Quantification of multicellular colonization in tumor metastasis using exome sequencing data

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

[1] Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), 1-5-45 Yushima, Bunkyo-ku, Tokyo, 113-8510, Japan

[2] Department of Hepatobiliary and Pancreatic Surgery, Tokyo Medical and Dental University (TMDU), 1-5-45 Yushima, Bunkyo-ku, Tokyo, 113-8510, Japan

[3] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehirocho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan

[4] CREST, JST, Tokyo, 113-8510, Japan

[*] Address for correspondence:

Tatsuhiko Tsunoda, Ph.D. (Medicine) & Ph.D. (Engineering)
Department of Medical Science Mathematics
Medical Research Institute
Tokyo Medical and Dental University (TMDU)
1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan
Email: tsunoda.mesm@mri.tmd.ac.jp

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

Abstract

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 multi-cellular 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, 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 multi-cellular origin of metastasis and the founder size was quantified, ranging from 3 to 15 cells. Such a wide-ranging 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 might be involved in (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 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 (Figure 1). The drastic genetic changes in the metastatic tumor from the primary one, brought by the limited cell migration, i.e., 'bottleneck effect', might result in a difference in drug response between both tumors in the same patients.

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 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 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. The method uses the outputs from commonly used mutation callers, i.e., 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 for four colorectal cancer (CRC) patients.

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 the normal tissue (Figure 1A). 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 sites in the primary and metastatic tumors, respectively (Table 1 and Figure 1B; See supplementary Appendix and supplementary Figure S1 for more details of the input file). When the founder population size is large, the variant allele frequencies (VAFs) at given sites in the metastatic show high similarity to those in the primary tumor (Figure 1C). Conversely, when the size of founder cells is small, the VAFs in the primary and metastatic tumors are not so correlated (Figure 1C). In this case, due to the severe 'bottleneck effect', many variants can become extinct in the metastatic tumor or have a significantly higher VAF 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). There is no recurrence mutation at the same sites then the VAF is at most 0.5, $p_1 \leq 0.5$. The VAF follows some distribution, $p_1 \sim f(p_1)$, as is properly assumed by taking into account the nature of tumor cell evolution (see Implementation section in supplementary Appendix). In 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,

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

In bulk-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:

$$m_2 \sim \sum_{M_b=0}^{N_b} \{Bin(M_b|N_b, 2p_1)Bin(m_2|D_2, p_2)\},$$

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 randomly selected from the primary tumor and colonize a metastatic site. The M_b mutant cells in the metastatic site follows a binomial distribution with parameters N_b and $2p_1$ (mutant cell fraction). The sampled mutation read count, m_2, follows a binomial distribution with parameter D_2 and p_2, where p_2 is given by $p_2 = \frac{M_b}{2N_b}$.

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

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

For quality control, we use only the sites with or more than $m_{1(min)}$ (>0) 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 $m_1(\geq m_{1(min)})$ and $m_2(\geq 0)$ mutation reads in the primary and metastatic tumors, respectively, is expressed as

$$\frac{\int_{p_1=0}^{1} f(p_1)Bin(m_1|D_1,\ p_1) \sum_{M_b=0}^{N_b} \left\{Bin(M_b|N_b,\ 2p_1)Bin(m_2|D_2,\ \frac{M_b}{2N_b})\right\} dp_1}{\sum_{m'_1=m_{1(min)}}^{D_{1i}} \int_{p_1=0}^{1} f(p_1)Bin(m'_1|D_1,\ p_1)\, dp_1},$$

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

$$Likelihood(N_b)$$

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

By maximizing the likelihood (1), we obtain the maximum likelihood estimate of N_b (for implementation details, see supplementary Appendix). In reality, the independence assumption among variants does not hold since the 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} with $\gamma_1 p_{1i}$ and $\frac{M_b}{2N_b}$ with $\gamma_2 \frac{M_b}{2N_b}$, respectively.

