diploid regions (see "# of mutations" of Supporting Information Tables S2 and S3). The estimates using all WES data (Supporting Information Fig. S19) were consistent with the main results (Fig. 3) and the impact of a small portion of copy number aberrations was very small.

In the main results, only mutations with two or more than mutation reads in the primary tumor were considered, that is, $m_{1(\min)} = 2$. However, potential mutation sites supported with only a small number of mutation reads may be false positives. For example, the somatic mutation caller, Mutect,[15] with the default settings, does not call sites supported by <5 mutation reads as true mutations at a sequence depth of 100 since the "tumor LOD scores" (log-10 likelihood ratios of a model having mutations to no mutation model in the tumor population) fails to reach its default threshold of 6.3. To examine the impact of this uncertainty, we assumed that 5–20% of potential mutation sites supported by 2, 3 or 4 mutation reads in the primary tumor were calling errors (erroneous sites). In the case that the erroneous sites in the metastatic tumor were forced to have zero mutation read, the estimates did not differ greatly from the main results, although lower estimates were obtained as the error rates increased (Supporting Information Fig. S20). Next, the erroneous sites in the metastatic tumor were assumed to have the same number of mutation reads as the primary tumor. In this case, the (point) estimates were the same as the main result at any error rates (Supporting Information Fig. S21). Even when the erroneous sites in the metastatic tumor had double the number of mutation reads in the primary tumor, the estimates changed very little compared to the main results (Supporting Information Fig. S22).

As a third sensitivity analysis, since the proposed method uses changes in VAF from the primary to metastatic tumor, clonal mutations are not informative on the founder size could bias the estimate. In the simulation study, the number of clonal mutations was shown not to affect the estimate (Fig. 2*c*; Supporting Information Fig. S4) even when we arbitrarily assumed 10% of mutations in the primary tumor were clonal. Defining mutations with observed VAFs > purity estimate × 0.5 × 0.9 as clonal, we confirmed that the estimates without using clonal mutations were similar to the main results, except for the estimates for A02 which were larger than the main results (Supporting Information Fig. S23).

There are mutations that are absent in the primary tumor sample but are present in the metastatic one. For each primary and metastatic tumor pair, we compared such VAFs in the metastatic tumor samples to those from the corresponding simulation based on the "pure birth tumor evolution model" described above (Supporting Information Fig. S24). If the proposed model is relevant, the distributions of observed and simulated VAFs should correspond well with each other. For matched primary and metastatic pairs, the simulations were conducted using the estimated purities (Supporting Information Table S1) and founder size (Fig. 3; Supporting Information Fig. S18), and randomly assigning the depths from matched pairs for each mutation (for the summary of depths, see Supporting Information Table S2), keeping the all other settings as in Figure 2*a*, that is, $\mu = 2.5$, $K = 50$ and $m_{1(\min)} = 2$,

N_1 = 100,000. We ran 100 simulations for each parameter set. For many pairs, the distributions were relatively similar to each other: medians of the observed VAFs are generally close to simulated ones (Spearman's $\rho = 0.534$, $p = 1.6 \times 10{-}4$). However, in many cases, the variations of observed VAFs were different from simulated ones, and the observations tended toward larger VAFs.

Discussion

We developed a method to quantify the founder population size of metastasis using paired WES data from primary and metachronous metastatic tumors. This method, implicitly using the fact that higher (lower) genetic similarity between the primary and metastatic tumors results from a larger (smaller) founder size (Fig. 1*c*), unbiasedly estimates the founder population size with sufficient accuracy in the range of realistic founder sizes and settings, for example, sequencing depth, purity and number of clonal mutations (Fig. 2 and Supporting Information Figs. S2–S8). The method is also robust to the realistic model of primary tumor evolution, including cell death, selection among cells in primary tumor evolution and moderate selective colonization (Supporting Information Figs. S9–S14). Note that in the cases of numerous advantageous mutations and/or a strong selection coefficient, the proposed method underestimates the true founder size for the following reason: Cells with many advantageous mutations, which have high probabilities of colonization, tend to be close relatives of each other and have similar mutational composition; Thus, "effective" number of founder cells should be smaller than "actual" number.

Although relative estimation errors become worse as the founder size increases, this weakness is overcome by deeper sequencing, that is, WES data with ×150 depth give sufficient accuracy even for a founder size of 100 (Supporting Information Fig. S1). As several advanced studies have shown,[7–9,13,16,17] the proposed method also shows the advantage of using VAF information (mutation read counts and depths) rather than using only the presence or absence of mutations, to infer the tumor evolutionary process.

In real data analysis of four CRC patients, we restricted the analysis to pairs of primary and metastatic (or lymph node) tumors with averaged purity ≥0.3 (Fig. 3) since while the estimation is unbiased, variance becomes large when purity ~0.3 (Fig. 2) and further increases when purity <0.3. In fact, the 90% confidence intervals for the tumors with averaged purity <0.3 tended to be large, for example, for the pairs of P1 and L1 (L2) for the patient A2 (Supporting Information Fig. S18).

