Nucleic Acids Research, 2024, **52**, e62
https://doi.org/10.1093/nar/gkae527
Advance access publication date: 26 June 2024
Methods

CLEMENT: genomic decomposition and reconstruction of non-tumor subclones

Young-soo Chung [1,†], Seungseok Kang [1,†], Jisu Kim [2,3], Sangbo Lee [1] and Sangwoo Kim [1,*]

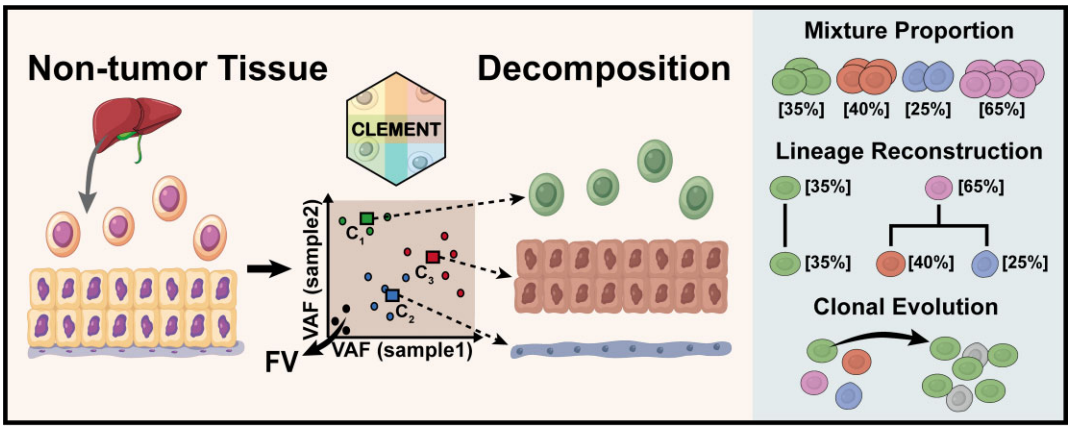
[1]Department of Biomedical Systems Informatics, Brain Korea 21 PLUS Project for Medical Science, Yonsei University College of Medicine, Seoul 03722, Republic of Korea
[2]DataShape team, Inria Saclay Île-De-France, Palaiseau 91120, France
[3]Department of Statistics, Seoul National University, Seoul 08826, Republic of Korea

[*]To whom correspondence should be addressed. Tel: +82 2228 2589; Fax: +82 2227 8308; Email: swkim@yuhs.ac
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Genome-level clonal decomposition of a single specimen has been widely studied; however, it is mostly limited to cancer research. In this study, we developed a new algorithm CLEMENT, which conducts accurate decomposition and reconstruction of multiple subclones in genome sequencing of non-tumor (normal) samples. CLEMENT employs the Expectation-Maximization (EM) algorithm with optimization strategies specific to non-tumor subclones, including false variant call identification, non-disparate clone fuzzy clustering, and clonal allele fraction confinement. In the simulation and *in vitro* cell line mixture data, CLEMENT outperformed current cancer decomposition algorithms in estimating the number of clones (root-mean-square-error = 0.58–0.78 versus 1.43–3.34) and in the variant-clone membership agreement (~85.5% versus 70.1–76.7%). Additional testing on human multi-clonal normal tissue sequencing confirmed the accurate identification of subclones that originated from different cell types. Clone-level analysis, including mutational burden and signatures, provided a new understanding of normal-tissue composition. We expect that CLEMENT will serve as a crucial tool in the currently emerging field of non-tumor genome analysis.

Graphical abstract



Introduction

Poly-clonality within a single specimen and its accurate decomposition have been important concerns in genomic analysis. Most research efforts to address this issue have focused on cancer, in which multiple subclones give rise to genetically distinct populations of a single tumor, resulting in intratumoral heterogeneity (ITH) that is responsible for drug resistance, tumor relapse, and poor clinical outcomes (1). Several methods, such as PyClone (2), SciClone (3), PyClone-VI (4) and QuantumClone (5), have been developed for the accurate decomposition and reconstruction of the cancer subclones. Although different statistical models and optimization strategies have been employed, the conceptual assumption is largely limited to the use of clonally expanded somatic mutations, which are clearly identifiable in conventional genome sequencing.

Recent advances in genomic analysis of non-tumor (i.e. normal) tissues pose a new challenge in genomic decomposition. Accurate clonal decomposition in normal tissue is necessary as it provides an understanding of the molecular-level landscape of the developmental process (6) or patterns of mosaicism (7). Additionally, clone-level analysis is applicable to various of non-cancer disease, such as early developmental disorders or borderline premalignancies (8,9). While both tumor and

Received: June 7, 2023. **Revised:** May 27, 2024. **Editorial Decision:** May 31, 2024. **Accepted:** June 12, 2024

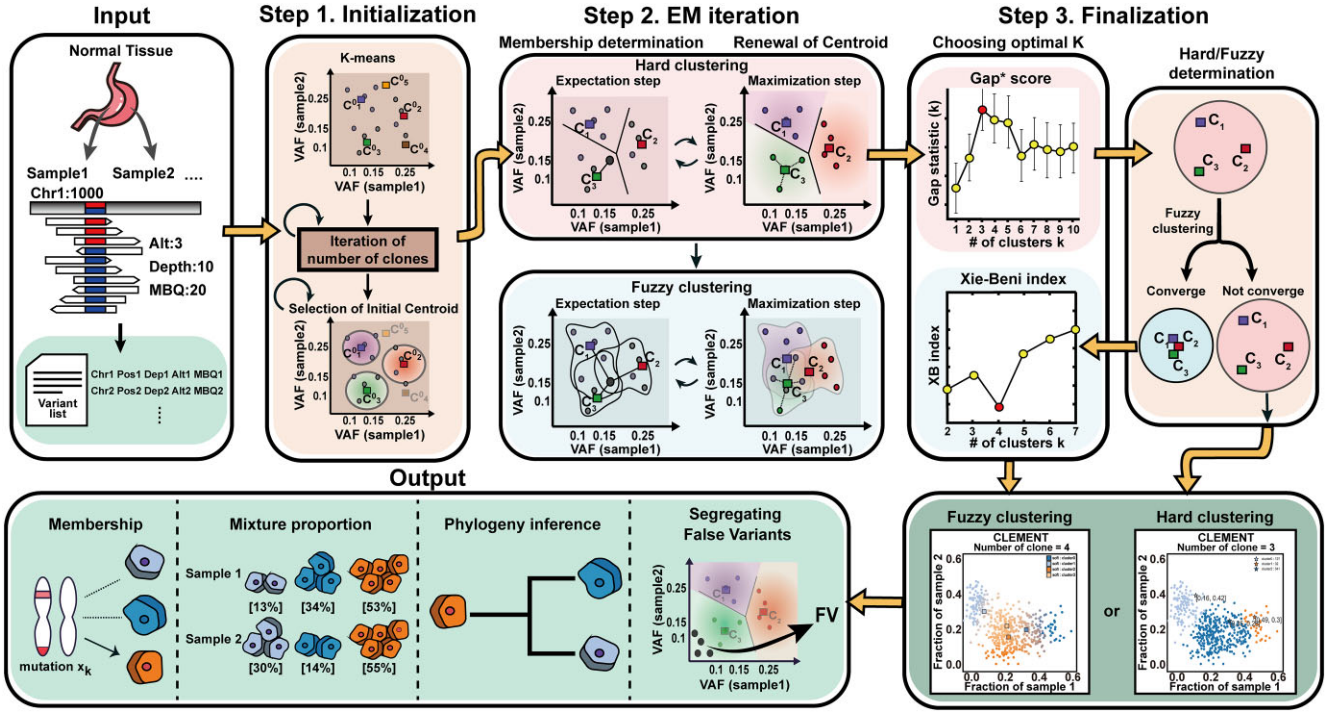


Figure 1. Overview of the CLEMENT workflow and core algorithms. Three steps (initialization, EM iteration and finalization) are depicted with defined input (variant information including total-, alternate read depth and base quality) and output (clone numbers, compositions, and variant membership). VAF: variant allele frequency.

non-tumor tissues in a single specimen have genetically distinct populations, fundamental differences in genomic characteristics and variant detectability lead to suboptimal results when applying existing methods to non-tumor decomposition. First, detection of somatic mutations in normal tissues is fraught with the low variant allele frequency (VAF) (<1–5%) (10), which causes numerous false calls. A series of brain mosaicism studies reported false positives ranging from 9.9 to 32.9% of total variants (11,12). As clone-specific mutations are the key evidence for decomposition, the erroneous variants should be considered and properly handled. Second, the genomic similarity among clones is higher in normal tissues due to the lower mutation rate and limited observable clone-specific mutations, making a deterministic assignment of clones difficult and negatively affects the estimation of clone numbers. Lastly, the absence of copy number alterations (CNA), important evidence of tumor decomposition, limits the information for clone identification and makes the entire algorithm rely solely on SNVs in normal tissues. Additionally, the lack of CNA alleviates the model complexity in relating VAF to cellular prevalence and warrants more efficient decomposition. These differences emphasize the need for a specialized method for the genomic decomposition of non-tumor samples.

In this study, we present a new method CLEMENT (CLonal decomposition using Expectation-Maximization algorithm Established in Non-Tumor diploid samples), for accurate decomposition and reconstruction of subclones in non-tumor tissues. We employed the following three core strategies to resolve the aforementioned problems: (1) measuring and parameterizing false positivity in the input variants to reduce noise in clone identification, (2) using fuzzy clustering to enable more flexible discrimination of genetically similar clones, and (3) setting restrictions on clonal fractions in the determination of clonal compositions (i.e. total clone fraction = 1, see Materials and methods for details) due to the absence CNAs. We observed the improved accuracy of CLEMENT in three independent, high-quality datasets: *in silico* simulations, *in vitro* cell-line mixture (13), and human datasets derived from multiple normal tissues using laser capture microdissection (LCM) (14). We anticipate that CLEMENT will provide a deeper understanding of genomic and tissue-level heterogeneity, mosaicisms, and the functional relatedness of somatic mutations in normal tissues.

Materials and methods

Overview of the CLEMENT algorithm

CLEMENT consists of three major steps, as follows: (i) the initialization step that determines the initial number of clones using K-means clustering, (ii) the iteration step that searches for the optimal compositions of clones based on the Expectation-Maximization (EM) algorithm in a given number of clones and (iii) the finalization step to determine the optimal number of individual and ancestral clones (undifferentiated clones that harbor two or more individual clones) and their hierarchical clone structures (Figure 1). CLEMENT uses a list (or lists, if two or more samples are provided) of established somatic variants and their total read counts and alternate read counts as input, and outputs a list of individual and ancestral clones, and the membership of somatic variants with a visual representation. Detailed methods are formulated in the following sections.