Results

Validation of our proposed method by simulations

We assessed our proposed method using simulated data (see Methods and supplementary Appendix for details; see also Table 1 for notations). Briefly, a single tumor cell with K mutations generates a daughter cell with an average μ new mutations and cell divisions repeat until the population has grown to the final primary tumor size, N_1. N_b cells from the N_1 cells make up a metastatic tumor. Exome samples in the primary and metastatic tumor have depth $\overline{D}$ and purity γ. Our proposed method was applied to the sites with $\geq m_{1(min)}$ mutant reads in the primary tumor. We ran 100 simulation for each parameter sets.

In Figure 2A-D, all simulations were performed under the conditions of $N_1 = 100{,}000$, μ=5. Firstly, the effect of varying mean depth, $\overline{D}$, on the estimation of N_b, was investigated under K=50, γ=1, and $m_{1(min)}$=2 (Figure 2A). The number of variants generated in the exome samples in the simulations were realistic, ranging from around one to five hundred (Supplementary Figure S2A). In cases of N_b=2, 5, 10 and 20, when $\overline{D}$ ≥50, the medians of estimates were very close to the true values, i.e., 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 $\overline{D}$ ≥50 held for larger N_b (for N_b=1-100, see supplementary

Figure S2B). The estimation accuracy got better as sequence depth increased. Even when the depth was $\overline{D} = 30$, the precision and accuracy were acceptable, the medians of estimates (IQRs) were 5.0 (5.0, 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 $\overline{D} = 30$, 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 supplementary Figure S2B). 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 1A (see also supplementary Figure S2A).

Next, the effects of the tumor purity, γ, on the estimation were investigated under K=50, $\overline{D}$=100, and $m_{1(min)}$=2 (Figure 2B). When $\gamma \geq 50\%$, the estimation was median-unbiased and the accuracy was acceptable. The medians of the estimates (IQRs) were 5.0 (5.0, 6.0), 10.0 (8.0, 12.0), and 20.0 (15.0, 28.0) for the true N_b=5, 10 and 20, respectively. In conjunction with the result of Figure 1A, 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 e.g., $\gamma \leq 40\%$ and $N_b \geq 30$, a reliable estimation was difficult obtain (for N_b=1-100, see supplementary Figure S3).

In the algorithm for N_b estimation, the proportion of clonal mutations in the primary tumors are fixed at 10%. However, the true clonal mutations vary among tumors. Thus, the impacts of the number of clonal mutations were investigated under $\overline{D} = 100$, γ=1 and $m_{1(min)}$=2 (Figure 2C). 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 supplementary Figure S4).

As 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, $\overline{D}$=100, and γ=1 (Figure 2D). The estimation results for up to N_b=100 were also assessed (supplementary Figure S5). In the case of $m_{1(min)} = 5$, the estimation accuracy was worse than those for $m_{1(min)} < 5$. The worse accuracy was not due to lower numbers of variants used for input (for the case of larger number of variants, see supplementary Figure S6 replacing μ=5 with μ=25). 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 supplementary 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.

Simulations were performed mainly under the conditions of the primary tumor size, $N_1 = 100{,}000$ and mutation rate, μ=5. When values of N_1 ranging from 1,000 to 300,000 were used under $\overline{D}$=100, μ=5, K=50, γ=1, and $m_{1(min)}$=2, the behavior of estimates were generally the same as that under $N_1 = 100{,}000$ (supplementary Figure S7). When values of μ ranging from 1 to 20 were used under $\overline{D}$=100, $N_1 = 100{,}000$, K=50, γ=1, and $m_{1(min)}$=2, the behavior

of estimates were generally the same as that under $\mu = 5$ although the estimation accuracy was a little lower as the mutation rate is small (supplementary Figure S8).

Real data analysis

We used the high-depth WES data from a study for four colorectal cancer (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). Called mutation summary information for each tumor were derived from the article [8]. Our method was applied to the four paired primary and metastatic tumors' data in the four patients, the data from Pri-1 and Met-1 for each patient (the primary and metastatic tumors were arbitrarily labeled Pri-x and Met-x, respectively, per patient). We conducted quality-controls and used the called mutations satisfying the following criteria: within 1,000 sequencing depths in the primary and metastatic tumors, having at least two mutation reads in the primary tumor, i.e., $m_{1(min)}$=2, having no mutation read in the normal sample in the primary and metastatic tumors, and without no copy number aberrations. The last criterion ensured diploid tumor populations, which is assumed in the current model, and copy number aberrations were retrieved from the article[8].

The proposed method was applied to the mutations that passed the quality-controls using the estimated tumor purities by PurBayes[9]. Mutations considered to be possible errors, many of which had higher VAFs (supplementary Figure S9) were removed. We then performed the definitive analyses to estimate the founder population size of metastatic tumors. The average number of variants used ranged from 79-195 for the four samples, with average sequence depths of 83.4-146.7 and 83.4-146.7 in the primary and metastatic tumor exomes, respectively (Table 2). The tumor purities were 0.23–0.72 and 0.29-0.85 in the primary and metastatic tumor samples, respectively (Table 2). The observed VAFs looked to be somewhat correlated in the primary and metastatic tumor in each patient (supplementary Figure S10). Estimated founder population sizes (80% confidence intervals) were estimated to be 15 (13.0, 17.0), 3 (2.0, 4.0), 8 (6.0, 9.0), and 14 (9.0, 21.1) for subject A01, A02, A03, A04, respectively (Figure 3). Consistent results were obtained when the same analyses were carried out using all variations without limiting diploid regions (Table 2 for the mutation summary, supplementary Figure S9 for removed outliers, supplementary Figure S10 for VAFs, and Figure 3 for the estimated founder sizes).

Discussion

We developed a method to quantify the founder population size in metastasis using a paired WES data from primary and metachronous metastatic tumors. This method, implicitly uses the fact that higher (lower) genetic similarity between the primary and metastatic tumors is come from larger (smaller) founder size (Figure 1C), enables us to unbiasedly estimate the founder population size

with sufficient accuracy in the range of realistic founder size and settings, e.g., sequencing depth, purity and number of variations (Figure 2 and supplementary Figure 1-7). Although relative estimation errors become worse as the founder size became larger, this weakness is overcome by deeper sequencing, i.e., WES data with $\times 150$ depth give sufficient accuracy even for the founder size 100 (supplementary Figure S1). 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, as has been applied so far[10-12].

In real data analysis of four colorectal cancer patients, our method supported the multi-cellular origin of metastatic tumors, which is consistent with the observation of recent mouse model studies[4-6] and the suggestion from WES studies[7, 8]. Our method further quantified the founder population sizes ranging 3 to 15 cells for CRC subjects[8]. The wide-ranging founder size in metastasis might result in large variations of genetic similarity between primary and metastatic tumors, which might cause variation in drug responses 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 or quantifying natural selection acting on the genome[13, 14]. The same should be true for the evolution of a tumor population. A potential advantage of the proposed method is to lead to identify selectively recruited mutations in the metastatic tumors under the inferred demographic model for tumor populations, i.e., the estimated founder size.

The limitations of our method are that it does not consider the time between the first exome sampling and metastatic occurrence and the time between metastatic occurrence and the second exome sampling. Particularly, in latter time periods, genetic drift in a small population of new metastatic tumor might not negligible. Our model does not distinguish between such genetic drifts and the bottleneck effect of metastatic colonization, and the estimated of 'the founder size' reflects both these effects. A method that distinguish both effects will be future work. In addition, there are possibly more complex cell migration patterns than our model, including reseeding or multisource seeding[15, 16], which are beyond the present study but worth investigating.

Funding

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

Disclosure

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 research* 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, Donaldson MC, Wittner BS, Spencer JA, Yu M, Pely A, Engstrom A, Zhu H. 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, Schipper K, Cohen JD, Fairchild AN, Gorin MA, Verdone JE, Pienta KJ, Bader JS, Ewald AJ. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc Natl Acad Sci U S A* 2016;**113**: E854-63.

7. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, Brewer DS, Kallio HML, Hognas G, Annala M, Kivinummi K, Goody V, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* 2015;**520**: 353-7.

8. Wei Q, Ye Z, Zhong X, Li L, Wang C, Myers RE, Palazzo JP, Fortuna D, Yan A, Waldman SA, Chen X, Posey JA, 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. Larson NB, Fridley BL. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* 2013;**29**: 1888-9.

10. Sun R, Hu Z, Sottoriva A, Graham TA, Harpak A, Ma Z, Fischer JM, Shibata D, Curtis C. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet* 2017;**49**: 1015-24.

11. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nat Genet* 2016;**48**: 238-44.

12. Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, Graham TA. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat Genet* 2018;**50**: 895-903.

13. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. *Nat Rev Genet* 2003;**4**: 99-111.

14. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. *Nat Rev Genet* 2007;**8**: 857-68.

15. El-Kebir M, Satas G, Raphael BJ. Inferring parsimonious migration histories for metastatic

cancers. *Nat Genet* 2018;**50**: 718-26.

16. Sanborn JZ, Chung J, Purdom E, Wang NJ, Kakavand H, Wilmott JS, Butler T, Thompson JF, Mann GJ, Haydu LE, Saw RP, Busam KJ, et al. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc Natl Acad Sci U S A* 2015;**112**: 10995-1000.

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 with or more than $m_{1(min)}$ mutation reads in the primary tumor.
m_1, m_2	Mutation read counts for the primary (m_1) and metastatic tumors (m_2).
$m_{1(min)}$	Minimum mutation read count in WES data from the primary tumor. For estimating N_b, we use only the sites with or more than $m_{1(min)}$ mutant reads.
D_1, D_2	Sequence depths for the primary (D_1) and metastatic tumors (D_2).
p_1, p_2	Population VAFs in the primary (p_1) and metastatic tumor (p_2).
$f(p_1)$	Probability density 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 in 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.
$\overline{D}$	Mean sequence depth in the primary and metastatic tumor.
γ	Tumor purity in the WES samples from the primary and metastatic tumors (γ_1=γ_2).

Table 2. Summary of WES data for CRC subjects from Wei et al. (2017).

Subject	# of SNV (R)	Tumor	Mean Depth	Purity (95%CI)
A01 - Diploid	195	Pri-1 (colon)	137.6	0.46 (0.44, 0.48)
		Met-1 (liver)	153.1	0.48 (0.47, 0.50)
- All	233	Pri-1 (colon)	138.4	0.46 (0.44, 0.48)
		Met-1 (liver)	154.4	0.48 (0.46, 0.49)
A02 - Diploid	165	Pri-1 (colon)	130.3	0.23 (0.22, 0.24)
		Met-1 (lung)	129.7	0.29 (0.28, 0.30)
- All	209	Pri-1 (colon)	139.5	0.45 (0.43, 0.48)
		Met-1 (lung)	141.1	0.29 (0.28, 0.30)
A03 - Diploid	72	Pri-1 (colon)	83.4	0.25 (0.23, 0.26)
		Met-1 (liver)	102.6	0.59 (0.58, 0.61)
- All	110	Pri-1 (colon)	87.4	0.46 (0.43, 0.49)
		Met-1 (liver)	105.2	0.80 (0.76, 0.85)
A04 - Diploid	193	Pri-1 (colon)	146.7	0.72 (0.68, 0.76)
		Met-1 (lung)	148.2	0.85 (0.81, 0.89)
- All	224	Pri-1 (colon)	149.6	0.70 (0.65, 0.74)
		Met-1 (lung)	150.0	0.84 (0.80, 0.88)

Purities and 95% credible intervals (CIs) were estimated by PurBayes [9]. Diploid: Results using only diploid regions (excluding copy number aberrations). All: Results using all exome regions (including copy number aberrations).

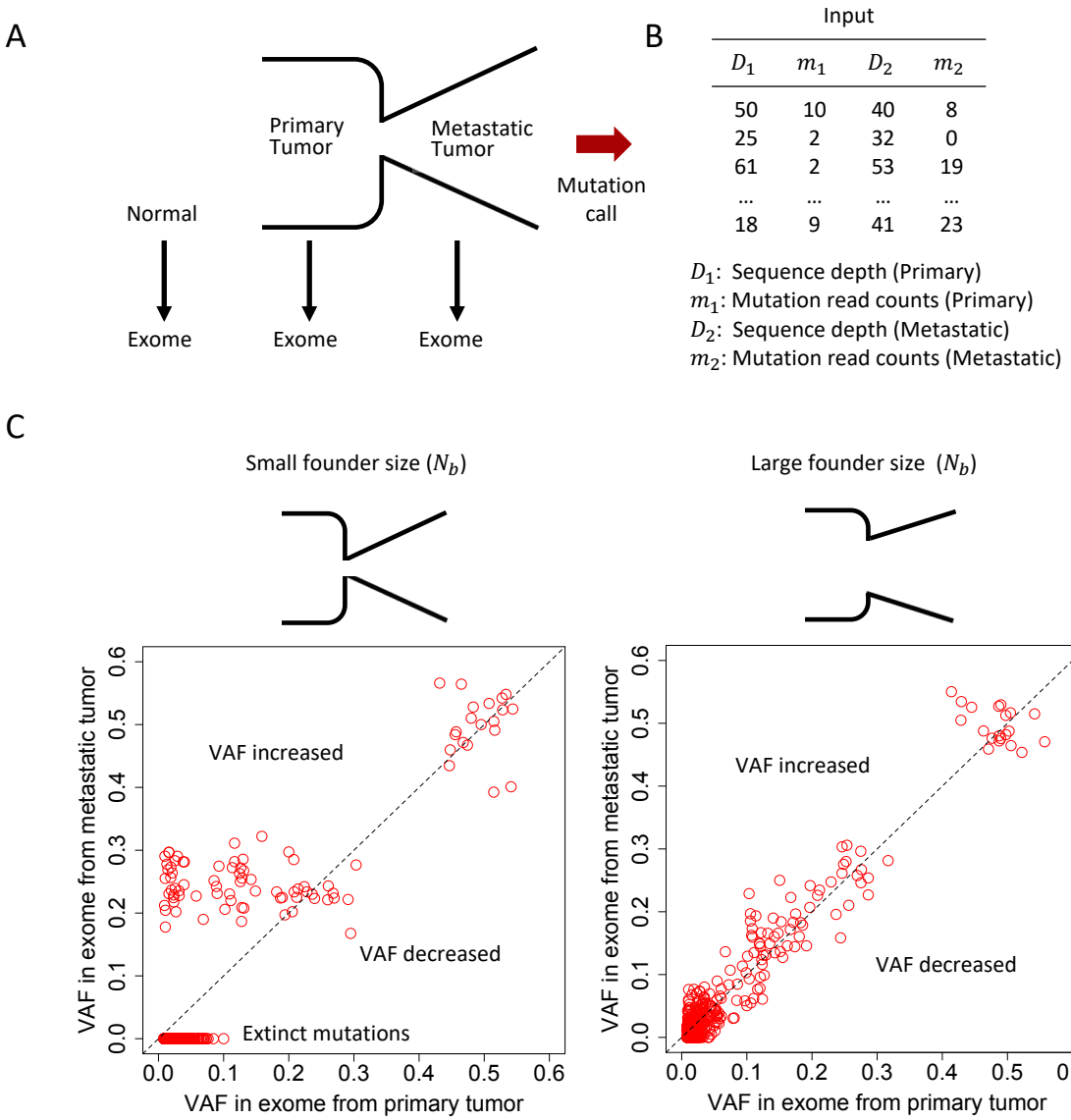

Figure 1. A schematic view of the proposed methodology. (A) Exome data from paired primary and metastatic tumors, and normal tissue. (B) Input of the method. (C) Illustration of basic premise for the estimation of founder sizes by computer simulations. Low correlation of observed VAFs between the primary 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).

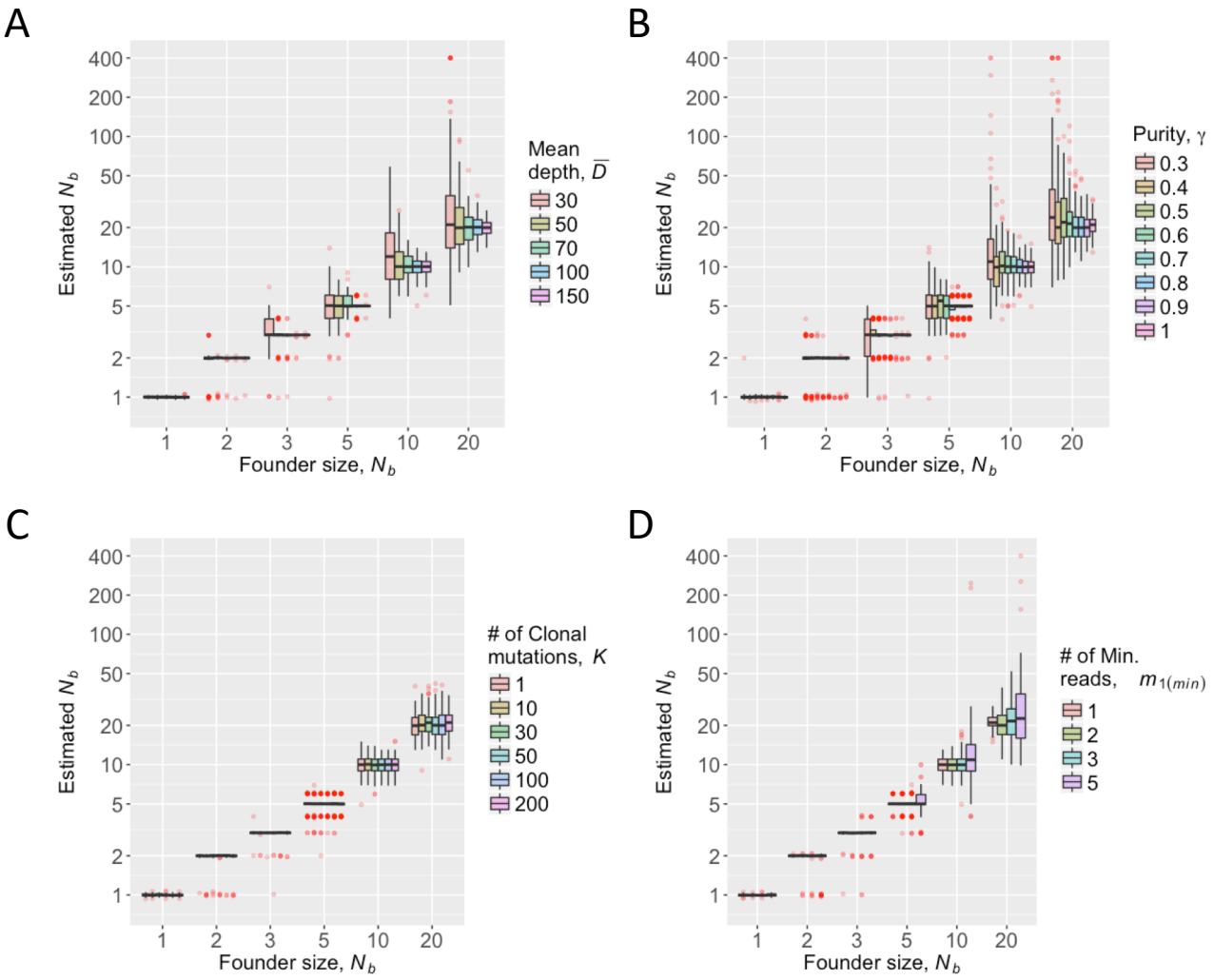

Figure 2. Valid quantification of founder size, N_b, confirmed by simulations. All simulations used the same primary tumor population size, $N_1 = 100{,}000$, and mutation rate per cell division per exome, μ = 5. (A) Varying mean sequencing depth, $\overline{D}$ for K=50, γ=1, and $m_{1(min)}$=2. (B) Varying tumor purity, γ, for K=50, $\overline{D}$=100, and $m_{1(min)}$=2. (C) Varying number of clonal mutations, K, for $\overline{D} = 100$, γ=1 and $m_{1(min)}$=2. (D) Varying minimum number of mutation reads, $m_{1(min)}$, for K=50, $\overline{D}$=100, and γ=1. (Variants with $m_{1(min)}$ or more mutation reads were used.)

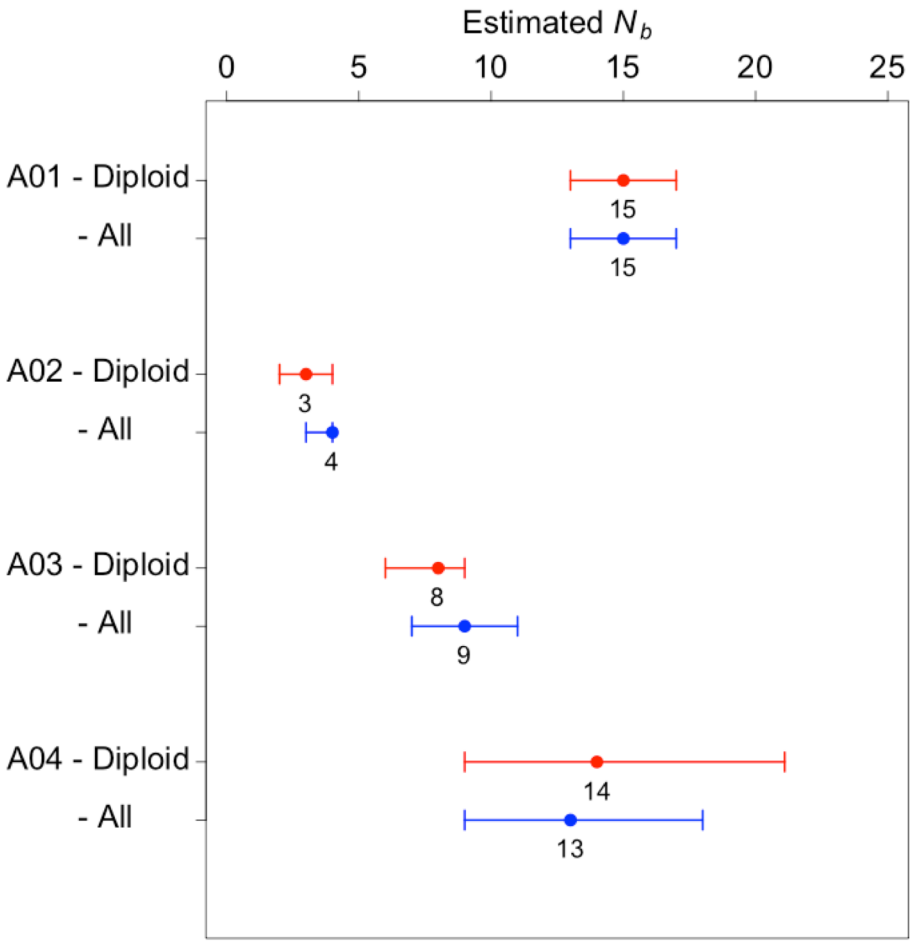

Figure 3. Estimated founder sizes (N_b) for the four colorectal cancer reported by Wei et al. (2017). Results using only diploid regions (excluding copy number aberrations) are shown in red. Results using all exome regions (including copy number aberrations) are shown in blue. Circles with bars indicate maximum likelihood estimates of N_b and these 80% confidence intervals, based on 100 non-parametric bootstrap samples.