Our method supports the multi-cellular origin of metastatic tumors, which is consistent with the observations of recent mouse model studies[4–6] and WES studies.[7,8] Our method further quantified the founder population sizes to ranging from 3 to 17 cells for CRC subjects (Fig. 3).[8] The wide-range of founder sizes in metastasis might result in large variations of genetic similarity between primary and metastatic tumors and cause variation in drug response between primary and metastatic tumors. In particular, when the founder population size is small, variants with drastically increased VAFs in the metastatic tumors might lead to difficulty in treatment.

In the context of population genetics, demographic history is a confounding factor for detecting and quantifying natural selection acting on the genome.[18,19] The same should be true for the evolution of a tumor population. A potential advantage of the proposed method is to identify selectively recruited mutations in the metastatic tumors under the inferred demographic model for tumor populations, that is, the estimated founder size.

The limitations of our method are that it assumes the model of single bottleneck occurs just after WES in the primary tumor and is followed by rapid growth for metastatic colonization. However, genetic drift (randomly fluctuation of VAFs) may occur in the period between the first exome sampling and metastatic occurrence or between metastatic occurrence and the second exome sampling. Particularly in the latter, genetic drift may substantially shift the VAFs, or a substantial fraction of founder cells might die off due to genetic drift early after metastatic colonization. Our model attributes such genetic drifts to a "single bottleneck," and the estimate of "the founder size" reflects those drifts. In the simulation for stochastic metastatic tumor evolution showed that for a death rate $\leqq$0.1 against unit the birth rate, the proposed method gives nearly unbiased estimates of founder size (Supporting Information Fig. S15). For death rate >0.1, the true founder size was underestimated due to non-negligible genetic drift (resulting in reducing the metastatic founder cells) during the early stages of development of metastatic tumor (~100-cell stage; Supporting Information Fig. S16). In addition, with respect to mutations that are absent in the primary tumor sample but are present in the metastatic tumor sample, while the central tendencies of the observed VAF distributions in the metastatic tumors were close to the simulated ones, the observed variabilities were larger than simulated ones (Supporting Information Fig. S24). This may show that there is room for improvement on the present simple model, for example, we should incorporate the stochastic metastatic evolution into the present model. Furthermore, there are possibly more complex cell migration patterns, including reseeding or multisource seeding,[17,20] which are also beyond the present study, but worth investigating.

Acknowledgements

This research was partially supported by JST CREST Grant Number JPMJCR1412, Japan, and JSPS KAKENHI Grant Numbers 17H06307 and 17H06299, Japan.

Conflicts of interest

The authors have declared no conflicts of interest.

References

1. Fidler IJ, Talmadge JE. Evidence that intravenously derived murine pulmonary melanoma metastases can originate from the expansion of a single tumor cell. *Cancer Res* 1986;46:5167–71.
2. Maddipati R, Stanger BZ. Pancreatic cancer Metastases Harbor evidence of Polyclonality. *Cancer Discov* 2015;5:1086–97.
3. Talmadge JE, Fidler I. Evidence for the clonal origin of spontaneous metastases. *Science* 1982;217:361–3.
4. Aceto N, Bardia A, Miyamoto DT, et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* 2014;158:1110–22.
5. Cheung KJ, Ewald AJ. A collective route to metastasis: seeding by tumor cell clusters. *Science* 2016; 352:167–9.
6. Cheung KJ, Padmanaban V, Silvestri V, et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc Natl Acad Sci USA* 2016; 113:E854–63.
7. Gundem G, Van Loo P, Kremeyer B, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* 2015;520:353–7.
8. Wei Q, Ye Z, Zhong X, et al. Multiregion whole-exome sequencing of matched primary and metastatic tumors revealed genomic heterogeneity and suggested polyclonal seeding in colorectal cancer metastasis. *Ann Oncol* 2017;28:2135–41.
9. Williams MJ, Werner B, Barnes CP, et al. Identification of neutral tumor evolution across cancer types. *Nat Genet* 2016;48:238–44.
10. Ohtsuki H, Innan H. Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theor Popul Biol* 2017;117:43–50.
11. Bozic I, Antal T, Ohtsuki H, et al. Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci USA* 2010;107:18545–50.
12. Diaz LA Jr, Williams RT, Wu J, et al. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 2012;486:537–40.
13. Williams MJ, Werner B, Heide T, et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat Genet* 2018;50:895–903.
14. Larson NB, Fridley BL. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* 2013; 29:1888–9.
15. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;31:213–9.
16. Sun R, Hu Z, Sottoriva A, et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet* 2017;49: 1015–24.
17. El-Kebir M, Satas G, Raphael BJ. Inferring parsimonious migration histories for metastatic cancers. *Nat Genet* 2018;50:718–26.
18. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. *Nat Rev Genet* 2003;4:99–111.
19. Nielsen R, Hellmann I, Hubisz M, et al. Recent and ongoing selection in the human genome. *Nat Rev Genet* 2007;8:857–68.
20. Sanborn JZ, Chung J, Purdom E, et al. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc Natl Acad Sci USA* 2015;112:10995–1000.