Definitions of subclone and superclone

The term 'clone' refers to a set of cells that harbor unique characteristics in terms of mutation (5). In CLEMENT, we used variants as a genomic feature to define clones. Among the clones, we defined '*subclones*' or *individual* clones, as clones that are genetically mutually exclusive (Supplementary Figure S1). The total cellular prevalence of subclones is 1.0 by definition. With the lack of CNAs and homozygosity of somatic mutations in normal tissue, the sum of VAFs is 0.5. In contrast, *ancestral* clones, or '*superclones*', possess a clonal mutation that has been dispersed among their own subclones, where the proportions of clonal mutations are the sum of the proportions of subclones (15).

Basic mathematical definitions

Let $S = \{s_1, \ldots, s_m\}$ and $V = \{v_1, \ldots, v_n\}$ be the set of m samples and n somatic variants given to CLEMENT, respectively. User input

$$\boldsymbol{N}^{total} = \left\{ n_{i,j}^{total} : n_{i,j}^{total} = \ total\ read\ count\left(s_i, v_j\right),\ 1 \le i \le m,\ 1 \le j \le n \right\}$$

$$\boldsymbol{N}^{alt} = \left\{ n_{i,j}^{alt} : n_{i,j}^{alt} \ = \ alternate\ allele\ count\left(s_i, v_j\right),\ 1 \le i \le m,\ 1 \le j \le n \right\}$$

are multisets of the total read count (i.e. read depth) and alternate allele count of each sample and genomic position of variants. $n_{i,j}^{total}$ is doubled when s_i is male and v_j located at sex chromosome, to calibrate from the haploid to diploid data.

From them, we define $\boldsymbol{F} = \{ f_{i,j} : f_{i,j} = \mathrm{VAF}(s_i,\ v_j) = \frac{n_{i,j}^{alt}}{n_{i,j}^{total}}, 1 \le i \le m,\ 1 \le j \le n, 0 \le f_{i,j} \le 1 \}$ as a multiset of VAF values of the variant set V, allowing for duplication. During the algorithm, $k+1$ clusters composed of k true biologic clones and a cluster of false variant (FV) $C = \{c_1, \ldots, c_k, c_{FV}\}$ are assumed, where each cluster occupies a subset of variants.

$\boldsymbol{\lambda}^i$ and $\boldsymbol{\Lambda}$ are defined as a posterior probability matrix with elements denoted as $\boldsymbol{\lambda}^i(\mathrm{v_j}, c_y)$ and $\boldsymbol{\Lambda}(\mathrm{v_j}, c_y)$ $(1 \le j \le n,\ y = \{1, \ldots, k, FV\})$ with regard to single sample s_i $(1 \le i \le m)$ and whole sample respectively, satisfying

$$\sum_{y=1}^{k} \boldsymbol{\lambda}^i\left(\mathrm{v_j}, c_y\right) + \ \boldsymbol{\lambda}^i\left(v_j, c_{FV}\right) = 1$$

and

$$\boldsymbol{\Lambda}\left(v_j, c_y\right) = \frac{\prod_{i=1}^{m} \boldsymbol{\lambda}^i\left(\mathrm{v_j}, c_y\right)}{\sum_{z=\{1,\ldots,k,FV\}} \prod_{i=1}^{m} \boldsymbol{\lambda}^i\left(\mathrm{v_j}, c_z\right)}$$

Subsequently, we defined the membership function $\boldsymbol{\Theta} = V \times C \to \{\rho : 0 \le \rho \le 1\}$ by satisfying for each $v \in V$, $\sum_{y=1}^{k} \boldsymbol{\Theta}(v, c_y) + \boldsymbol{\Theta}(v, c_{FV}) = 1$, which is gathered from $\boldsymbol{\Lambda}$. For hard clustering, $\boldsymbol{\Theta}(v, c)$ is either 0 or 1; $\boldsymbol{\Theta}(v, c) = 1$ if $v \in V$ is a member of $c \in C$, and $\boldsymbol{\Theta}(v, c) = 0$ if $v \in V$ is not a member of $c \in C$ (see below). For fuzzy clustering, we set $\boldsymbol{\Theta} = \boldsymbol{\Lambda}$ to allow partial membership by assigning a posterior probability between 0 and 1 in E step (see below).

Lastly, let $\mu^{\mathrm{i}}(\mathrm{c_y})$ $(y = \{1, \ldots, k, FV\})$ be the centroid of a cluster c_y in sample i, and $\mathbf{M}(\mathrm{c_y})$ $(y = \{1, \ldots, k, FV\})$ be the centroid in m-dimensional vector, comprised of $(\mu^1(\mathrm{c_y}), \ldots, \mu^{\mathrm{m}}(\mathrm{c_y}))$, which is recalibrated in M step.

Step 1: Initialization

In the Initialization step, CLEMENT selects k initial centroids, k is iterated within range of [2, 10] (user adjustable), on the given data S , V , *and* F, to provide a rough estimate of the clone structure for the next EM-iteration step.

First, CLEMENT performs the K-means clustering using the scikit-learn (version 1.0.2) package (16) to partition somatic variants ${}^{\forall}v_j \in V (1 \ \le \ j \ \le \ n)$, an m-dimensional VAF vector $(f_{1,j}, \ \ldots, f_{m,j})$, into T (user adjustable; default = 10) clusters. Then, CLEMENT randomly selects k $(k \le T)$ initial TP clusters out of T given centroids, which implies a provisional clone set $C^0 = \{c_1^0, \ldots, c_k^0, c_{FV}^0\}$. This random selection step is repeated for 10 times (user adjustable) for each k to ensure extensive exploration.

Step 2: EM Iteration

This EM-iteration step takes the initial clone set $C^0 = \{c_1^0, \ldots, c_k^0, c_{FV}^0\}$ from the aforementioned step, together with S, V, and F, and outputs a clone set $C^{Max} = \{c_1^{Max}, \ldots, c_k^{Max}, c_{FV}^{Max}\}$ of the Maximum a Posteriori (MAP) probability. In this step, an alternative EM process is iterated 20 times (user adjustable) until it satisfies the stopping conditions (see the end of the section). The whole EM iteration step is conducted in two ways, as follows: first, in a hard clustering, and then in followed a fuzzy clustering (Figure 1, top middle). The final output of the EM iteration can be either from the initial hard clustering or the fuzzy clustering, which is determined later in Step 3 (Figure 1, bottom right).

E step

In the E step, CLEMENT assigns each data point to the k clones, given the centroids of clusters as a latent variable. CLEMENT also considers the probability that a given variant is a sequencing artifact, or falsely called, aiming to exclude its assignment to true clones. To achieve this, we created an additional cluster with a fixed centroid (c_{FV}) at the origin $(0, \ldots, 0)$, consisting of false variants. This cluster is distinguishable from other clones $(c_1, \ldots, \ c_k)$ because it does not represent a true biologic clone.

For $\forall i,\ j$, and $y = \{1, \ldots, k, FV\}$, posterior probability of data point belonging to clone j in sample i are calculated based on Bayesian theorem.

$$P\left(c_y|v_j\right) = \lambda^i\left(v_j, c_y\right) = \frac{P\left(v_j|c_y\right)P\left(c_y\right)}{\sum_z P\left(v_j|c_z\right)P(c_z)} \tag{1}$$

where $P(v_j|c_y)$, $P(c_y)$ denotes likelihood and prior probability, respectively.

Meanwhile, a variant v_j in sample i can be categorized as true positive (TP), false positive (FP), true negative (TN) or false negative (FN).

If $\mathrm{n}_{\mathrm{i,j}}^{\mathrm{alt}} \neq 0$, v_j is regarded either TP or FP. For subclone c_y ($y = \{1, \ldots, k, FV\}$) that satisfies $\boldsymbol{\mu}^i(c_y) \neq 0$ in sample i (TP), the likelihood of v_j in subclone c_y ($= P(v_j|c_y)$) follows beta-binomial distribution, represented as $\mathbf{L}_{\mathbf{BetaBin}}(n_{i,j}^{alt}|n_{i,j}^{total}, \alpha_{\mathrm{i,y}}, \beta_{\mathrm{i,y}}) = P(X = n_{i,j}^{alt}),\ X \sim \mathbf{BetaBin}(n_{i,j}^{total}, \alpha_{\mathrm{i,y}}, \beta_{\mathrm{i,y}})$, where parameters are set as $\alpha_{\mathrm{i,y}} = \mu^{\mathrm{i}}(\mathrm{c_y}) \cdot \mathrm{n}_{\mathrm{i,j}}^{\mathrm{total}}$ and $\beta_{\mathrm{i,y}} = (1 - \mu^{\mathrm{i}}(\mathrm{c_y})) \cdot \mathrm{n}_{\mathrm{i,j}}^{\mathrm{total}}$ to maximize the likelihood if v_j is located in the centroid. Because setting $\hat{\alpha},\ \hat{\beta}$ where $0 = \frac{\partial \mathbf{L}_{\mathbf{BetaBin}}}{\partial \alpha}|_{\hat{\alpha},\hat{\beta}}$ and $0 = \frac{\partial \mathbf{L}_{\mathbf{BetaBin}}}{\partial \beta}|_{\hat{\alpha},\hat{\beta}}$ requires a another computational load represented as Newton-Raphson method, we simply approximated to mean of the beta-binomial model coincides the mean of the observed data, $\mu^{\mathrm{i}}(\mathrm{c_y})$ (17). Users can set the multiplication constant $\boldsymbol{c}$ to $\alpha_{\mathrm{i,y}}$ and $\beta_{\mathrm{i,y}}$ through the user input if input data is significantly condensed or dispersed. Meanwhile, for subclone c_y ($y = \{1, \ldots, k, FV\}$) that satisfies $\boldsymbol{\mu}^i(c_y) = 0$ in sample i (FP), the likelihood follows binomial distribution, $\mathbf{B}(n_{i,j}^{total},\ p_{SE})$, where p_{SE} stands for sequencing error probability of v_j in sample i, inferred from base quality (BQ) score provided by the user input. If the users do not provide any input, it is set to 0.01 by default (18). So, likelihood of v_j in cluster c_y ($= P(v_j|c_y)$) in sample i is represented as $\mathbf{L}_{\mathbf{B}}(n_{i,j}^{alt}|n_{i,j}^{total}, p_{SE}) = P(X = n_{i,j}^{alt}),\ X \sim \mathbf{B}(n_{i,j}^{total}, p_{SE})$. Eq. 1 is rewritten in eq. 2–1 and eq. 2–2.

For c_y that satisfies $\boldsymbol{\mu}^i(c_y) \neq 0$ (TP),

$$\boldsymbol{\lambda}^i\left(v_j, c_y\right) = \frac{\mathbf{L}_{\mathbf{BetaBin}}\left(\mathrm{n}_{\mathrm{i,j}}^{\mathrm{alt}}|n_{ij}^{total}, \alpha_{\mathrm{i,y}},\ \beta_{\mathrm{i,y}}\right) \cdot \mathrm{P}\left(c_y\right)}{\sum_{\boldsymbol{\mu}^i(c_z)\neq 0} \mathbf{L}_{\mathbf{BetaBin}}\left(\mathrm{n}_{\mathrm{i,j}}^{\mathrm{alt}}|n_{i,j}^{total}, \alpha_{\mathrm{i,z}},\ \beta_{\mathrm{i,z}}\right) \cdot \mathrm{P}\left(c_z\right) + \sum_{\boldsymbol{\mu}^i(c_{z'})=0} \mathbf{L}_{\mathbf{B}}\left(n_{i,j}^{alt}|n_{i,j}^{total}, p_{SE}\right) \cdot \mathrm{P}\left(c_{z'}\right)} \tag{2-1}$$

and for c_y that satisfies $\boldsymbol{\mu}^i(c_y) = 0$ (FP),

$$\boldsymbol{\lambda}^i\left(v_j, c_y\right) = \frac{\mathbf{L}_{\mathbf{B}}\left(n_{i,j}^{alt}|n_{i,j}^{total}, p_{SE}\right) \cdot \mathrm{P}\left(c_y\right)}{\sum_{\boldsymbol{\mu}^i(c_z)\neq 0} \mathbf{L}_{\mathbf{BetaBin}}\left(\mathrm{n}_{\mathrm{i,j}}^{\mathrm{alt}}|n_{i,j}^{total}, \alpha_{\mathrm{i,z}},\ \beta_{\mathrm{i,z}}\right) \cdot \mathrm{P}\left(c_z\right) + \sum_{\boldsymbol{\mu}^i(c_{z'})=0} \mathbf{L}_{\mathbf{B}}\left(n_{i,j}^{alt}|n_{i,j}^{total}, p_{SE}\right) \cdot \mathrm{P}\left(c_{z'}\right)} \tag{2-2}$$

The prior probability $\mathrm{P}(c_{z'})$ for each $c_{z'}$ satisfying $\boldsymbol{\mu}^i(c_{z'}) = 0$ is set 0.01 by default (19), but the user can adjust this parameter by tuning the option. $\mathrm{P}(c_z)$ that satisfies $\boldsymbol{\mu}^i(c_z) \neq 0$ is set to $\frac{1-\sum \mathrm{P}(c_{z'})}{\mathrm{n(z)}}$ to satisfy the sum of prior to be 1.

If $\mathrm{n}_{\mathrm{i,j}}^{\mathrm{alt}} = 0$, v_j is either FN or TN for sample i. In case of FN, likelihood of v_j in subclone c_y ($= P(v_j|c_y)$) in sample i is $\mathbf{L}_{\mathbf{BetaBin}}(0|n_{i,j}^{total}, \alpha_{\mathrm{i,y}},\ \beta_{\mathrm{i,y}})$ of beta-binomial distribution as forementioned. Likewise, likelihood of v_j being TN is derived from binomial distribution mentioned in FP, calculated as $\mathbf{L}_{\mathbf{B}}(0|n_{i,j}^{total}, p_{SE})$.

For c_y that satisfies $\boldsymbol{\mu}^i(c_y) \neq 0$ (FN),

$$\boldsymbol{\lambda}^i\left(v_j, c_y\right) = \frac{\mathbf{L}_{\mathbf{BetaBin}}\left(0|n_{i,j}^{total}, \alpha_{\mathrm{i,y}},\ \beta_{\mathrm{i,y}}\right) \cdot \mathrm{P}\left(c_y\right)}{\sum_{\boldsymbol{\mu}^i(c_z)\neq 0} \mathbf{L}_{\mathbf{BetaBin}}\left(0|n_{i,j}^{total}, \alpha_{\mathrm{i,z}},\ \beta_{\mathrm{i,z}}\right) \cdot \mathrm{P}\left(c_z\right) + \sum_{\boldsymbol{\mu}^i(c_{z'})=0} \mathbf{L}_{\mathbf{B}}\left(0|n_{i,j}^{total}, p_{SE}\right) \cdot \mathrm{P}\left(c_{z'}\right)} \tag{2-3}$$

and for c_y that satisfies $\boldsymbol{\mu}^i(c_y) = 0$ (TN),

$$\boldsymbol{\lambda}^i\left(v_j, c_y\right) = \frac{\mathbf{L}_{\mathbf{B}}\left(0|n_{i,j}^{total}, p_{SE}\right) \cdot \mathrm{P}\left(c_y\right)}{\sum_{\boldsymbol{\mu}^i(c_z)\neq 0} \mathbf{L}_{\mathbf{BetaBin}}\left(0|n_{i,j}^{total}, \alpha_{\mathrm{i,z}},\ \beta_{\mathrm{i,z}}\right) \cdot \mathrm{P}\left(c_z\right) + \sum_{\boldsymbol{\mu}^i(c_{z'})=0} \mathbf{L}_{\mathbf{B}}\left(0|n_{i,j}^{total}, p_{SE}\right) \cdot \mathrm{P}\left(c_{z'}\right)} \tag{2-4}$$

Regarding the prior probability, sum of $\mathrm{P}(c_{z'})$ where $\boldsymbol{\mu}^i(c_{z'}) = 0$ (TN) is basically set 0.99 with an identical value for each clone (user adjustable), and $\mathrm{P}(c_z)$ where $\boldsymbol{\mu}^i(c_z) \neq 0$ (FN) is set to $\frac{1-\sum_{\mu^i(c_{z'})=0} P(c_{z'})}{n(z)}$ to reflect the real-world knowledge (19).

Finally, v_j is assigned to clone $c_{\hat{y}}$.

$$\hat{y} = \underset{y}{argmax} \prod_{i=1}^{m} \boldsymbol{\lambda}^i\left(v_j, c_y\right) = \underset{y}{argmax}\ \Lambda\left(v_j, c_y\right) \tag{3}$$

In hard clustering,

$$\Theta\left(v_j, c_y\right) = \begin{bmatrix} 1\ , & y = \hat{y} \\ 0\ , & y \neq \hat{y} \end{bmatrix}\ for\ \forall j, y \tag{4}$$

In fuzzy clustering,

$$\Theta\left(v_j, c_y\right) = \Lambda\left(v_j, c_y\right)\ \ for\ \forall j, y \tag{5}$$

M step

In the M step, CLEMENT updates the centroid of each clone using the membership calculated in the E step as below:

$$\mu^{\mathrm{i}}\left(\mathrm{c_y}\right) = \frac{\sum_{\mathrm{j}=1}^{\mathrm{n}} f_{i,j}\boldsymbol{\Theta}\left(\mathrm{v_j},\mathrm{c_y}\right)}{\sum_{\mathrm{j}=1}^{\mathrm{n}} \boldsymbol{\Theta}\left(\mathrm{v_j},\mathrm{c_y}\right)} \quad for\ \forall i, y \tag{6}$$

Distinguishing the individual clone and ancestral clone

After updating the centroids, CLEMENT classifies all the clones into either (i) an *individual clone* C^{ind} ($C^{ind} \subseteq C$, $n(C^{ind}) = k'$), which is an independent clone that is separated from the other clones, or (ii) an *ancestral clone* C^{anc} ($C^{anc} \subseteq C$, $n\ (C^{anc} \cap C^{ind}) = 0,\ n(C^{anc}) = k - \ k'$, Supplementary Figure S2), which is an undifferentiated clone that is composed of two or more individual clones. As mentioned earlier (see Basic mathematical definitions), CLEMENT chooses a set of clones from all possible combinations, the sum of whose centroids is 0.5. These set of clones are regarded as individual clones, or subclones (*subclone rule*, Eq. (7-1)).

$$\sum_{y\ =\ 1}^{k'} \mu^{i}\left(\mathrm{c_y}\right) = 0.5\ \ for\ \mathrm{c_y} \in C^{ind},\ \forall i \tag{7-1}$$

CLEMENT employs multisample *t*-test with null hypothesis $\sum_{y\ =\ 1}^{k'} \mu^{i}(\mathrm{c_y}) - 0.5 = 0$, with significance level = 0.01 and degree of freedom = $\sum_{y=1}^{k'} n(c_y) - k'$. If variance of each cluster is not identical, an alternative degree of freedom is used (20). *t* values are derived from difference of group averages by dividing standard error of difference. If the statistics do not reject the null hypothesis, it ensures the sum of centroids is regarded as 0.5. When more than one combination of clusters satisfies the condition, CLEMENT selects the set with the highest *P* value on the multisample t-test, which supports the null hypothesis most strongly.

Ancestral clone $\mathrm{c_z}$ is cluster of clonal mutations, whose proportion is sum of its subclones $\mathrm{c_u}$. CLEMENT establishes superclone–subclone structure (phylogeny inference) by (eq. 7–2), in other words, sum rule (15).

$$M(\mathrm{c_z}) = \sum_{c_u \in C'^{ind}} M(c_u) \text{for } c_z \in C^{anc},\ C'^{ind} \subset C^{ind} \tag{7-2}$$

CLEMENT also employs another multisample *t*-test, assuming a null hypothesis that the mean value of the superclone is equal to the sum of mean values of the subclones, as previously described.

If any of (Eq. (7-1)) or (Eq. (7-2)) is unsatisfied, the iteration stops and restarts with Step 1 using another provisional clone set $C^0 = \{c_1^0, \ldots, c_k^0, c_{FV}^0\}$. Otherwise, the E–M process continues until convergence.

Determination of convergence

The EM iteration stops if the following conditions are satisfied:

- A. The number of EM iterations is >10 times (user adjustable), AND:
- B-1. The gap between the Maximum a Posteriori probability of two successive steps is <1%, OR
- B-2. The gaps between all centroids of two successive steps are <0.01, OR
- B-3. The membership matrix $\boldsymbol{\Theta}$ retains same for two successive steps.

Step 3: Finalization

After two rounds of the EM iteration, CLEMENT determines whether the clustering results from the initial hard clustering or the secondary fuzzy clustering will be used as output (Figure 1, bottom right). Among the hard clustering results, CLEMENT uses Gap* Statistics (21) to choose optimal *k*, where the intra-cluster variation is minimized and inter-cluster variation is maximized. If Jaccard similarity between $\boldsymbol{\Lambda}(v, c_{y_1})$ and $\boldsymbol{\Lambda}(v, c_{y_2})$ exceeds 0.2 for $\exists y_1, y_2\ (1 \leq y_1, y_2 \leq k)$, CLEMENT selects the result from fuzzy clustering; otherwise, the hard clustering results are retained. Finally, variants in c_{FV} are designated as false variants (FV), in the other words, sequencing artifacts. The remaining variants are treated as true variants (TV).

Test set preparation

Test set preparation

We used three independent test sets to measure the performance of CLEMENT and other tools, as follows: (i) a simulated dataset (SimData), (ii) an *in vitro* cell line mixture dataset (CellData) and (iii) real human multi-clone microdissected tissues (BioData).

Simulated dataset (SimData)

The SimData was made up of hypothetical clones and computationally introduced clone-specific somatic mutations thereof, based on two assumptions for non-tumor conditions: (a) absence of CNAs, and (b) absence of homozygote variants. Test sets were constructed with a random choice of (i) the number of clones ($k : 2 \leq k \leq 7$) and (ii) the number of samples ($i :\ 1 \leq i \leq 3$) under the discrete uniform distribution. The number of clones and samples is limited arbitrarily for the economic computational burden of parallel benchmarking. The total number of variants and the mean read-depth for the benchmark were chosen as 500 and 125, to reflect the 125× DNA sequencing that is commonly used in real datasets.

Two types of datasets, SimData-decoy and SimData-lump, were generated computationally. First, for the SimData-decoy dataset, the proportional distribution of k clones in sample i was randomly drawn from the Dirichlet distribution with shape parameter $\alpha = (\alpha_1, \ldots, \alpha_k)$, where α_y $(1 \leq y \leq k)$ was randomly selected from a binomial distribution $\mathbf{B}(10y,\ 0.5)$. For each clone of a proportion ρ, somatic heterozygote mutations were generated at random genomic loci, with the total read-depth ($\boldsymbol{N}^{total}$) randomly chosen from normal distribution $\mathcal{N}(\mathbf{125}, \mathbf{8})$ and alternate allele count ($\boldsymbol{N}^{alt}$) following binomial distribution $\mathbf{B}(\boldsymbol{N}^{total} \times \rho/2, 0.5)$, making sure that VAF follows distribution with mean VAF as $\rho/2$. Then, false somatic variants were generated, the VAF of which were based on the reverse sigmoid function (eq. 8) to mimic the nature of real-world dataset.

$$P\left(X = f_{i.j} | v_{i.j} \in \boldsymbol{c}_{FV}^{A}\right) = \left(1 - \frac{1}{1 + e^{\left(100 f_{i.j} - 5\right)}}\right) \times Constant \tag{8}$$

where $\boldsymbol{c}_{FV}^{A}$ refers FV cluster of answer sets, and ***Constant*** is a constant to make the sum of probability density function 1. Mean and median VAF of the model were 0.029 and 0.027, respectively (see Supplementary Figure S3). Five different datasets, including a different number of false variants (0%, 2.5%, 5%, 7.5% or 10% of the total 500 variants), were evaluated accordingly.

Similarly, the SimData-lump dataset was generated by changing the shape parameter of the Dirichilet distribution $\alpha' = (\alpha_1', \ldots, \alpha_k')$, where $\alpha_y'(1 \leq y \leq k)$ is randomly selected from a binomial distribution $\mathbf{B}(10k, 0.5)$; this makes the clones more agglomerated. Likewise, five datasets with false variants added (0%, 2.5%, 5%, 7.5% or 10% of the total 500 variants) were prepared to evaluate the influence of false variants.

Here, $500 - \mathrm{n}(\boldsymbol{c}_{FV}^{A})$ somatic mutations were distributed to k clones, so the number of mutations per clone followed a multinomial distribution ($\mathbf{n} = 500 - n(\boldsymbol{c}_{FV}^{A})$, $\mathbf{p_i} = \frac{1}{k}$ for $i = (1,\ 2,\ \ldots,\ k)$). The VAF and number of mutations of each clone are depicted in Supplementary Figure S4. Base quality (BQ) of each variant was set to 20 (99% confidence).

Then, we expanded our benchmark by selecting the combination of total variants and mean read-depth among $Total\ variants = \{100,\ 500,\ 1000\}$ and $Read\_depth = \{30,\ 125,\ 250\}$, to verify the performance of CLEMENT according to the number of total variants and read-depths. In each dataset, we repeated the random sampling 30 times and evaluated the mean value.

Cell line mixture dataset (CellData)

CellData was constructed based on our previous study that provided 39 physical mixtures of six completely genotyped human cell lines (MRC5, RPE, CCD-18co, HBEC30-KT, THLE-2 and FHC) in various compositions (three or four cell lines out of the six) and cellular proportions (0.5–56%) (11). CellData consisted of fully diploid genomes by excluding sex chromosomes to ensure copy number neutrality.

In a mixture, cell line-specific variants form an individual clone. Also, variants that are shared between the cell lines form hypothetical ancestral clones. We downloaded the 1,100x whole-exome sequencing (WES) data of the 39 mixtures (M1-1 to M1-9, M2-1 to M2-12 and M3-1 to M3-18) from the Sequence Read Archive (SRA) repository database (SRP334852). Among the 39 mixtures, we used eight to construct the test sets. The exclusion criteria were as follows: (i) the presence of two uneven clones (clone size difference $>$ 5 times) at inseparable frequencies (VAF difference ≤ 0.03) (excluded M1-1,3,5,7 and M2-10,12), (ii) the presence of extremely small clones (clone size $\leq$ 30 variants; mean clone size = 1629) (excluded M1-9, M2-1,3,5,7,9,11) and (iii) redundancy of the clone composition (all 18 M3 mixtures). As a result, eight test sets with 3–4 individual and 0–1 ancestral clone were prepared. A multi-sample dataset was generated by combining 2 or 3 samples, irrespective of mixture category (M1 or M2). Accordingly, 28 ($=_8C_2$) two-sample and 56 ($=_8C_3$) three-sample test sets were prepared. The characteristics of CellData, including clonal proportion, variants counts, and false variants compositions are listed in Supplementary Table S1.

For each WES dataset of the selected mixtures, somatic single nucleotide variants (SNVs) were selected using GATK MuTect2 (version 4.2.3.0) and filtered by FilterMutectCalls (ver. 4.2.3.0) with default parameters (22). SNVs that did not match any of the cell line genotypes were marked as false variants. Similar to SimData, we chose the total number of variants and the mean read-depth for the benchmark as 500 and 125, by downsampling the initially downloaded 1,100x datasets using picard (v2.26.4, http://broadinstitute.github.io/picard).

We extended our test datasets by applying 0 (0%), 13 (2.5%), 25 (5%), 38 (7.5%) and 50 (10%) false variants. Then, we generated the datasets without the ancestral clone and with one ancestral clone added. The number of mutations and mean read-depths for simulations were extended to a combination of $Total\ variants = \{100,\ 500,\ 1000\}$ and $Read\_depth = \{30,\ 125,\ 250\}$ to assess the performance of CLEMENT in various conditions. We conducted repetitive randomized trials (30 times) and comparisons for each dataset. In each trial, we selected variants through random sampling while maintaining the overall proportions of each clone. The test datasets for CellData are available on https://github.com/Yonsei-TGIL/CLEMENT.

Human normal tissue dataset (BioData)

BioData was prepared using a recent study that conducted whole-genome sequencing of 561 laser capture microdissected patches from 29 microscopic histological structures from three individuals (14). The number of clones from the 29 tissues was estimated by the genomic VAF and histological assessment in the original study and used in the test. We downloaded 524 732 somatic SNVs and VAFs from the 29 tissues (5 mono-clonal, 4 bi-clonal and 20 poly-clonal; according to the author's estimation) from the Supplementary Information of the study (14) and used them as inputs for testing. We discarded samples that did not pass the following conditions: (i) average read depth $\geq$ 20 and (ii) total number of mutations $\geq$ 100. Finally, we obtained 224 samples from 24 tissues to perform single-sample decomposition.

In the evaluation, mono- and bi-clonal samples from the original datasets were marked as k (the number of clusters) = 1 *and* 2, respectively. For poly-clonal samples, a prediction of $k \geq 3$ was considered correct.

Performance measurement

Scoring index

The test performance was measured in two terms, as follows: (i) the accuracy of the clone number estimation and (ii) the accuracy of variant membership. For (i), the deviation of the estimated clone number from the true number was scored in the root mean square error (RMSE) of all the trials (30 times). For (ii), the Adjusted Rand Index (ARI) (23) and membership score (S_M) were used to performance measurement.

Let $C_A = \{c_1^A, \ \ldots, c_{\kappa_A}^A\}$ and $C_P = \{c_1^P, \ \ldots, c_{\kappa_P}^P\}$ be a set of clusters in answer and predicted output, and their indices as $I_A = \{1, 2, \ldots, \kappa_A\}$ and $I_P = \{1, 2, \ldots, \kappa_P\}$. Then, we define $u_{i,j}$ $(1 \le i \le \kappa_A, \ 1 \le j \le \kappa_P)$ as the number of common variants in $c_i^A \in C_A$ and $c_j^P \in C_P$. Also, let a_i and p_j be the number of variants in c_i^A and c_j^P, respectively. ARI is defined as below:

$$ARI = \frac{\sum_{ij}\binom{u_{i,j}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{p_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{p_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{p_j}{2}\right] / \binom{n}{2}} \tag{9}$$

where n is the number of variants as forementioned.

However, for the test sets with false variants, ARI is not applicable due to its inability to discriminate the set of false variants from true clones. Symmetricity of ARI provides an advantage in not needing precise labeling of clusters, but there is also a limitation in its applicability when exact discrimination of FV is required. Therefore, we defined an additional score S_M that measures the maximum matches of variant membership from all possible injection functions (between the answer and the predicted clusters) in answer set C_A and the predicted set C_P.

$$S_M = \max_{R \in \mathcal{R}} \left\{ \sum_{(i,j) \in R} u_{i,j} \right\} \times \frac{100}{n} \tag{10}$$

where $\mathcal{R}$ is a set of relations $R \subset I_A \times I_P$ satisfying that, if $\kappa_A \le \kappa_P$, then R is an injective function ($R : I_A \to I_P$ with $(x, y) \in R$ and $(z, y) \in R$ implies $x = z$), and if $\kappa_P \le \kappa_A$, then R^T is an injective function, where $R^T = \{(y, x) : (x, y) \in R\}$. To normalize it, we recalibrated by the number of variants, n. An example of the determination of S_M is described in Supplementary Figure S5.

Testing of cancer decomposition tools

Three cancer decomposition tools were prepared for the test. PyClone-VI (version 0.1.1) was downloaded from the GitHub repository (https://github.com/Roth-Lab/pyclone-vi) and installed using Conda. SciClone (version 1.1.0) was downloaded and installed on R (version 4.2.0) following the installation instructions in (https://github.com/genome/sciclone). QuantumClone (version 1.0.0.9) was downloaded and installed on R (ver. 4.2.0) using CRAN (https://CRAN.R-project.org/package=QuantumClone). For SciClone and PyClone-VI, parameters for the copy number ('major_cn' and 'minor_cn') and tumor content ('tumour_content') were set to 1 for optimization. For QuantumClone, the 'Genotype' parameter was set to 'AB'. Default values were used for all other parameters.

Analysis in the real-world datasets

Mutational burden and signature analysis in bi- or poly-clonal samples

Out of a total 224 samples, CLEMENT identified 136 samples as monoclonal. For the remaining 88 bi- or poly-clonal samples, we measured the mutational burden in each clone. In cases where CLEMENT chose fuzzy clustering which does not provide binary membership, we assigned the membership of v_j as y where $\Theta(v_j, c_y)$ is maximized. We used standard deviation as a metric for dissimilarity of mutational burden. To determine the threshold where clones are not genetically identical, we employed a bootstrap approach with 100 times of iteration. The null hypothesis assumed that the mutational burden for each clone is the same. If the standard deviation of the proportions of each clone within the sample is beyond the 95% confidence interval (CI) of the distribution obtained through bootstrapping, we considered that sample to be highly dissimilar.

We analyzed the spectrum of mutational signatures in BioData following clonal decomposition using CLEMENT. For signature extraction and matrix formation, we utilized SigProfilerMatrixGenerator (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator). Signature extraction, assignment to the COSMIC database, and visualization were performed using SigProfilerAssignment (version 0.0.29, https://github.com/AlexandrovLab/SigProfilerAssignment) (24). We specifically focused on SBS1, 2, 4, 5/40, 7a, 7b, 13, 16, 18, 32, 35, 88 and 91, as outlined in Moore *et al.*'s paper (14), to maintain the pattern of signatures from the original paper. Notably, we combined SBS5 and SBS40 into SBS5/40, following the previous work. We explored the percentage of SBS1 and SBS5/40 in each clone, as discrepancies among tissues were noted in the previous study.

Clonality analysis in adrenal glands

We obtained three layered tissues (Zona Glomerulosa (ZG), Zona Fasciculata (ZF) and Zona Reticularis (ZR)) from one donor (PD28690) at five different regions (L1–L5), resulting in a total of 15 samples. One-sample and two-sample clonal decompositions were performed using CLEMENT, PyClone-VI, SciClone and QuantumClone.

We performed unsupervised hierarchical clustering and visualization between the samples using scipy (ver. 1.10. https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html) based on Jaccard similarity. By comparing the mutation sets and decomposed clones using CLEMENT, we analyzed the clonality in two adjacent tissues (ZG – ZF). We used BioRender (http://app.biorender.com/) to create the illustration.

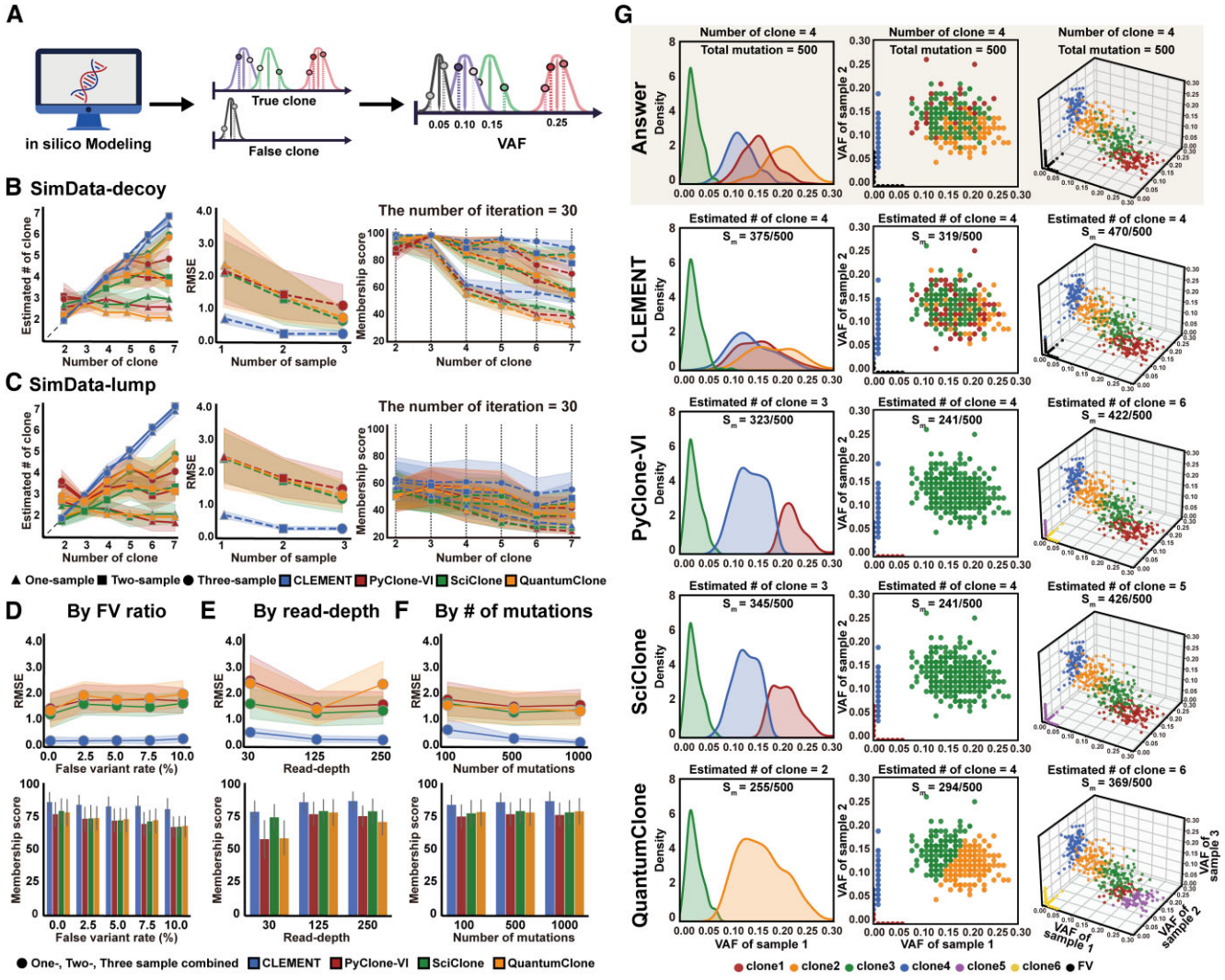


Figure 2. Test on simulated data. (**A**) Establishment of an *in silico* mixed non-tumor diploid model using Dirichlet's random sampling. (B, C) Performance comparison for estimated number of clones (left), RMSE (middle) for number of clones, and membership score (S_M, right) with 500 mutations and mean depth 125 in (**B**) SimData-decoy without false variants, (**C**) SimData-lump without false variants. Extended benchmark of SimData-decoy without false variants by (**D**) false variant ratio, (**E**) read-depth and (**F**) the number of mutations. Each point represents the mean value of 30 times of iteration, and 95% confidence interval (CI) is depicted as a shadow. (**G**) Examples of one-sample (left), two-sample (middle), and three-sample (right) decomposition that show the superiority of CLEMENT. False variants are depicted as black dots. FV: false variants, VAF: variant allele frequency, RMSE: root mean square error.

Results

Clonal decomposition in simulated data

We tested CLEMENT on simulated data sets (SimData-decoy and SimData-lump) over various conditions (Figure 2a). The primary SimData is a collection of 18 simulated samples consisting six different numbers of clones (k = 2, 3, 4, 5, 6, and 7), and three different numbers of samples (m = 1, 2, and 3) containing 500 variants (n = 500) whose mean read-depths are 125, without the embedment of false mutations (see Materials and metthods). The mean VAFs of each cluster exhibited divergence in SimData-decoy and agglomerated in SimData-lump, although the difference was narrowed as the number of clones increased (Supplementary Figure S4). Three cancer decomposition tools (PyClone-VI, SciClone and QuantumClone) were also tested on the datasets and compared with CLEMENT.

Firstly, in the SimData-decoy dataset, we observed a nearly perfect correlation between the estimated number of clones and answers in CLEMENT (Figure 2b, left). Specifically, we observed a remarkable improvement in terms of RMSE (0.06–0.55 versus 0.96–2.09, 0.58–2.27 and 0.48–2.03 in CLEMENT versus PyClone-VI, SciClone, and QuantumClone, respectively; ranges within one-, two-, and three-sample datasets, Figure 2b, middle) and the membership agreement metrics (27.0–58.3% and 28.8–40.0% increase in membership score S_M and ARI, respectively; ranges within one, two, and three-sample datasets, Figure 2b, right). The improved accuracy was intensified with the increase in sample and clone numbers. In particular, CLEMENT maintained its performance, while all the cancer decomposition tools showed severe underestimation of clone numbers in $k \geq 4$ (Figure 2B, right).

We also observed the superiority of CLEMENT compared to the other tools in the SimData-lump dataset, in which the VAF values of mutations were more condensed, making clustering more challenging. CLEMENT demonstrated superior performance in estimating the number of clones, achieving nearly perfect matches for each clone (Figure 2C, left). In terms of RMSE, CLEMENT exhibited better results ranging 0.05–0.49, compared to 1.33–2.37, 1.12–2.36 and 1.01–2.31 in PyClone-VI, SciClone and QuantumClone, respectively (Figure 2C, middle). Additionally, membership agreement to clones was higher in SimData-lump (40.2–60.2% higher in S_M and 43.1–53.3% higher in ARI, Figure 2C, right, Supplementary Table S2). Similar to that in SimData-decoy, the superiority of CLEMENT over other tools was remarkable as the number of clones increased (11.3–33.9% higher in $k = 2$ and 38.5–60.2% higher in $k = 7$ regarding S_M). Additionally, the accuracy of CLEMENT was even more significant as more samples were included. Especially in $m = 3$, the performances were sustained even when more clones were introduced where accurate decomposition was harsher.

We extended our evaluations of the SimData-decoy and SimData-lump datasets under various conditions, including changing the ratio of inserted false variants, read-depths, and the number of mutations (Figure 2D–F, Supplementary Figure S6a–c). CLEMENT consistently outperformed the other tools across all conditions, showing particularly reliable outputs in estimating the number of clones. The superiority of CLEMENT became more pronounced with an increase in the higher read-depths (11.8–35.4% higher in S_M and 1.10–2.14 in ΔRMSE), greater number of mutations (11.6–13.6% higher in S_M and 1.17–1.44 in ΔRMSE), and more embedded false variants (11.8–12.0% higher in S_M and 1.24–1.70 in ΔRMSE). CLEMENT successfully isolated false variant cluster in 65.2% of whole simulations, whereas the other tools erroneously allocated the false variants into clusters or considered them as true biologic clone. Identification of false variant clusters became more feasible when more samples were provided (see Supplementary Table S3).

A more detailed inspection of the predicted cluster compositions provides a better understanding of the underlying characteristics of CLEMENT and cancer decomposition algorithms (Figure 2G). In the presence of multiple clones with overlapping VAF ranges, cancer decomposition tools are more inclined to predict a larger, merged cluster instead of smaller individual clones, resulting in an underestimation of clone numbers. The strong feature of the CLEMENT is its superior performance in more agglomerated datasets that reflect real-world biology, via fuzzy clustering.

Test on *in vitro* cell line data

We tested CLEMENT on another test set (CellData) constructed from mixtures of pre-genotyped cell lines in various proportions (Figure 3A, see Materials and methods for details). Unlike the simulation datasets, CellData contains a series of false negatives when performing two-sample or more decomposition because low VAF mutations are easily missed in single-sample calling (Supplementary Figure S7). Additionally, several samples in CellData consists of different type of mixtures (e.g. M1-2 of M1 and M2-4 of M2, see Supplementary Table S1), resulting in a series of sample-restricted clones that reflect real-world biology. Therefore, direct sequencing of physical clones and conventional variant calling offers a most realistic scenario for non-tumor decomposition. Like the SimData dataset test, CLEMENT showed superior performance.

In the absence of false variant and ancestral clones, CLEMENT exhibited better performance in both RMSE and S_M. CLEMENT demonstrated the highest accuracy in estimating clone numbers, particularly in three-sample decomposition (RMSE = 0.58–0.78) (Figure 3b, left). Conversely, cancer decomposition tools either under- or over-estimated clone numbers, resulting in significantly higher RMSE values (0.85–1.79, 1.18–1.43 and 1.11–3.34 in PyClone-VI, SciClone and QuantumClone, respectively; ranges within one-, two- and three-sample datasets). Additionally, when one superclone was added, CLEMENT demonstrated the best performance, except for one-sample decomposition (Figure 3B, middle). We speculate that this is because most of the clones in CellData have extremely low prevalence, making their superclone challenging to discern in one-sample decomposition. Generally, the accuracy of CLEMENT improved as more samples were provided, whereas the other tools exhibited inconsistency. CLEMENT also showed the superiority in terms of S_M regardless of the presence of a superclone, although the differences among the tools were not as significantly pronounced as those in SimData (∼7.4%, ∼9.1% and ∼21.2% higher for one-, two- and three-sample decomposition, Figure 3b, right). The same phenomenon was observed in terms of ARI (Supplementary Table S4)

In the presence of false variant data, ranging from 2.5% to 10% of the entire datasets, CLEMENT outperformed other cancer decomposition tools significantly. When 10% false variants were added, CLEMENT showed the best RMSE and S_M for all sample numbers (Figure 3C). Its superiority over the other cancer decomposition tools was more pronounced than that in the absence of false variants (∼7.9%, ∼14.2% and ∼26.7% higher for one-, two- and three-sample decomposition). All the tools tended to be less accurate when more false variants were included, whereas CLEMENT maintained relative overall performance (Figure 3D). CLEMENT successfully identified false variant data in two- and three- sample test sets (∼29.5 and ∼49.1% detection rate, respectively, Supplementary Table S5). In one-sample data, discrimination of false variants was challenging because the mean VAF of false variants (0.029, see Supplementary Figure S3) made identification alongside other low VAF clones (0.01–0.04) extremely unrealistic.

In the extended test sets that include various combinations of read-depths and the number of mutations, CLEMENT generally outperformed the other tools (Figure 3E, F, Supplementary Figure S8a, b). In 30× downsample and 100 mutations datasets, the superiority of CLEMENT was not as remarkable as that in other conditions, because the densely populated clones in low VAF of CellData are easily influenced by the harsh condition, making decomposition much more challenging. Especially, among the 30× data, variants of low VAF were filtered in the variant calling step, distorting the input information and resulting in right-shifting of the clusters. However, CLEMENT still maintained competitiveness compared to the others and outperformed them in all the other tests.

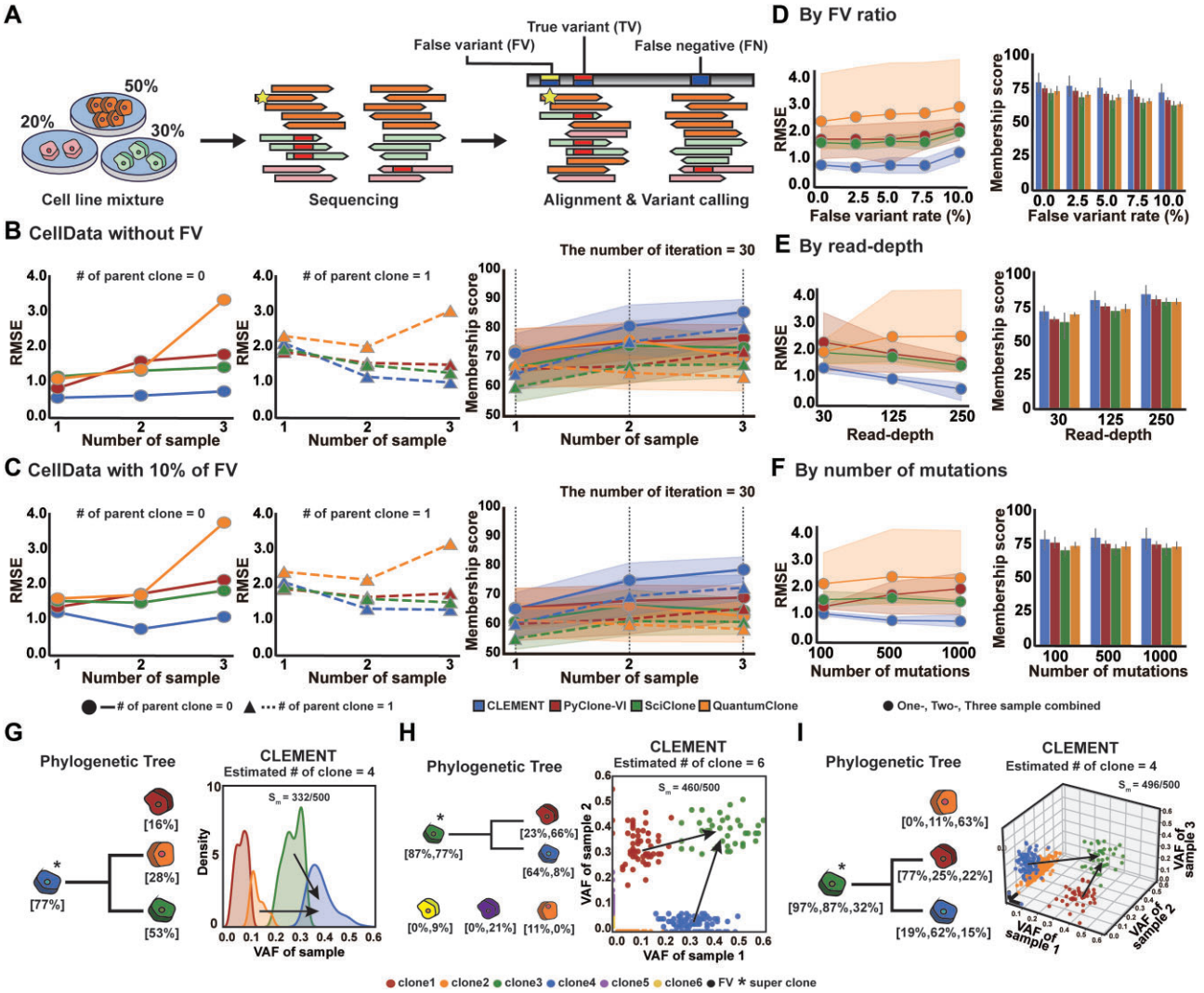


Figure 3. Test on pre-genotyped mixed cell line data. (**A**) Illustrative diagram of analysis utilizing mixed cell line data. (**B**, **C**) Comparison of RMSE (left and middle) for the number of clones and mean membership score (S_M, right) with 95% confidence interval (CI, shadow) in one-sample ($n = 8$), two-sample ($n = 28$), and three-sample ($n = 56$) cell line datasets with (B) no false variant or (C) 10% false variants. Extended benchmark by (**D**) FV ratio, (**E**) read-depth and (**F**) the number of mutations are also described. (**G–I**) Phylogeny reconstruction of superclone-added cell line data in one-sample input (G, M1-6), two-sample input (H, M1-8 + M2-4) and three-sample input (I, M1-4 + M1-6 + M1-8). False variants are depicted as black dots. FV: false variants, VAF: variant allele frequency, RMSE: root mean square error.

When a superclone was added to CellData, CLEMENT appropriately reconstructed superclone-subclone structures. The examples of reconstructed clone structures confirmed the accurate discrimination of ancestral clones from individual ones in various conditions, which was consistent with the answer datasets (Figure 3G: one-, 3H: two-, 3I: three-samples).

Test on human multi-clonal normal tissues

Finally, we applied CLEMENT to the sequencing of 24 microdissected human normal tissues that showed mono- to poly-clonal microstructures in Moore *et al.* (12) (BioData) (Figure 4A). The number of clones identified in the original study based on genomic profiling and histological assessment served as the gold standard. Clonality analysis was performed based on a sample level that evaluated the predicted number of clones. Only CLEMENT provided the fuzzy clustering that reflected the agglomerated nature of human data (Figure 4B). In total, CLEMENT estimated the exact clone numbers in 204 samples (91.1%) out of 224, with an RMSE of 0.30, outperforming the other tools (# exact match = 44 (19.6%), 57 (25.4%), and 66 (29.4%); RMSE = 1.12 (0.84–1.51), 1.45 (0.91–1.86) and 0.85 (0.0–1.03) in PyClone-VI, SciClone, and QuantumClone, respectively) (Figure 4C, D). We found a high correlation (Pearson's $r = 0.94$) between the predicted numbers and the gold standard in CLEMENT, whereas almost no correlation (r = -0.45—0.20) was observed in the three cancer decomposition tools (Figure 4E). Notably, QuantumClone converged to bi-clonality in nearly all samples, which clearly demonstrated the weakness of the Bayesian Information Criterion (BIC) method when determining the number of clones (Figure 4F). Conversely, SciClone tended to produce a large number of clones in nearly all cases, indicating that the RMSE of SciClone in poly-clonal ($k \geq 3$) tissues were erroneously overestimated.

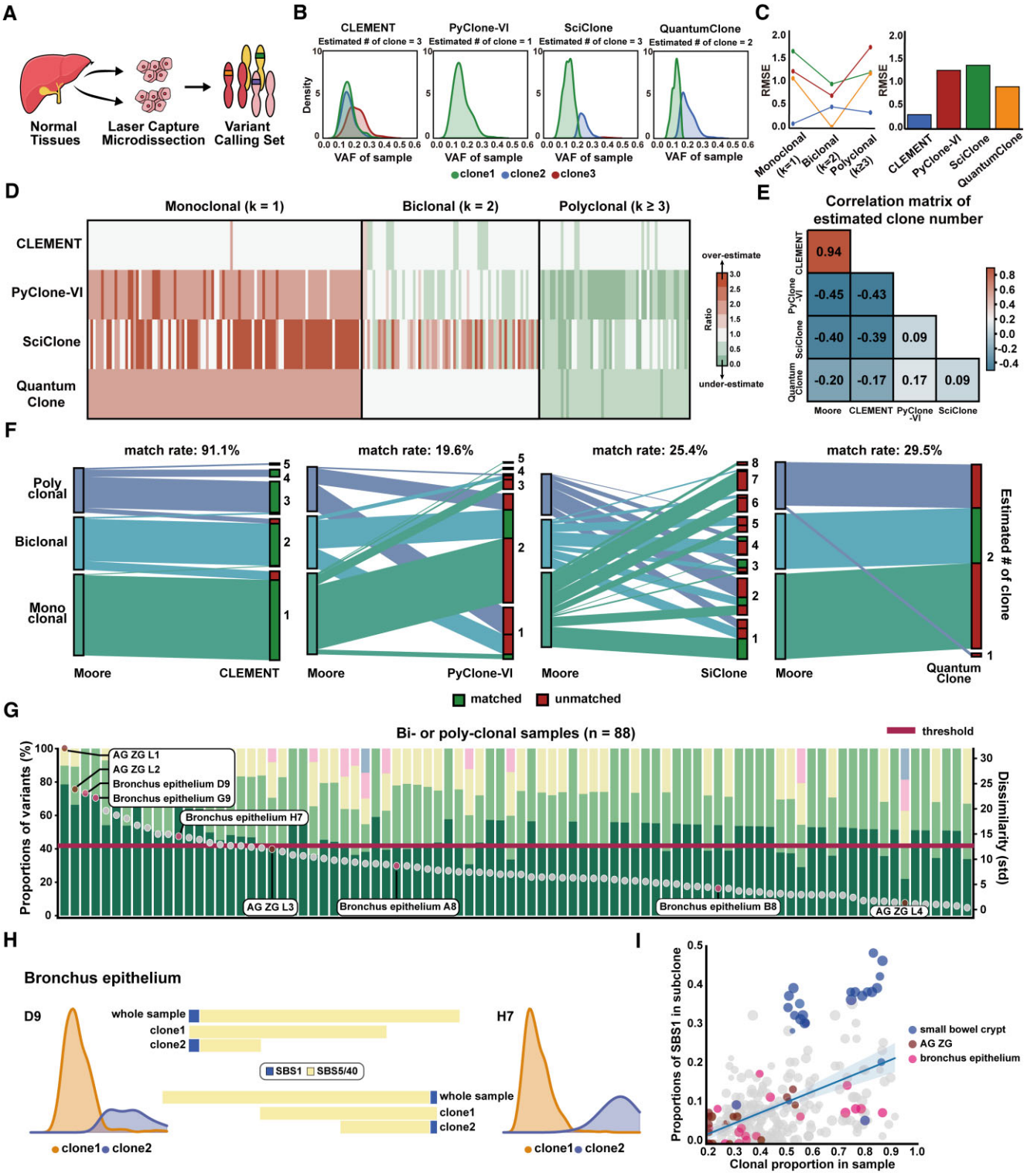


Figure 4. Test on human microdissected tissue data. (**A**) Schematic figure for obtaining clonal data in normal tissue, which shows high genomic similarity. (**B**) Confirmation of successful separation of CLEMENT through one-sample decomposition example (PD43850-pancreas duct-D7). (**C**) RMSE in each clonality (mono-clonal, bi-clonal, and poly-clonal) for four decomposition methods (left). Barplot (right) describes the RMSE for all samples. (**D**) Heatmap depicting the relationship between Moore *et al.*'s conjecture and other tools. Red indicates an overestimation of the number of clones compared with Moore's, and green refers to underestimation. (**E**) Correlation matrix and coefficients of estimated number of clones between the tools, including Moore *et al.*'s conjecture. (**F**) Alluvial plots for the number of estimated clones between Moore's conjecture and decomposition tools. (**G**) Stacked bar plot depicting the proportions of subclones for 88 bi- or poly-clonal samples. Circle indicates the dissimilarity of number of mutations calculated by standard deviation. (**H**) Composition of mutational signatures in two bronchus epithelium tissues. (**I**) Correlation was noted between clonal proportion in sample and proportions of SBS1 in each subclone. Linear regression was plotted as blue line ($r = 0.49$). VAF: variant allele frequency.

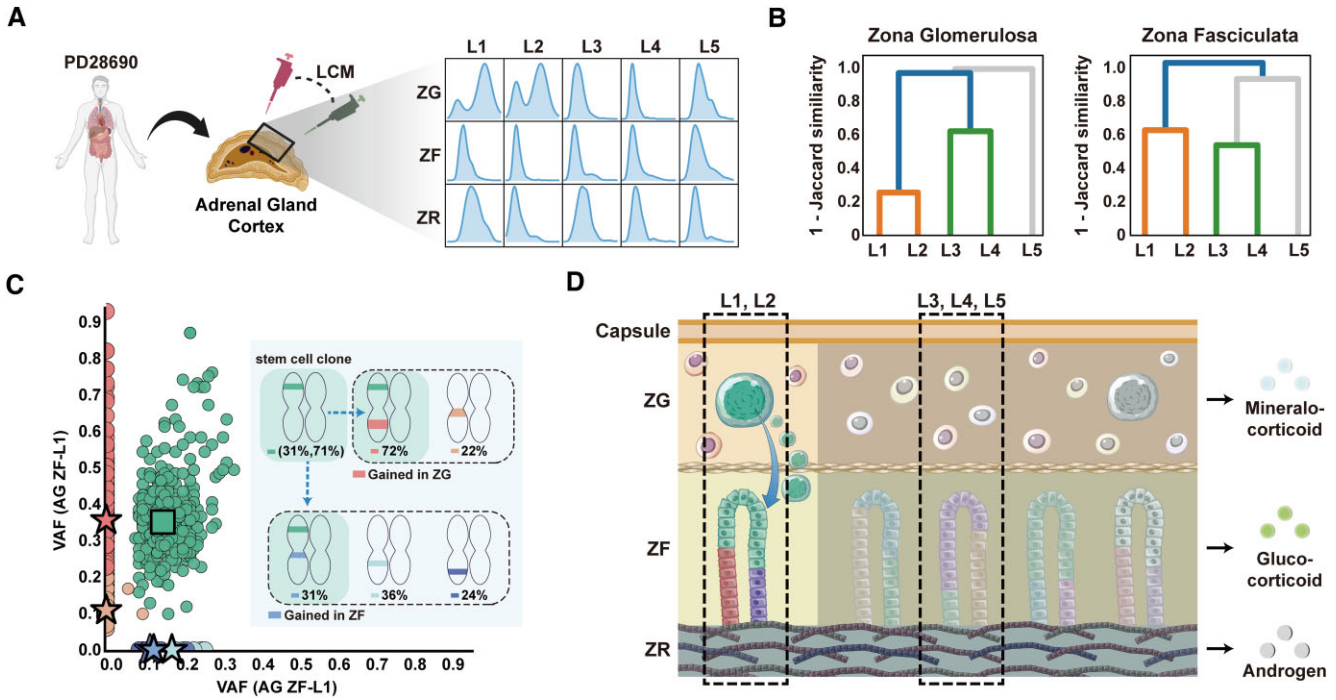


Figure 5. Clonality analysis of adrenal gland cortex. (**A**) Further adrenal gland analysis in 15 tissues. The blue figures show the distribution of variant allele frequency (VAF). (**B**) Unsupervised hierarchical clustering based on Jaccard similarity in ZG and ZF. (**C**) Two-sample decomposition revealing the presence of superclone, indicating the adult stem cell in adrenal gland cortex. (**D**) Proposed concepts of clonal migration in the adrenal gland. Large blue cell implies an adult stem cell in ZG. ZG: Zona Glomerulosa, ZF: Zona Fasciculata, ZR: Zona Reticularis Adapted from 'Organs, multiple systems', 'Pipette (symbol)' and 'Adrenal Gland Structure and Hormones Production', by BioRender.com (2024). Retrieved from https://app.biorender.com/biorender-templates.

For bi- or poly-clonal samples ($n = 88$, Supplementary Figure S9), we investigated the genomic profile for each clone and found dissimilar features within the clones. First, we examined the evenness of the number of mutations among the clones by calculating the standard deviation. We observed discrepancies in clonal mutational burden distribution among the samples (Figure 4G). The dissimilarity was well observed in samples with multiple peaks. For example, in adrenal gland zona glomerulosa (ZG) L1, the green clone included most of the mutations, whereas two condensed clones (light green and beige) had fewer mutations. Clones in normal samples have long been thought to be homogenous, but we witnessed that some populations (18%) exhibited distinguished discrepancies beyond the threshold, necessitating precise clonal decomposition. Additionally, we found that the dissimilarity varied even within the same tissue, such as ZG and bronchus epithelium. Clonal inference (proportions and number of clones) in normal tissue bulk datasets has been made based on the assumption that all clones are homogeneous, but we suggest clonal inference after exact clonal decomposition.

Next, we investigated the pattern of mutational signatures at the clone-level. The average number of mutations per clone was 415, which is sufficient to decompose the mutational signatures according to the genomic context. SBS5/40 was the dominant mutational pattern in most tissues, except for the small bowel crypt where SBS1 was the major mutational pattern. The discrepancy among the tissues was mentioned in a previous publication (14), and we reaffirmed the same phenomenon at the clone-level. Interestingly, we found that bronchus epithelium showed a different pattern of mutational signatures between the clones. For example, bronchus epithelium H7 and D9, which showed a clear bimodal peak and were decomposed by CLEMENT as bi-clonal, were divided as a major clone with SBS1 and a minor clone consisting only SBS5/40 (Figure 4h). We expanded our inspection for all clones of all samples. We noted the positive relationship between the clonal prevalence and the proportion of SBS1 (correlation coefficient $r = 0.49$, Figure 4I). As clones with higher VAFs indicate mutations occurred earlier (25), we hypothesized that the pattern of mutational signatures differs by the timing of mutations acquisition. A single-cell genome sequencing of the forebrain revealed the C > T mutations are enriched in early mutations (26), supporting our finding in that SBS1 signature mostly represents C > T mutations. Although direct evidence of high allelic fraction clones possessing high stemness is limited, we offer a glimpse of the relationship between mutational contexts and developmental dynamics at the clone level.

Analysis on microscopic tissue of adrenal gland cortex

Further analysis of 15 adrenal gland tissues revealed the heterogeneous nature of clonal compositions (Figure 5a). Cortex of adrenal glands consists of three layers—Zona Glomerulosa (ZG), Zona Fasciculata (ZF) and Zona Reticularis (ZR)—from the outer to the inner layer, each producing different steroid hormones. Since 1883, the presence of adult stem cells in the periphery of the adrenal gland cortex and formation of the ZG–ZF axis have been hypothesized ('centripetal migration model') (27), which was validated through BrdU staining (28).

In BioData, we observed that the outermost layer, ZG of L1 and L2 had definite dual peaks, implying more than bi-clonality. Conversely, ZF of L1 and L2 had a single peak, suggesting highly agglomerated poly-clonal tissues. Jaccard similarity within ZG and ZF revealed that L1 and L2 were genetically equivalent, as well as L3 and L4 (Figure 5B, Supplementary Figure S10). We noted substantial shared mutations between ZG-L1 and ZF-L1 (Jaccard similarity = 0.38), whereas there was no shared mutation between ZG-L3 and ZF-L3 (Jaccard similarity = 0.0, Supplementary Figure S11). Interestingly, VAFs of shared mutations were similar to the dominant peak in ZG (0.38) and one of the homogenous clones in ZF (0.17), indicating the clonality between the ZG-L1 and ZF-L1. In a two-sample decomposition (Figure 5c), CLEMENT revealed the presence of a superclone that has the major clone of ZG-L1 and one of the homogenous poly-clonal backgrounds in ZF-L1 as its subclones. From these observations, we confirmed the presence of clonality in ZG-L1 and ZF-L1, indicating superclone-subclone structures. In normal tissue, a superclone in localized tissue is equivalent to the adult stem cell. Therefore, we concluded that the adult stem cell clone resides in the adrenal gland, migrates to ZG and ZF and proliferates. This clone is clearly distinguished from the poly-clonal background in ZF-L1, which seems to be multifurcated from the developmental period. Conversely, in L3, there was no clonality between ZG and ZF, indicating the absence of an adult stem cell clone. The absence of clonality in ZG-L3 and ZF-L3 supports independent clear zonation in the developmental stage (27). These findings are in accordance with previous findings that postmeiotic stem cells are found in *localized* areas of ZG (29) (Figure 5D). Throughout the entire process, we took advantage of CLEMENT, which provided (i) clonal reconstruction and (ii) homogenous poly-clonality in most samples, unlike the other tools. Clonality analysis in ZG-L5 or ZR were unavailable due to the limited number of mutations and extremely low cellular fractions (Supplementary Figure S12).

Discussion

In this study, we pioneered a novel method of genomic decomposition in non-tumor samples. The EM-based algorithm with additional strategies for the proper handling of non-tumor sequencing data led to a substantial improvement in estimating the clone composition and structures and was validated in three independent test sets. In-depth analysis under various conditions, including the presence of false variants and different inter-clone similarities confirmed the effectiveness of the strategies, as well as the limitation of current cancer decomposition tools in normal clone analysis.

Recent efforts in identifying clonal heterogeneity (9,30,31) and developmental lineage (6,32) in normal tissues identified essential characteristics of non-tumor subclones in multiple aspects, especially against traditional cancer-derived samples. While the post-embryonic somatic and mosaic mutations are the major sources for both subclones, the differences in mutation rates (1–100 per Mb in cancer versus <1 per MB in normal tissues), mutation types (frequent CNAs and chromosomal instability in cancer), VAFs (very low in normal tissues), and the colonization path (fast clonal expansion in cancer versus slow to no clonal expansion in normal tissues) confer the intrinsic differences between their subclones, which set the basis for our study. In addition to the genomic properties that are already formulated in CLEMENT (false variants, clone similarity, and absence of CNAs), more sophisticated features can also be employed for further improvement, such as the distribution of VAFs within a clone and the mutation signatures.

As in the tumor decomposition, there is a growing interest in the use of non-tumor clonal analysis to understand disease pathogenesis. For example, the existence of the stem cell niche and the regeneration of the cells *in situ* has been widely studied to assume pathogenic clones in normal tissues, as shown in the colonic crypt, esophagus epithelium, and the subventricular and subgranular zone of the brain (33,34). In addition, recent studies adopt a new perspective on genomic regeneration and clonal evolution in investigating neurodegenerative diseases, including Alzheimer's disease (35), hippocampal sclerosis (36), and schizophrenia (37). As we discovered the localized stem cell in the adrenal gland cortex, understanding the clonal structure using appropriate clonal decomposition may be greatly helpful in expanding the knowledge of normal or non-tumor tissue. We expect that CLEMENT will provide a better assessment of the compositions and microscopic structures, as well as the number, dispersion, and phylogeny of these clones.

Despite the substantial achievements of CLEMENT, there are a few remaining technical issues to be resolved. First, the robust mathematical background for the modified E-M algorithm applied in CLEMENT has not been well discussed. Unlike the orthogonal E-M algorithm, CLEMENT interrupts the iteration if the *subclone rule* or *sum rule* is unsatisfied. Additionally, CLEMENT employs multiple probabilistic models in the E step considering FP and FN, to reflect real-world biology. However, we observed the convergence of E–M iteration and the achievement of optimal results by simulating more than thousands of datasets (Supplementary Figure S13). Second, proper discrimination of false variants from true low-allele frequency mutation is still a challenging problem. We noted that significant portions of false variants were not properly identified, especially in a single-sample case. We expect that the incorporation of recent technical advances, such as duplex sequencing (38) or a bioinformatics approach (39), could address the problem. Third, performance is limited when low read-depth is provided because of a lack of read information to be classified as multiple clones, and uncalled or filtered low VAF mutations distorting the clonal structure, resulting in the right-shifting of low prevalent clones. We hope that further advance in biotechnology and computational algorithms will improve the forementioned problems.

Data availability

CLEMENT with all the code and data used in this manuscript is available at the FigShare repository (https://figshare.com/s/12f35ff0785ad5c27921) and GitHub repository (https://github.com/Yonsei-TGIL/CLEMENT).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

Figure 5A and D were created with BioRender.com.

Funding

MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development Institute (KHIDI), National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [RS-2023-00278314]; Korea Health Technology R&D Project through the KHIDI and Korea Dementia Research Center (KDRC), funded by the Ministry of Health & Welfare, Republic of Korea [RS-2021-KH113577]. Funding for open access charge: NRF of Korea grant funded by the Korea government (MSIT) [RS-2023-00278314t].

Conflict of interest statement

None declared.

References

1. Hinohara,K. and Polyak,K. (2019) Intratumoral heterogeneity: more than just mutations. *Trends Cell Biol.*, **29**, 569–579.
2. Roth,A., Khattra,J., Yap,D., Wan,A., Laks,E., Biele,J., Ha,G., Aparicio,S., Bouchard-Cote,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. Gillis,S. and Roth,A. (2020) PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinf.*, **21**, 571.
5. Deveau,P., Colmet Daage,L., Oldridge,D., Bernard,V., Bellini,A., Chicard,M., Clement,N., Lapouble,E., Combaret,V., Boland,A., *et al.* (2018) QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. *Bioinformatics*, **34**, 1808–1816.
6. Park,S., Mali,N.M., Kim,R., Choi,J.W., Lee,J., Lim,J., Park,J.M., Park,J.W., Kim,D., Kim,T., *et al.* (2021) Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature*, **597**, 393–397.
7. Hsieh,A., Morton,S.U., Willcox,J.A.L., Gorham,J.M., Tai,A.C., Qi,H., DePalma,S., McKean,D., Griffin,E., Manheimer,K.B., *et al.* (2020) EM-mosaic detects mosaic point mutations that contribute to congenital heart disease. *Genome Med.*, **12**, 42.
8. Coorens,T.H.H., Moore,L., Robinson,P.S., Sanghvi,R., Christopher,J., Hewinson,J., Przybilla,M.J., Lawson,A.R.J., Spencer Chapman,M., Cagan,A., *et al.* (2021) Extensive phylogenies of human development inferred from somatic mutations. *Nature*, **597**, 387–392.
9. Brunner,S.F., Roberts,N.D., Wylie,L.A., Moore,L., Aitken,S.J., Davies,S.E., Sanders,M.A., Ellis,P., Alder,C., Hooks,Y., *et al.* (2019) Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*, **574**, 538–542.
10. Roberts,N.D., Kortschak,R.D., Parker,W.T., Schreiber,A.W., Branford,S., Scott,H.S., Glonek,G. and Adelson,D.L. (2013) A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*, **29**, 2223–2230.
11. Kim,J.H., Hwang,S., Son,H., Kim,D., Kim,I.B., Kim,M.H., Sim,N.S., Kim,D.S., Ha,Y.J., Lee,J., *et al.* (2022) Analysis of low-level somatic mosaicism reveals stage and tissue-specific mutational features in human development. *PLoS Genet.*, **18**, e1010404.
12. Kim,M.H., Kim,I.B., Lee,J., Cha,D.H., Park,S.M., Kim,J.H., Kim,R., Park,J.S., An,Y., Kim,K., *et al.* (2021) Low-level brain somatic mutations are implicated in schizophrenia. *Biol. Psychiatry*, **90**, 35–46.
13. Ha,Y.J., Oh,M.J., Kim,J., Kim,J., Kang,S., Minna,J.D., Kim,H.S. and Kim,S. (2022) Establishment of reference standards for multifaceted mosaic variant analysis. *Sci. Data*, **9**, 35.
14. