https://doi.org/10.1093/nargab/lqab093

Clone decomposition based on mutation signatures provides novel insights into mutational processes

Taro Matsutani[1,2,*] and Michiaki Hamada[1,2,3,*]

[1]Graduate School of Advanced Science and Engineering, Waseda University, 55N-06-10, 3-4-1, Okubo Shinjuku-ku, Tokyo 169–8555, Japan, [2]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 169–8555, Japan and [3]Graduate School of Medicine, Nippon Medical School, Sendagi, Bunkyo, Tokyo 113-8602, Japan

Received May 08, 2021; Revised September 17, 2021; Editorial Decision September 26, 2021; Accepted October 19, 2021

ABSTRACT

Intra-tumor heterogeneity is a phenomenon in which mutation profiles differ from cell to cell within the same tumor and is observed in almost all tumors. Understanding intra-tumor heterogeneity is essential from the clinical perspective. Numerous methods have been developed to predict this phenomenon based on variant allele frequency. Among the methods, CloneSig models the variant allele frequency and mutation signatures simultaneously and provides an accurate clone decomposition. However, this method has limitations in terms of clone number selection and modeling. We propose SigTracer, a novel hierarchical Bayesian approach for analyzing intra-tumor heterogeneity based on mutation signatures to tackle these issues. We show that SigTracer predicts more reasonable clone decompositions than the existing methods against artificial data that mimic cancer genomes. We applied SigTracer to whole-genome sequences of blood cancer samples. The results were consistent with past findings that single base substitutions caused by a specific signature (previously reported as SBS9) related to the activation-induced cytidine deaminase intensively lie within immunoglobulin-coding regions for chronic lymphocytic leukemia samples. Furthermore, we showed that this signature mutates regions responsible for cell–cell adhesion. Accurate assignments of mutations to signatures by SigTracer can provide novel insights into signature origins and mutational processes.

INTRODUCTION

Intra-tumor heterogeneity (ITH) is a phenomenon in which the mutation profiles differ from cell to cell within the same tumor and is observed in almost all tumors. In clinical practice (especially for treatment strategies), understanding heterogeneity is an important task because cell populations with heterogeneous genetic profiles make it challenging to determine which drugs are effective for a particular tumor (1). In addition, heterogeneity represents how tumors have evolved. Hence, the accurate estimation of heterogeneity is essential to elucidate cancer dynamics. Multi-region sampling and single-cell DNA sequencing are effective in estimating the heterogeneity of a single tumor because they directly provide region-by-region and cell-by-cell mutational profiles. However, because of the high cost, low sequencing depth and small amount of data acquired, *bulk* sequencing data are used in numerous cases for comprehensive analysis instead of single-cell sequencing.

ITH is mainly estimated via bulk sequencing by decomposing all mutations into temporally similar populations (called clones) based on the cancer cell fraction representing a hypothetical time axis. Numerous methods to predict ITH have been developed (2–7), and most depend on clustering mutations (i.e. decomposing into clones) by probabilistically modeling the variant allele frequency (VAF) for each sequenced mutation and calculating the cancer cell fraction (CCF) for each mutation by correcting the VAF using copy number aberrations (CNA) according to structural variants. However, the VAF of a single mutation is a noisy observation due to some technical limitations such as low sequencing depth, and it is difficult to accurately reconstruct the tumor evolution using only VAF.

When modeling mutations probabilistically, it is natural to focus on the cause of mutations, in other words, the mutational processes. In general, each mutational process leaves a specific fingerprint. This mutation spectrum can be formulated as a probabilistic distribution called the mutation signature (8,9). For instance, the deamination of 5'-methylcytosine results in the characteristic single base substitution, N[C>T]G; hence, its mutation signature tends to have a higher proportion of these substitutions than others. The probabilistic distribution representing such a fea-

[*]To whom correspondence should be addressed to Michiaki Hamada. Tel: +81 3 5286 3130; Fax: +81 3 5286 3130; Email: mhamada@waseda.jp
Correspondence may also be addressed to Taro Matsutani. Tel: +81 3 5286 3130; Fax: +81 3 5286 3130; Email: taro.matsutani@hamadalab.com

© The Author(s) 2021. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

ture is registered in the COSMIC database as a signature (namely, SBS1), and >50 other signatures for single base substitutions have been reported. Several studies have suggested that each clone in the tumor has different active signatures (10–12), and incorporating signatures to the model for estimating ITH was found to be useful. Abécassis *et al.* proposed CloneSig (13) that models mutation signatures and VAF simultaneously. CloneSig was found to outperform conventional methods of clone estimation with simulated mutation profiles that mimic whole-exome sequencing samples. While jointly modeling VAF and signatures, CloneSig considers that all mutations that accumulate in one tumor occur due to multiple signatures and each clone differs in the signature composition. For example, the clones with a strong SBS1 signature are expected to carry more N[C>T]G point mutations than other clones, as mentioned above. Therefore, mutation-clone matching can be achieved by following multiple clues including VAF and the substitution type and observing the surrounding bases, which improves the accuracy of clone estimation.

CloneSig has enabled many prospects for clone decomposition, but there are still some limitations. One of the drawbacks of CloneSig is the inaccurate model selection in terms of the number of clones using the Bayesian information criterion (BIC). BIC does not support rigorous validity of singular models (14) or mixture models, which have hidden variables. Another limitation of CloneSig is the instability of the estimation due to parameters with no prior distributions. In this study, we developed SigTracer, a hierarchical Bayesian extension of CloneSig that provides a method of valid clone number selection and more robust clone decomposition, which selects the clone number using the evidence lower bound (ELBO), the lower bound of the Kullback–Leibler (KL) divergence between the true distribution and the approximate posterior distribution. Besides, SigTracer prepares Dirichlet distributions as conjugate priors for the signature activity of each clone. This extension can be regarded as a generalization of the CloneSig model, and other studies on similar tasks such as signature extraction have already highlighted the effectiveness of Dirichlet priors (15–17). Here, we aimed to show if SigTracer provides more reasonable clone estimations than CloneSig for artificial tumors.

Clone decomposition based on mutation signatures also has significant potential in terms of signature analysis. Although methods to estimate which signatures are active in a given tumor have been proposed (18,19), they usually do not consider VAF. In other words, they infer which signature leads to a certain mutation from only its trinucleotide type and the mutational distribution of the signature. This can lead to incorrect assignment of mutations to signatures. In fact, in the original paper reporting CloneSig, an example of a sarcoma patient was provided who had clones with differently dominant signatures within a single tumor, and other studies have also suggested the transition of signature activities in coordination with clones (10–12). All these results indicate the effectiveness of considering VAF for the signature analysis. We applied SigTracer to single-cell sequences of the ovarian cancer sample and whole-genome sequences of blood cancer samples, and then reported the findings through accurate signature assignment.

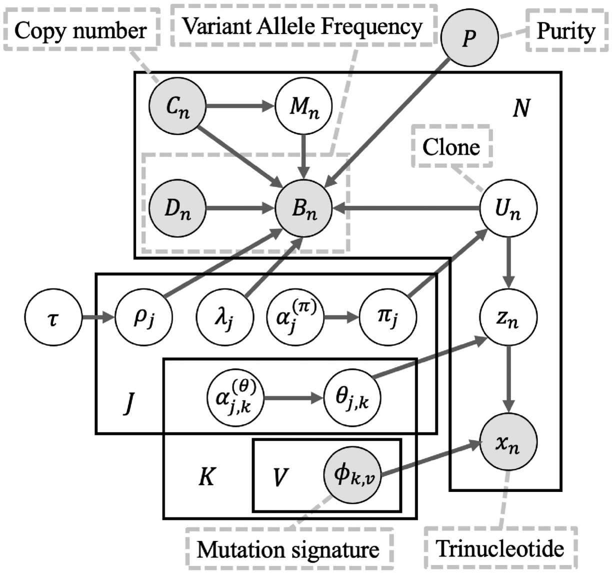

Figure 1. Graphical model of SigTracer for one tumor sample. This representation follows the plate notation, in which the variables shaded in black represent constants that can be observed in advance. Notations for all variables are explained in the main text and Supplementary Table S1.

MATERIALS AND METHODS

Overview of the SigTracer model and the generative process

Figure 1 shows the graphical model of SigTracer for a single tumor, and we have summarized all notations in Supplementary Table S1. The tumor contained a total of N point mutations, and for every mutation, we simultaneously modeled the mutation type x_n ($1 \le n \le N$) and the number of reads overlapping x_n. B_n and D_n indicate the number of reads with mutated and total alleles, respectively. We considered six types of single base substitutions and 16 different neighboring bases around the mutated base using the known SBS signature set; thus, x_n was a categorical variable, and it took $V = 96$ different values. Relatedly, assuming that the subset consisting of K active signatures in the tumor was known and each mutational distribution of the k-th signature was denoted by $\boldsymbol{\phi}_k \in \mathbb{R}^V$ for $1 \le k \le K$, we easily estimated the subset by applying various fitting methods (18,19) to the mutation set in advance. For the genomic locus in which each mutation existed, we assumed that the copy number in a cancer cell $C_n^{(\text{tumor})}$, the copy number of a major allele in a cancer cell $C_n^{(\text{major})}$, the copy number in a normal cell $C_n^{(\text{normal})}$ (these are collectively denoted as C_n in Figure 1), and the sample purity P were also known, and we must set these values in some way to predict CNA considering structural variants (20,21).

For each mutation, the SigTracer model had three latent variables—z_n, U_n, and M_n—indicating the signature via which the mutation occurred (a categorical variable with K types), the clone carrying that mutation (a categorical variable with J types where J is the number of clones), and the copy number of the mutation ($M_n \in \mathbb{N}$ satisfying $M_n \le C_n^{(\text{major})}$), respectively. The clone U_n and signature z_n were generated using categorical distributions with parameters $\boldsymbol{\pi} \in \mathbb{R}^J$ and $\boldsymbol{\theta}_{j=U_n} \in \mathbb{R}^K$, which followed prior Dirichlet distributions with $\boldsymbol{\alpha}^{(\pi)}$ and $\boldsymbol{\alpha}^{(\theta)}_{j=U_n}$. Note that each clone *j* had a different signature activity $\boldsymbol{\theta}_j$. M_n was also modeled to follow a categorical distribution so that any possible natural number less than or equal to $C_n^{(\text{major})}$ was equally sampled. Then, the number of reads, B_n and D_n, was probabilistically modeled as follows:

$$B_n \sim \text{BetaBinomial}(D_n, \mu_n, \nu_n) \tag{1}$$

where

$$\mu_n = \rho_{U_n} \times \lambda_{U_n} \times \eta_{n,M_n}, \quad \nu_n = \rho_{U_n}(1 - \lambda_{U_n} \times \eta_{n,M_n}),$$
$$\eta_{n,M_n} = \frac{P \times M_n}{P \times C_n^{(\text{tumor})} + (1-P) \times C_n^{(\text{normal})}}.$$

Here, λ_j and ρ_j are the CCF and overdispersion parameters for the *j*-th clone, respectively. BetaBinomial(N, μ, ν) is the probabilistic distribution of the number of successes in N trials, where success probability is sampled from a prior beta distribution with shape parameters μ and ν. $\eta_{n,m}$ is the normalization term for the copy number of the mutation n in a sampled cell when $M_n = m$ and the expected value of Beta(μ, ν) becomes $\lambda_{U_n} \times \eta_{n,M_n}$. CCF, which is the VAF corrected for copy number, is defined as the proportion of sequenced cancer cells that contain mutations. λ_j, which is the CCF for the *j*-th clone, shows when the clone was established (the larger the λ, the older the clone) under certain assumptions, including the infinite-site model.

In summary, the entire generative process is given as follows:

```
$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\pi)})$
for each clone $j = 1, \cdots, J$ do
    $\boldsymbol{\theta}_j \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(\theta)}_j)$
for each mutation $n = 1, \cdots, N$ do
    $U_n \sim \text{Categorical}(\boldsymbol{\pi})$
    $z_n \sim \text{Categorical}(\boldsymbol{\theta}_{U_n})$
    $\boldsymbol{\zeta} = \{1/C_n^{(\text{major})}, \cdots\}, |\boldsymbol{\zeta}| = C_n^{(\text{major})}$
    $M_n \sim \text{Categorical}(\boldsymbol{\zeta})$
    $x_n \sim \text{Categorical}(\boldsymbol{\phi}_{z_n})$
    draw $B_n$ according to Equation (1)
```

Inference algorithm

Estimation of hidden variables and parameters. We predicted the responsibilities of the latent variables using the collapsed variational Bayesian (CVB) inference, and for marginalized parameters (i.e., $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$), we computed the estimates using the expected values obtained from the approximate posterior distributions and hyper-parameters. $q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M})$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\theta})$ were the approximated posterior distributions of the latent variables and the corresponding parameters. Regarding latent variables, when we used the mean-field approximation, such as $q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M}) \approx q(\boldsymbol{z})q(\boldsymbol{U})q(\boldsymbol{M})$, preliminary experiments showed that the prediction accuracy was significantly worse. Therefore, while preserving the structure among the latent variables (i.e. joint posterior $q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\theta})$ is factorized into $q(\boldsymbol{\pi})q(\boldsymbol{\theta})\prod_{n=1}^{N} q(z_n, U_n, M_n)$), we derived the objective function, ELBO, as follows:

$$F[q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M})] = \sum_{n=1}^{N}\sum_{z_n}\sum_{U_n}\sum_{M_n} q(z_n, U_n, M_n)$$
$$\times \log \frac{p(x_n, B_n, z_n, U_n, M_n \mid D_n, C_n, P, \boldsymbol{\lambda}, \boldsymbol{\rho}, \boldsymbol{\alpha}^{(\pi)}, \boldsymbol{\alpha}^{(\theta)}, \boldsymbol{\phi})}{q(z_n, U_n, M_n)}. \tag{2}$$

According to this objective function, we can obtain the updated formula for $q(z_n, U_n, M_n)$ which takes a stationary point to give the extreme value of $F[q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M})]$ as follows:

$$\begin{aligned} &q(z_n = k, U_n = j, M_n = m) \\ &\propto \exp\Bigg[\log \phi_{k,x_n} + \log \frac{\Gamma(B_n + \mu_{n,j,m})\Gamma(D_n - B_n + \nu_{n,j,m})\Gamma(\rho_j)}{\Gamma(D_n + \rho_j)\Gamma(\nu_{n,j,m})\Gamma(\mu_{n,j,m})} \\ &+ \log \frac{\alpha^{(\theta)}_{j,k} + \sum_{n' \neq n}\sum_{M_{n'}} q(z_{n'} = k, U_{n'} = j, M_{n'})}{\sum_{k'}\left\{\alpha^{(\theta)}_{j,k'} + \sum_{n' \neq n}\sum_{M_{n'}} q(z_{n'} = k', U_{n'} = j, M_{n'})\right\}} \\ &+ \log \frac{\alpha^{(\pi)}_j + \sum_{n' \neq n}\sum_{z_{n'}}\sum_{M_{n'}} q(z_{n'}, U_{n'} = j, M_{n'})}{\sum_{j'}\left\{\alpha^{(\pi)}_{j'} + \sum_{n' \neq n}\sum_{z_{n'}}\sum_{M_{n'}} q(z_{n'}, U_{n'} = j', M_{n'})\right\}}\Bigg] \end{aligned} \tag{3}$$

where $\Gamma(\cdot)$ denotes the gamma function and

$$\mu_{n,j,m} = \rho_j \times \lambda_j \times \eta_{n,m} \text{ and } \nu_{n,j,m} = \rho_j(1 - \lambda_j \times \eta_{n,m}). \tag{4}$$

The detailed derivation is provided in Section S1. For $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, by taking the expected values with respect to $q(z_n, U_n, M_n)$ and hyper-parameters, we can estimate the following:

$$\pi_j \propto \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{C_n^{(\text{major})}} q(z_n = k, U_n = j, M_n = m) + \alpha^{(\pi)}_j \tag{5}$$

$$\theta_{j,k} \propto \sum_{n=1}^{N}\sum_{m=1}^{C_n^{(\text{major})}} q(z_n = k, U_n = j, M_n = m) + \alpha^{(\theta)}_{j,k}. \tag{6}$$

After estimating the responsibility and parameters, we calculated the expected CCF for each mutation as follows:

$$\mathbb{E}[\text{CCF}_n] = \frac{B_n}{D_n} \times \frac{P \times C_n^{(\text{tumor})} + (1-P) \times C_n^{(\text{normal})}}{P \times \mathbb{E}[M_n]} \tag{7}$$

where

$$\mathbb{E}[M_n] = \sum_{m=1}^{C_n^{(\text{major})}} m \times \left\{\sum_{k=1}^{K}\sum_{j=1}^{J} q(z_n = k, U_n = j, M_n = m)\right\}.$$

Estimation of hyper-parameters. We predicted the hyper-parameters, including $\boldsymbol{\alpha}^{(\pi)}$, $\boldsymbol{\alpha}^{(\theta)}$, λ_j and ρ_j, using fixed-point iterations to maximize ELBO. In fixed-point iterations, we first derived the lower bound of ELBO using the gamma function and used the stationary points to maximize it for each parameter. For the ρ_j update to control overdispersion, we employed the exponential distributions $p(\rho|\tau) = \tau\exp(\tau\rho)$ as priors to achieve a stable estimation and confirmed that this modification prevented divergence of parameter learning. We have provided details of the updated formulas for all hyper-parameters in the Supplementary Data (see Equations (S5)-(S12)).

Model selection using the variational Bayesian (VB) inference. To select a plausible number of clones, J, we used ELBO as the criterion for model selection in this framework. In CVB, although parameter estimation is possible using Equations (3)–(6), we could not explicitly calculate the ELBO value (as shown in Equation (2)) because the approximate posteriors of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ were marginalized. Therefore, we used the VB method to derive the ELBO and performed tentative parameter estimation for model selection, which yielded the predicted number of clones for each tumor in advance of CVB. In the VB method, ELBO was formulated as shown in Equation (S4), and we computed this value using the predicted parameters according to the update rules: Equations (S5)–(S7).

Different properties of SigTracer from CloneSig. In terms of modeling, one difference was that we prepared Dirichlet distributions as priors of signature activities for each clone ($\boldsymbol{\theta}_j$) and the clone proportion ($\boldsymbol{\pi}$). These categorical distribution parameters yielded z_n and U_n, and CloneSig predicted these via an EM algorithm. Therefore, the update formula of the EM algorithm was equivalent to that of the VB method when the priors were Dirichlet distributions with all parameters set to 1.0. Another improvement in modeling was in terms of the setting for CCF overdispersion (ρ). CloneSig set the same value for all clones, whereas SigTracer controlled the overdispersion by each clone for the beta distribution, which is the prior of binomial distributions that determined B_n against D_n.

Regarding an inference algorithm, SigTracer adopted ELBO as a criterion to select the number of clones instead of BIC used in CloneSig. In addition, we used fixed-point iteration to estimate hyper-parameters instead of the projected Newton method used by CloneSig.

Summary of the algorithm to infer parameters. The bottleneck of inference for both CVB and VB was calculated using $q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M})$, and the following time complexity $\mathcal{O}(NKJ \times \max(\boldsymbol{C}^{(\text{major})}))$. To terminate CVB learning, we used the approximated ELBO. As described above, ELBO derived using CVB could not be calculated because the posteriors of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ were marginalized; hence, we approximated ELBO by substituting the expected value of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ into Equation (S4). This value could not be used for model selection because it was not the objective of CVB, but it was used to confirm if training was saturated. This inference could not possibly converge to the global optimal solution because it is a non-convex optimization. To address this potential problem, we initialized the parameters five times in the following experiments and subsequently selected the solution with the highest approximated ELBO. Finally, we created Algorithm 1 to summarize the inference.

Algorithm 1 Parameter estimation of SigTracer

```
Require $N, K, V, J_{\max}, \boldsymbol{x}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{C}, \boldsymbol{\phi}, P$
$\mathcal{H} \equiv \{\boldsymbol{\alpha}^{(\pi)}, \boldsymbol{\alpha}^{(\theta)}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \tau\}$
Model selection with VB :
for each clone number $J = 1, \cdots, J_{\max}$ do
    Iterate to estimate $q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M}), q(\boldsymbol{\pi}), q(\boldsymbol{\theta})$ and $\mathcal{H}$
                                        by Eqs.(S5)-(S12)
    Derive ELBO for $J$ by Eq.(S4)
The predicted number of clones $\hat{J} \leftarrow \arg\max$ ELBO
Parameter estimation with CVB :
Initialize $q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M})$ and $\mathcal{H}$ with $\hat{J}$
for each iteration $r = 1, \cdots, 1000$ do
    for each mutation $n = 1, \cdots, N$ do
        Update $q(z_n, U_n, M_n)$ by Eq.(3)
    Update $\boldsymbol{\pi}, \boldsymbol{\theta}$ by Eqs. (5) and (6)
    Calculate approximate ELBO by Eq. (S4)
    if $r \geq 500$ and ELBO converged then
        Terminate iteration
    Update $\mathcal{H}$ by Eqs. (S5)-(S12)
for each mutation $n = 1, \cdots, N$ do
    Calculate $\mathbb{E}[\mathrm{CCF}_n]$ by Eq. (7)
```

Statistical test for measuring the relationship between mutations and signatures

We assumed that a particular somatic mutation drove tumor evolution and induced mutational processes of an individual signature, or conversely, a certain mutational process caused particular mutations. In that case, such mutations were likely to be concentrated in clones in which the relevant signature was highly active. Based on this idea, we implemented the following pipeline of statistical tests to measure the relevance between mutations and signatures at a genetic level.

First, for all mutations, we determined the clone $\hat{j}$ to which the mutation n belonged, as follows:

$$\hat{j} = \arg\max_{j} \left\{ \sum_{k=1}^{K} \sum_{m=1}^{C_n^{(\text{major})}} q(z_n = k, U_n = j, M_n = m) \right\}.$$

The next step was to determine the active signature k for each clone j based on whether or not the following threshold was satisfied:

$$\sum_{n=1}^{N} \sum_{m=1}^{C_n^{(\text{major})}} q(z_n = k, U_n = j, M_n = m) \geq 100.$$

We divided all clones into active or inactive groups in terms of the signature k to be tested according to the above threshold, and we calculated the ratio of active group sizes for the signature k against all groups (denoted as r_k) by adding the number of mutations in the clones belonging to each group.

If we knew where the mutations occurred at the genetic level, we estimated whether the clones carrying mutations on a certain gene were active or inactive for a particular signature k according to the above procedures. If there was no

Table 1. Artificial datasets and the setting of essential parameters

ID	Parameter setting and remarks
WGS-1	Baseline that mimics **whole-genome** sequenced tumors, $N = 2000$, $J = 2$, $K \sim 2 + \text{Poisson}(3)$, $D_n \sim \lfloor\text{Uniform}(5, 50)\rfloor$, $\rho_j \sim \text{Uniform}(5, 100)$
WGS-2	$J = 1$, following WGS-1 for other parameters
WGS-3	$J = 3$
WGS-4	$J = 4$
WGS-5	ρ_j is equal between clones
WGS-6	$K \sim 2 + \text{Poisson}(5)$
WGS-7	$K \sim 2 + \text{Poisson}(1)$
WES-1	Baseline that mimics **whole-exome** sequenced tumors, $N = 1000$, $J = 2$, $K \sim 2 + \text{Poisson}(3)$, $D_n \sim \lfloor\text{Uniform}(50, 500)\rfloor$, $\rho_j \sim \text{Uniform}(5, 100)$
WES-2	$N = 500$, following WES-1 for other parameters
WES-3	$N = 200$
WES-4	$N = 100$
WES-5	ρ_j is equal between clones
Ideal-1	$N = 2000$, $J = 3$, $K \sim 2 + \text{Poisson}(3)$, $D_n \sim \lfloor\text{Uniform}(50, 500)\rfloor$, $\rho_j \sim \text{Uniform}(5, 100)$, CCF between clones is separated by >0.2
Ideal-2	$J = 4$, following Ideal-1 for other parameters

All datasets contained 100 samples, and the mutation profiles followed the generative process of SigTracer. N, K and J are the number of mutations, the number of active signatures and the number of clones, respectively. The datasets can be divided into three categories: WGS-mimicking (low coverage and a large number of mutations), WES-mimicking (high coverage and few mutations) and ideal (high coverage and a large number of mutations). The parameters of the datasets from WGS-2 to WGS-7 follow WGS-1, except the ones explicitly described in the table. Similarly, the parameters of the datasets from WES-2 to WES-5 follow WES-1.

relationship between the gene of interest and the signature k, the mutations would be distributed into active and inactive groups according to the ratio r_k. To test this null hypothesis, we performed a binomial test for all gene-signature combinations against a binomial distribution with a success probability of r_k. In the actual implementation, we used scipy.stats.binom_test with a significance level of $\alpha < 0.05$ in Python.

Datasets

Simulation data. To evaluate the performance of SigTracer, we artificially produced 14 datasets following the generative process in SigTracer. All datasets contained 100 samples, and we used 67 signatures registered in COSMIC ver3.1 as the reference mutational distributions. The datasets were divided into three categories: whole-genome sequencing (WGS) mimicking (low coverage and a large number of mutations), whole-exome sequencing (WES) mimicking (high coverage and a small number of mutations) and ideal (high coverage and a large number of mutations). Detailed information regarding how the datasets were produced is provided in Section S2, and Table 1 summarizes the differences in essential parameters for each dataset.

The single-cell sequenced ovarian cancer sample. As a proof of concept, we analyzed a single-cell sequenced ovarian cancer sample collected from ascites (22). This dataset includes three samples, OV2295 sequenced at diagnosis, and OV2295(R2) and TOV2295(R) sequenced at relapse. Since the removal of cancer cells targeted by chemotherapy result in severe noise for the observed VAF of a relapsed tumor, we have chose the primary tumor, OV2295, as the representative to be used for our analyses. Then, we considered all the sequenced cells in OV2295 as a pseudo-bulk sample for SigTracer analysis. The raw data used in this experiment can be downloaded from https://zenodo.org/record/3445364.

Application using blood cancer samples. We applied SigTracer to blood cancer samples of the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort. Tumors used in this study were subjected to whole-genome sequencing, and the tumor types were roughly divided into two categories: chronic lymphocytic leukemia (CLL) and B-cell non-Hodgkin lymphoma (BNHL), including Burkitt lymphoma, diffuse large B-cell lymphoma, follicular lymphoma, and marginal lymphoma. All sources are available for download from the ICGC Data Portal: https://dcc.icgc.org/releases/PCAWG. We obtained 95 CLL samples and 100 BNHL samples. Similar to the simulation, we used COSMIC ver3.1 as the reference SBS signature set: https://cancer.sanger.ac.uk/cosmic/signatures/SBS/index.tt. For each tumor, previous studies using SigProfiler reported the types of active signatures (23); four signatures, SBS1, SBS5, SBS9 and SBS40, were active in CLL, and 14 signatures, SBS1, SBS2, SBS3, SBS5, SBS6, SBS9, SBS13, SBS17a, SBS17b, SBS34, SBS36, SBS37, SBS40 and SBS56, were active in BNHL. This result is available in synapse.org ID syn11801889: https://www.synapse.org#!Synapse:syn11804040, and we utilized them as the model input. Furthermore, the CNA and purity were estimated using several methods and are provided in the PCAWG database (24); we also used these as the model input. In addition, when we applied the statistical test described above, we utilized the locus of all somatic mutations from mapping results uploaded on the PCAWG database, which were already annotated using Hugo symbols.

RESULTS AND DISCUSSION

Simulation experiments

Evaluation of model selection. To evaluate the model selection performance and compare SigTracer with CloneSig, we applied each method to the datasets presented in Table 1. We could easily identify active signatures in each sample using different fitting methods. Hence, we assumed that the type and number of signatures were known in this simulation. In addition, we elected $J_{\min} = 1$ and $J_{\max} = 5$ as the range of the clone number. Table 2 summarizes the results of model selection. Out of the 100 samples in each dataset, the bold characters indicate the number of samples for which the correct number of clones could be estimated using each method. We have presented the results of WES-2 to WES-5 in Supplementary Table S2.

Table 2 shows that SigTracer consistently estimated the correct number of clones compared to CloneSig for $J \leq 2$. When a tumor included multiple clones, such as WGS-3 and WGS-4, the CloneSig estimation was more accurate than the SigTracer estimation. However, with the Ideal-1/2 datasets (high coverage and a high number of mutations),

Table 2. The number of clones predicted with artificial mutation profiles

ID	Method	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$
WGS-1	SigTracer	16	**81**	3	0	0
	CloneSig	1	**25**	42	23	8
WGS-2	SigTracer	**87**	13	0	0	0
	CloneSig	**23**	44	22	10	1
WGS-3	SigTracer	6	72	**21**	1	0
	CloneSig	0	12	**39**	35	14
WGS-4	SigTracer	0	63	32	**5**	0
	CloneSig	1	7	36	**38**	18
WGS-5	SigTracer	23	**77**	0	0	0
	CloneSig	0	**30**	49	20	1
WGS-6	SigTracer	19	**80**	1	0	0
	CloneSig	1	**31**	43	19	6
WGS-7	SigTracer	7	**91**	2	0	0
	CloneSig	0	**25**	49	21	5
WES-1	SigTracer	12	**82**	6	0	0
	CloneSig	1	**21**	27	17	34
Ideal-1	SigTracer	0	20	**78**	2	0
	CloneSig	0	0	**21**	30	49
Ideal-2	SigTracer	0	8	50	**39**	3
	CloneSig	0	0	7	**29**	64

All IDs indicate the name of simulation datasets which are provided in Table 1. Columns with $J = 1 \sim 5$ indicate the number of samples whose estimated numbers are J. Out of the 100 samples in each dataset, the bold characters indicate the number of samples for which the correct number of clones could be estimated by each method.

SigTracer estimated the correct number of clones in numerous samples. Even with an ideal dataset like Ideal-1/2, CloneSig predicted a larger number of clones than was true, indicating that BIC did not work correctly with the singular model. In addition, the CloneSig implementation adopted heuristics to compensate for the degrees of freedom in BIC, which might not be suitable for these cases. Notably, SigTracer exhibited better performance of model selection than CloneSig using the ideal datasets.

To investigate whether the tendency of SigTracer to estimate a small number of clones for the datasets with low coverage and many clones could be improved, we calculated the log-likelihood for each sample. Using artificial data, we could derive the 'true' likelihood because we knew the correct latent variables ($\boldsymbol{z}$, $\boldsymbol{U}$ and $\boldsymbol{M}$) and true parameters ($\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$). Figure 2 shows the comparison between the log-likelihood based on the estimated parameters and the true log-likelihood for WGS-3, WGS-4, Ideal-1 and Ideal-2. Figure 2A and C show that the log-likelihood of SigTracer with fewer clones than the true number in WGS exceeded the log-likelihood with the true clone composition. In contrast, in the ideal datasets, the number of clones required by SigTracer was more than or equal to the true number to exceed the true log-likelihood in many samples. This result indicated that accurate clone number estimation in low-coverage data was challenging using the criteria based on the likelihood including BIC and ELBO.

Through these experiments, we highlighted the quantitative limitations of the current model for low-coverage data. However, its usefulness was still high, as evidenced by the fact that SigTracer could accurately estimate the clone number for numerous samples under an ideal setting. This method could become more critical as the number of high-coverage data will increase with the development of sequencing technologies in the future.

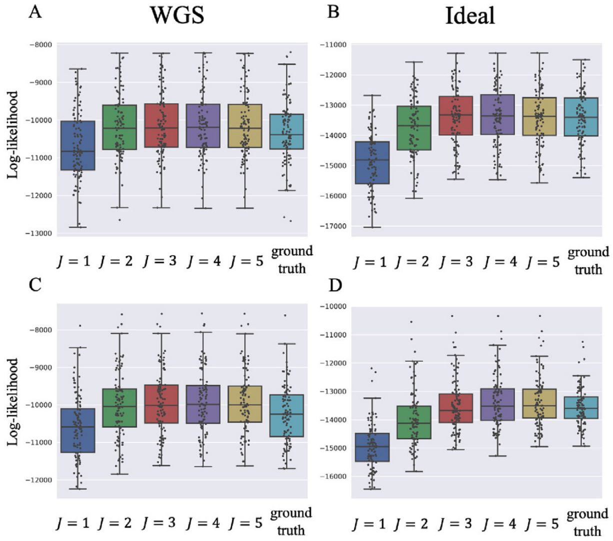

Figure 2. Log-likelihood with the estimated and ground-truth parameters in SigTracer. Each scatter plot shows the log-likelihood of each sample with the estimated and ground-truth parameters for $q(\boldsymbol{z}, \boldsymbol{U}, \boldsymbol{M})$, $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$. For the estimated parameters, box plots are drawn separately for the clone number. (**A**–**D**) differ in terms of the dataset, and they show the results of WGS-3, Ideal-1, WGS-4 and Ideal-2, respectively. The correct number of clones in (A) and (B) is $J = 3$, and that in (C) and (D) is $J = 4$.

Evaluation of clone estimation. Next, we examined the accuracy of parameter estimations by SigTracer and CloneSig using the artificial datasets shown in Table 1. As measures of the estimation accuracy, we focused on whether they could correctly estimate CCF ($\boldsymbol{\lambda}$) and the signature activity ($\boldsymbol{\theta}$) for each clone j. In this simulation, we provided the true number of clones and only evaluated the parameter estimation performance.

We denoted $\hat{\boldsymbol{\lambda}} = \{\hat{\lambda}_j\}_{j=1}^{J}$ as the true CCF. For a single tumor, we defined the minimum value obtained by summing the CCF distance $| \lambda_j - \hat{\lambda}_{j'} |$ considering all possible combinations of the true and predicted clones as the evaluation criterion. For the signature activity by each clone, we denoted $\hat{\boldsymbol{\theta}}_j \in \mathbb{R}^K$ as the true activity for the j-th clone. We calculated the sum of the cosine distanced between $\boldsymbol{\theta}_j$ and $\hat{\boldsymbol{\theta}}_{j'}$ for all combinations, similar to the case of CCF, and used the minimum value among them as the evaluation criterion. Both criteria were desired to be small.

Figure 3 shows the results for a part of WGS dataset. We evaluated the results on an average for 100 samples included in each dataset. We have presented the results for other datasets in Supplementary Figure S1. For CCF, except for WGS-4 with many clones and low coverage, SigTracer achieved comparable or better accuracy than CloneSig. In addition, the accuracy in terms of signature activity for each clone of SigTracer was comparable or better than that of CloneSig for all datasets. Finally, we have summa-

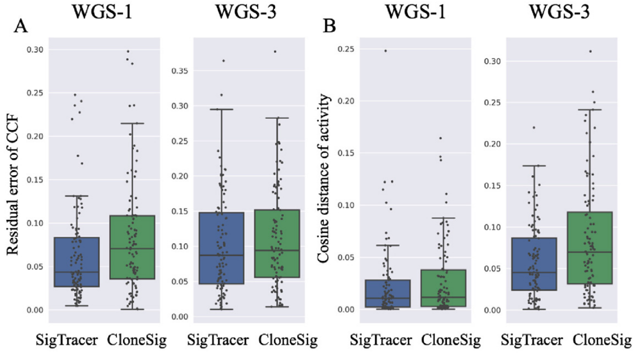

Figure 3. Accuracy of estimation for CCF and signature activity for each clone with the WGS-1/3 datasets. For the names of datasets, refer to Table 1. Panels (**A** and **B**) show the residual error of the estimated CCF and the cosine distance of the estimated activity against the ground-truth values, respectively (the lower values are better). In each panel, the left box plots are the results of SigTracer, and the right box plots are those of CloneSig. Each scatter plot shows the results of each sample included in each dataset. The results of other datasets are shown in Supplementary Figure S1.

rized the numerical statistics, including the mean and median values in Supplementary Tables S3 and S4 for all simulation results.

These improvements in accuracy were due to the differences in modeling as explained in Materials and Methods section. A clear example is the comparison between WES-1 and WES-5 shown in Supplementary Figure S1. The only difference between these two datasets was whether the variance was set for each clone or not. SigTracer outperformed CloneSig in numerous samples in WES-1 in terms of CCF estimation, whereas there was almost no difference for WES-5. These differences were caused by the fact that SigTracer prepared the overdispersion parameters for each clone. Besides, in WGS with low coverage, SigTracer outperformed both CCF and the signature activity estimation for WGS-5 with the same variance between clones, suggesting that other modifications also contributed to the improvement.

Evaluation of signature assignment. In terms of signatures, it is vital to analyze their accuracy of generating the corresponding mutations. Therefore, we applied six pipelines including SigTracer, CloneSig, SigLASSO, deconstructSigs, Ccube + SigLASSO and Ccube + deconstructSigs to some of the datasets shown in Table 1 (WGS-1/2, WES-1/2, Ideal-1/2) to compare the accuracy of signature assignment. SigLASSO (19) and deconstructSigs (18) are frequently used signature fitting methods. Ccube (7) is a recently developed method of clonal decomposition based on CCF. Ccube + SigLASSO and Ccube + deconstructSigs consider the clones estimated by Ccube as new samples and subsequently apply SigLASSO and deconstructSigs, respectively. The posterior probability of a signature producing the corresponding mutation is calculated in all pipelines for all mutations. We compared their accuracy as shown in Supplementary Table S5. Consequently, SigTracer exhibited the best accuracy for all datasets. Interestingly, Ccube + SigLASSO and Ccube + deconstructSigs exhibited poor accuracy compared to that of SigLASSO and deconstructSigs. This could be primarily attributed to the classification of mutations by Ccube that reduces the number of mutations contained per clone and does not provide a sufficient number of mutations for analysis by SigLASSO and deconstructSigs. These results suggest that a joint modeling-based approach like SigTracer and CloneSig is necessary to effectively utilize VAF in signature fitting.

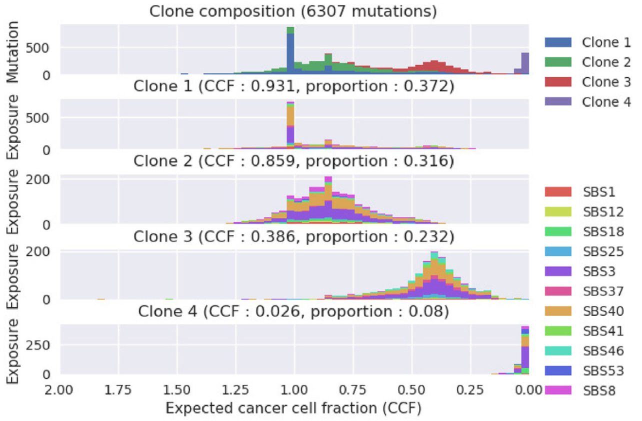

Figure 4. SigTracer output example with the single-cell sequenced ovarian cancer sample. The number of clones included in this tumor was estimated to be $J = 4$. The horizontal axis in all panels shows the expected cancer cell fraction. The top panel shows a histogram of all mutations with regard to the expected CCF, and it is color-coded to indicate the clone that the mutations belong to. Subsequent panels show histograms of mutations included in each clone, and they are color-coded to indicate which signature exposure is responsible.

Analysis of the single-cell sequenced ovarian cancer sample

We applied SigTracer to the single-cell sequenced ovarian cancer sample as described above. For a preliminary analysis, we applied SigLASSO (19) to determine the active signature set and used them as the input signatures for the application of SigTracer. Figure 4 shows an example of the output from SigTracer, in which the expected value of CCF for mutation n was computed using Equation (7) and was visualized using a histogram of latent variables. SigTracer has succeeded in decomposing mutations to clones with different CCF.

When using single-cell data, preliminary mutation clustering using predicted copy number may enhance the interpretability. A previous study has clustered all the sequenced cells into four populations based on predicted copy numbers (25). Then, we considered the clustered cell populations as pseudo-bulk samples and applied SigTracer to these. Supplementary Figure S2A–D shows the results of SigTracer against the four cell populations. SigTracer-based analysis revealed two clones from all four cell populations, which could be classified into primary clones that arise in the early stages of carcinogenesis (i.e. with higher CCF) and subclones that arise in the later stages (i.e., with lower CCF). Although these experiments were performed independently between cell populations, similar signature activities of the primary clones were observed. In fact, when we performed hierarchical clustering of these clones based on the cosine distance of the predicted activity, the primary clones accumulated in single cluster as shown in Supplementary Figure S2E. This result suggests that each cell population acquired a common primary clone and subsequently branched off to acquire a subclone that characterizes each cell population. It is a reasonable result that captures branching evolutionary process. Supplementary Figure S2E also shows that if we consider all the cells as one pseudo-bulk sample like Figure 4, signature activities of every clone including three subclones (denoted as Figure 4-Sub1, 2 and 3 in Supplementary Figure S2E) are similar to the primary one of each cell population. This indicates that the information of clustered cell populations increased the resolution of subclones in SigTracer's analysis, and such preprocessing is important in the analysis with real data.

Real data analysis with blood cancer samples

We applied SigTracer to CLL and BNHL samples described in the Materials and Methods section. For model selection, the predicted numbers of clones are summarized in Supplementary Figure S3. As observed in the simulation experiments, we must be aware that the predicted number of clones for WGS data might be lower than the actual number.

Before we focus on the specifics, we provide some evidence to ensure the reliability of our results. First, we showed the validity of the SigTracer extension for preparing overdispersion parameters by each clone using the results obtained for CLL. Figure 5 summarizes the visualization of the reconstructed VAF distribution based on the estimated parameters when SigTracer was applied to a certain CLL sample with the number of clones, $J = 3$. Figure 5 A shows the observed VAF (i.e. a histogram for B_n/D_n). Intuitively, when the VAF distribution was observed for a given number of clones, $J = 3$, it was desirable to decomposed the distribution into three elements with the expected VAF values of 0.1, 0.25 and 0.55. Figure 5B shows the result obtained when the variance of VAF was controlled by each clone (i.e. the model includes the extension of interest), which yielded an intuitively correct decomposition. In contrast, Figure 5C shows the result of SigTracer with the same value of overdispersions between clones; SigTracer predicted two clones with an expected VAF of almost 0.0. This tendency was not due to a simple difference in the initial values because the same results were obtained even after changing the initial values 10 times. The model with the same overdispersion parameters between clones could not capture the difference in the accumulation of mutations around the expected VAF. For instance, the red clone mutations in Figure 5B intensively occurred when the VAF was approximately 0.1, and the variance was smaller than those of the other two clones. However, in Figure 5C, this clone is shown to be split into two clones to adjust the variance between clones. Similar to this sample, we observed the effectiveness of the SigTracer extension in a number of CLL/BNHL samples.

Next, we provide quantitative evidence that the clone decomposition by SigTracer for actual data was reliable. Because all signature-based methods include unsupervised learning and we cannot find true parameters, rigorously verifying the correctness of the estimated results for real data is not easy. However, based on the estimated parameters, we could quantify how accurately the model represents original data. We defined a measure called the 'reconstruction rate: *RR*' to evaluate how many observed variables could be reconstructed. This measure was calculated by each sample. Two types of *RR*: RR_{mutation} were used to indicate how well the model reconstructed the mutation type (the *V*-dimensional categorical distribution) and RR_{VAF} to show how well the model reconstructed the VAF distribution (the beta mixture distribution); detailed definitions are provided in Section S5. For instance, the RR_{VAF} calculation included quantifying the overlapping histograms in Figure 5A and B. By comparing the *RR* values calculated from the estimators using artificial and real data, we can see whether SigTracer was suitable for real data. In other words, if the *RR* value for the simulated data was close to that for real data, we could indirectly state that the real data could be represented by SigTracer. Supplementary Table S6 summarizes the *RR* values calculated from all experiments. Although the *RR* values of real data (CLL and BNHL) tended to be lower than those of the simulated data that followed the generative process of SigTracer completely, we found the results of this study to be reliable to some extent. In summary, SigTracer is effective for not only artificial data but also for actual data.

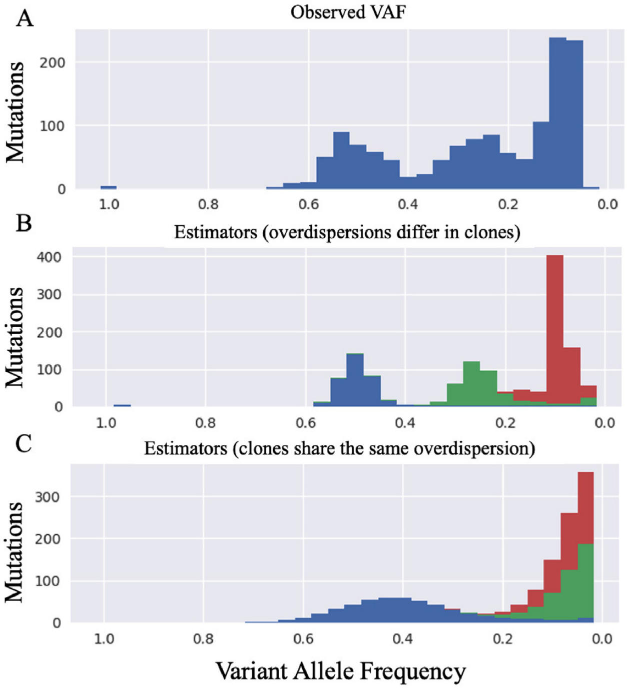

Figure 5. Comparison of the observed VAF in a certain CLL sample and the expected VAF based on the estimators of SigTracer. The horizontal axis shows the VAF, and all panels show histograms for all mutations with respect to VAF. Panel (**A**) is based on the observed VAF (i.e. a histogram with regard to B_n/D_n). Panels (**B** and **C**) are histograms based on the estimated parameters under the observation of (A). In panel (**B**), SigTracer prepared overdispersion parameters for each clone ($J = 3$), and the figure is based on the estimators under those settings. In contrast, panel (C) is drawn based on the estimators obtained when SigTracer shares the same overdispersion across clones.

SBS9 is initially active in CLL. SBS9 is the signature related with activation-induced cytidine deaminase (AID) or

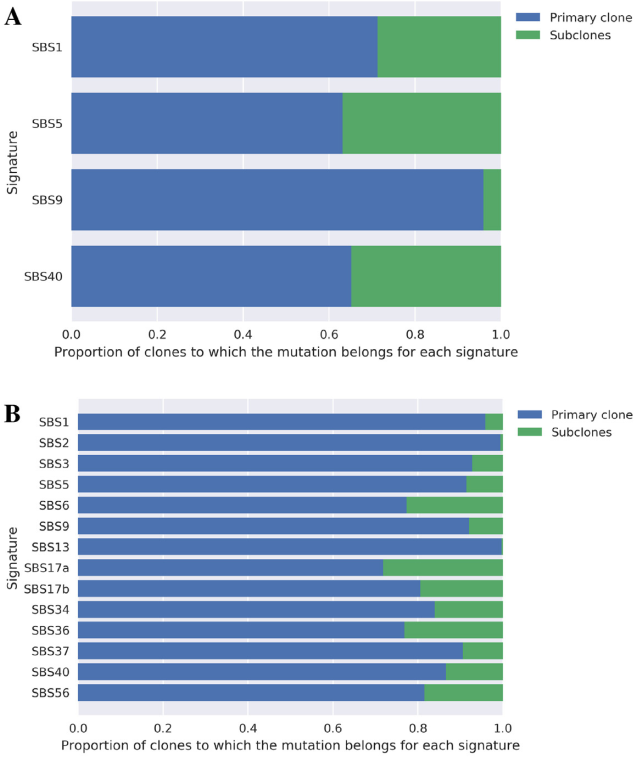

Figure 6. Types of clones that are likely to carry the mutations contributed by each signature. Panels (**A** and **B**) show the results of CLL and BNHL, respectively. The clones were divided into two groups primary and subclones—according to the procedure described in the text. This figure summarizes which signature-derived mutations they have. In CLL, the mutations in subclones contain a few SBS9-derived ones (cf. the third bar in panel A).

polymerase eta working in association with AID (9,26). The estimated CCF of the clones with high SBS9 activity tended to be close to 1.0 for most cases among the CLL samples. To quantify this trend, we divided the predicted clones into two categories: primary clones and subclones. We defined primary clone as clones with the largest size (π_j) among clones with a CCF >0.95 in one tumor. If no clone with a CCF >0.95 existed, we defined the clone with the highest CCF in that tumor as the primary clone. Subclones were all clones except the primary clone. When we calculated which of the two types of clones was more likely to carry the mutations attributed to each signature from the estimated responsibility, we found that SBS9 was particularly active in the primary clone in CLL compared with other signatures (Figure 6A). This result was consistent with that of a previous report (24). In contrast, this tendency was not observed in BNHL samples, and SBS9 was active in both primary and subclones (Figure 6B); hence, it remains to be elucidated what caused such a difference between the two types of blood cancer samples.

Somatic mutations strongly associated with each signature. Using the statistical test pipeline described in the Materials and Methods section, we attempted to comprehensively identify mutations associated with each signature in CLL and BNHL samples at the genetic level. First, as a proof of concept, we have summarized the list of genes that were significantly associated with SBS9 in CLL samples in Table 3. We adopted the multiple test correction with familywise error rate (FWER) to avoid false positives. At the first glance, we observed that mutations were concentrated in the IG region coding immunoglobulin. AID, the mutational process of SBS9, is required for class-switching of immunoglobulin (27); thus, it is reasonable that mutations from SBS9 are concentrated in this region. Furthermore, Supplementary Table S7 summarizes the mutated regions associated with other signatures active in CLL. SBS1, SBS5 and SBS40 were all considered clock-like signatures (28); therefore, it was unlikely that these signatures would act on specific regions of the genome. As shown in Supplementary Table S7, the number of mutated regions associated with these signatures was reasonably small compared to that for SBS9, the mutational process for which a specific target region existed.

Table 3. Frequently mutated regions in clones with high SBS9 activity for CLL samples

Hugo symbol	Variant classification	*P*-value < FWER
IGHJ6	RNA	1.22E-33
IGLL5	5'Flank	8.13E-18
IGLL5	Intron	5.82E-17
BCL6	Intron	4.25E-15
AC018717.1	lincRNA	3.98E-14
AC096579.7	RNA	3.60E-11
IGLV3-25	RNA	3.82E-11
IGLV3-1	RNA	1.08E-10
IGKJ2	RNA	5.53E-10
KIAA0125	5'Flank	8.89E-10
Unknown	IGR	1.05E-09
IGKV4-1	RNA	1.76E-09
DMD	Intron	1.24E-08
TCL1A	Intron	1.60E-08
IMMP2L	Intron	6.96E-08
FSTL5	Intron	1.12E-07
IGKC	RNA	2.72E-07
IGKV1-5	RNA	2.81E-07
IGKJ3	RNA	4.28E-07
BCL2	5'UTR	1.42E-06
IGHD3-10	RNA	1.58E-06
FHIT	Intron	1.73E-06

This table lists the mutated regions significantly correlated with SBS9 in CLL samples detected according to the test pipeline. The left column indicates the gene name, and the middle column the region of the gene. They are listed in order of decreasing *P*-value, and only those that were determined to be significant by Bonferroni correction are shown.

We also applied the same test to BNHL results to investigate the relationship between signatures and mutations in blood cancer. Supplementary Table S8 shows the mutated regions that were significantly enriched with clones with high SBS9 activity. Combined with Table 3, the FWER-based test yielded only one region associated with SBS9 in common with CLL, which was an intronic region of FHIT. In addition, the mutations in the immunoglobulin-coding region were not enriched in the clone with high SBS9 activity for BNHL samples, which significantly differed from the result obtained for CLL samples.

Then, we focused on the mutated regions in both CLL and BNHL samples among those detected using a simple significance level $\alpha = 0.05$ rather than FWER. This ap-

Table 4. Significant GO terms related with SBS9 in both CLL and BNHL samples

GO biological process	Corresponding genes	FDR
Cell–cell junction organization	CDH12, CDH18, CDH19, CADM2	4.32E-02
↪ cell junction organization	+ CNTN5, GRID2, GRM5, PCLO	1.90E-03
cell–cell adhesion via plasma-membrane adhesion molecules	CDH12, CDH18, CDH19, GRID2, PCDH9, PCDH15, ROBO2	2.18E-04
↪ cell–cell adhesion	+ CTNNA2,LPP, NEGR1	3.77E-05
↪ cell adhesion	+ CADM2, CNTN5, CNTNAP5	7.55E-05
Regulation of neuron projection development	CSMB3, CTNNA2, EPHA7, GRID2, NEGR1, ROBO2, SEMA3A, SEMA3C	5.26E-03

The table shows results of GO analysis of the mutated regions of the clones with high activity of SBS9 in both CLL and BNHL samples. Among the 74 mutated regions, 32 were not mapped due to the lack of annotation (e.g. unannotated lncRNAs), and the remaining gene set was used for analysis.

proach did not include multiple testing corrections, which might have led to false positives. However, this ensured a certain degree of reliability because two results that were obtained independently were compared. All four signatures (SBS1, SBS5, SBS9 and SBS40) active in the CLL samples were also active in some of the BNHL samples, and we focused on these signatures. We have summarized the number of significantly mutated regions and their Hugo symbols in Supplementary Table S9 and S10. The number of mutated regions associated with SBS9 was high compared to other signatures, and interestingly, a missense mutation for KLHL6 was identified in the gene set corresponding to SBS9; thus, we could hypothesize that the etiology of SBS9, probably AID or polymerase eta, might be associated with KLHL6. Recent studies suggest that KLHL6 may be an essential tumor suppressor gene in B-cell lymphoma (29,30). Therefore, KLHL6 aberration may lead to high activity of the mutational process of SBS9, resulting in a large number of somatic mutations. However, further validations are required to reveal if this hypothesis is true because we only considered somatic mutations not including germline mutations lying on the actual genome in this study.

We performed gene ontology (GO) analysis on the resulting gene set to determine if there were any common features of each signature (GO Enrichment Analysis powered by PANTHER: http://geneontology.org/). Thus, significant terms were detected only for the gene set corresponding to SBS9, and the most interesting ones are shown in Table 4. For example, ‘cell junction organization’ and ‘cell adhesion’ annotations were detected, and they included cadherin-coding regions. These functions are closely related to the immune system, and cell adhesion is related to whether or not a tumor can acquire metastatic potential (31,32); hence, SBS9 activity might also be closely associated with a poor prognosis of blood cancer. Notably, these mutations also occurred in BNHL samples, compared to the coding region of immunoglobulin, where mutations were concentrated only in CLL samples. Because mutations on these genes were concentrated in the intronic region, we suspected that they were the consequence of SBS9 rather than the cause. Additionally, Supplementary Figure S4 shows a Venn diagram of the overlaps between signatures, indicating that the mutated regions for SBS9 were particularly unique.

Future work

SigTracer, which performs clone decomposition based on mutation signatures, showed significant potential of providing novel insights into mutational processes because of its improved accuracy of assigning mutations to signatures by considering VAF. However, the method suffers from some limitations. An issue in the modeling is that there is no established method that can predict the correct number of clones for arbitrary data. In our simulation experiments, we showed that the model achieved a sufficiently high likelihood with fewer clones than the true number for low-coverage data, which highlighted the limitation in terms of the amount of information in the input. To solve this problem, it is necessary to incorporate other data sources that are effective for inference in addition to mutation types and VAF. One possible solution is the extension of SigTracer to support multi-region sampling data. Clone decomposition against sequences obtained from multi-region sampling is less likely to neglect clones and provides higher resolution than that procured against bulk sequencing (33,34). By hierarchizing the parameter $\boldsymbol{\pi}$ representing the clone composition into the total number of regions, it is feasible to support multi-region sampling in SigTracer, and this is likely to gain importance with an increase in the number of multi-region sampling data.

Moreover, there is room for reconsideration of the type of signature used in the analysis of real datasets. SBS84 and SBS85, the existence of which has been demonstrated in the latest studies (23), are signatures known to be active in blood cancer. The inclusion of these signatures may lead to novel findings. When more signatures are included as active signature candidates, a phenomenon called ‘signature bleeding’ occurs, in which signatures that are not actually active are erroneously presumed to be active (35). To avoid such a situation, we must introduce a regularization to the model to reduce the number of active signatures used in the fitting method, such as sigLASSO (19).

Regarding the current test pipeline, we need to conduct further validation such as consideration of the evolutionary background of a tumor sample. For example, if tumors follow a branching evolutionary process (36), each tumor cell may have mutations that originate from multiple clones. In such cases, it is difficult to capture the causality of signatures due to mutations because there is a possibility that the active signature in a certain clone is activated by mutations belonging to other clones. Thus, the causality of signatures due to mutations can be detected using the method described here only if the tumor evolution follows a specific process, such as the Big Bang dynamics which state that cancer cells evolve independently (37), where each clone directly represents a mutational population carried by each cell. Recent reports concerning colorectal cancer support the neutral evolution and the Big Bang model (38,39), and future accumulation of knowledge on cancer evolution will resolve this issue.

CONCLUSION

We developed SigTracer, a signature-based method for estimating clonal evolution based on mutation types and VAF observed via bulk sequencing. In computational simulations, SigTracer outperformed the existing method in terms of model selection and accuracy for ideal artificial data. In addition, we applied SigTracer to CLL samples; the results were consistent with previous findings that SBS9, which is associated with AID or polymerase eta, intensively causes mutations in the immunoglobulin-coding regions. Furthermore, we performed the same analysis on BNHL samples using SigTracer and found that SBS9 also includes intensive mutation of the regions coding for cadherins and other genes regulating cell adhesion. Our results indicate that AID or polymerase eta activity may be induced in more regions related to the immune system than previously known. These new observations were obtained because of the improved accuracy of assigning mutations to signatures by considering not only mutation types but also VAF. We believe applying the proposed method to other cancer types may lead to the annotation of signatures for which mutational processes and target regions are unknown. Our results provide an excellent prospect for understanding the mechanism of carcinogenesis.

DATA AVAILABILITY

Our implementation of SigTracer in C++ and custom scripts in Python is available at GitHub repository: https://github.com/qkirikigaku/SigTracer.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Computation for this study was partially performed on the NIG supercomputer at ROIS National Institute of Genetics. We thank members at Hamada Laboratory for the valuable discussions regarding this study.

FUNDING

Ministry of Education, Culture, Sports, Science and Technology (KAKENHI) [JP17K20032, JP16H05879, JP16H01318, JP16H02484, JP16H06279, JP20H00624 to M.H.; JP20J20016 to T.M.].

Conflict of interest statement. None declared.

REFERENCES

1. Fittall,M.W. and Van Loo,P. (2019) Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Med.*, **11**, 20.
2. Roth,A., Khattra,J., Yap,D., Wan,A., Laks,E., Biele,J., Ha,G., Aparicio,S., Bouchard-Côté,A. and Shah,S.P. (2014) PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–398.
3. Miller,C.A., White,B.S., Dees,N.D., Griffith,M., Welch,J.S., Griffith,O.L., Vij,R., Tomasson,M.H., Graubert,T.A., Walter,M.J. *et al.* (2014) SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.*, **10**, e1003665.
4. Jiao,W., Vembu,S., Deshwar,A.G., Stein,L. and Morris,Q. (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinform.*, **15**, 35.
5. Deshwar,A.G., Vembu,S., Yung,C.K., Jang,G.H., Stein,L. and Morris,Q. (2015) PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, **16**, 35.
6. Gillis,S. and Roth,A. (2020) PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinform.*, **21**, 571.
7. Cmero,M., Yuan,K., Ong,C.S., Schröder,J., Corcoran,N.M., Papenfuss,T., Hovens,C.M., Markowetz,F. and Macintyre,G. (2020) Inferring structural variant cancer cell fraction. *Nat. Commun.*, **11**, 730.
8. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Campbell,P.J. and Stratton,M.R. (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.
9. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A.J.R., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
10. Rubanova,Y., Shi,R., Harrigan,C.F., Li,R., Wintersinger,J., Sahin,N., Deshwar,A. and Morris,Q. (2020) Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. *Nat. Commun.*, **11**, 731.
11. Harrigan,C.F., Rubanova,Y., Morris,Q. and Selega,A. (2019) TrackSigFreq: subclonal reconstructions based on mutation signatures and allele frequencies. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*. WORLD SCIENTIFIC, Singapore, pp. 238–249.
12. Antić,Ž., Lelieveld,S.H., van der Ham,C.G., Sonneveld,E., Hoogerbrugge,P.M. and Kuiper,R.P. (2021) Unravelling the sequential interplay of mutational mechanisms during clonal evolution in relapsed pediatric acute lymphoblastic leukemia. *Genes*, **12**, 214.
13. Abécassis,J., Reyal,F. and Vert,J.-P. (2021) CloneSig: joint inference of intra-tumor heterogeneity and mutational signatures' activity in tumor bulk sequencing data. *Nat. Commun.*, **12**, 1–16.
14. Watanabe,S. (2013) A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.*, **14**, 867–897.
15. Shiraishi,Y., Tremmel,G., Miyano,S. and Stephens,M. (2015) A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLOS Genet.*, **11**, e1005657.
16. Matsutani,T., Ueno,Y., Fukunaga,T. and Hamada,M. (2019) Discovering novel mutation signatures by latent dirichlet allocation with variational Bayes inference. *Bioinformatics*, **35**, 4543–4552.
17. Matsutani,T. and Hamada,M. (2020) Parallelized latent dirichlet allocation provides a novel interpretability of mutation signatures in cancer genomes. *Genes*, **11**, 1127.
18. Rosenthal,R., McGranahan,N., Herrero,J., Taylor,B.S. and Swanton,C. (2016) deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.*, **17**, 31.
19. Li,S., Crawford,F.W. and Gerstein,M.B. (2020) Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. *Nat. Commun.*, **11**, 3575.
20. Martincorena,I., Raine,K.M., Gerstung,M., Dawson,K.J., Haase,K., Van Loo,P., Davies,H., Stratton,M.R. and Campbell,P.J. (2017) Universal patterns of selection in cancer and somatic tissues. *Cell*, **171**, 1029–1041.

21. Dentro,S.C., Leshchiner,I., Haase,K., Tarabichi,M., Wintersinger,J., Deshwar,A.G., Yu,K., Rubanova,Y., Macintyre,G., Demeulemeester,J. *et al.* (2021) Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, **184**, 2239–2254.
22. Létourneau,I.J., Quinn,M.C., Wang,L.-L., Portelance,L., Caceres,K.Y., Cyr,L., Delvoye,N., Meunier,L., de Ladurantaye,M., Shen,Z. *et al.* (2012) Derivation and characterization of matched cell lines from primary and recurrent serous ovarian cancer. *BMC Cancer*, **12**, 379.
23. Alexandrov,L.B., Kim,J., Haradhvala,N.J., Huang,M.N., Tian Ng,A.W., Wu,Y., Boot,A., Covington,K.R., Gordenin,D.A., Bergstrom,E.N. *et al.* (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94–101.
24. Gerstung,M., Jolly,C., Leshchiner,I., Dentro,S.C., Gonzalez,S., Rosebrock,D., Mitchell,T.J., Rubanova,Y., Anur,P., Yu,K. *et al.* (2020) The evolutionary history of 2,658 cancers. *Nature*, **578**, 122–128.
25. Laks,E., McPherson,A., Zahn,H., Lai,D., Steif,A., Brimhall,J., Biele,J., Wang,B., Masud,T., Ting,J. *et al.* (2019) Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell*, **179**, 1207–1221.
26. Puente,X.S., Pinyol,M., Quesada,V., Conde,L., Ordóñez,G.R., Villamor,N., Escaramis,G., Jares,P., Beà,S., González-Díaz,M. *et al.* (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **475**, 101–105.
27. Muramatsu,M., Kinoshita,K., Fagarasan,S., Yamada,S., Shinkai,Y. and Honjo,T. (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, **102**, 553–563.
28. Alexandrov,L.B., Jones,P.H., Wedge,D.C., Sale,J.E., Campbell,P.J., Nik-Zainal,S. and Stratton,M.R. (2015) Clock-like mutational processes in human somatic cells. *Nat. Genet.*, **47**, 1402–1407.
29. Choi,J., Lee,K., Ingvarsdottir,K., Bonasio,R., Saraf,A., Florens,L., Washburn,M.P., Tadros,S., Green,M.R. and Busino,L. (2018) Loss of KLHL6 promotes diffuse large B-cell lymphoma growth and survival by stabilizing the mRNA decay factor Roquin2. *Nat. Cell Biol.*, **20**, 586–596.
30. Choi,J., Zhou,N. and Busino,L. (2019) KLHL6 is a tumor suppressor gene in diffuse large B-cell lymphoma. *Cell Cycle*, **18**, 249–256.
31. Bendas,G. and Borsig,L. (2012) Cancer cell adhesion and metastasis: selectins, integrins, and the inhibitory potential of heparins. *Int. J. Cell. Biol.*, **2012**, e676731.
32. Läubli,H. and Borsig,L. (2019) Altered cell adhesion and glycosylation promote cancer immune suppression and metastasis. *Front. Immunol.*, **10**, 2120.
33. Jamal-Hanjani,M., Wilson,G.A., McGranahan,N., Birkbak,N.J., Watkins,T.B., Veeriah,S., Shafi,S., Johnson,D.H., Mitter,R., Rosenthal,R. *et al.* (2017) Tracking the evolution of non–small-cell lung cancer. *New. Engl. J. Med.*, **376**, 2109–2121.
34. Liu,L.Y., Bhandari,V., Salcedo,A., Espiritu,S. M.G., Morris,Q.D., Kislinger,T. and Boutros,P.C. (2020) Quantifying the influence of mutation detection on tumour subclonal reconstruction. *Nat. Commun.*, **11**, 6247.
35. Maura,F., Degasperi,A., Nadeu,F., Leongamornlert,D., Davies,H., Moore,L., Royo,R., Ziccheddu,B., Puente,X.S., Avet-Loiseau,H. *et al.* (2019) A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.*, **10**, 2969.
36. Davis,A., Gao,R. and Navin,N. (2017) Tumor evolution: linear, branching, neutral or punctuated? *Biochim. Biophys. Acta*, **1867**, 151–161.
37. Sun,R., Hu,Z. and Curtis,C. (2018) Big bang tumor growth and clonal evolution. *CSH Perspect. Med.*, **8**, a028381.
38. Uchi,R., Takahashi,Y., Niida,A., Shimamura,T., Hirata,H., Sugimachi,K., Sawada,G., Iwaya,T., Kurashige,J., Shinden,Y. *et al.* (2016) Integrated multiregional analysis proposing a new model of colorectal cancer evolution. *PLOS Genet.*, **12**, e1005778.
39. Roerink,S.F., Sasaki,N., Lee-Six,H., Young,M.D., Alexandrov,L.B., Behjati,S., Mitchell,T.J., Grossmann,S., Lightfoot,H., Egan,D.A. *et al.* (2018) Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, **556**, 457–462.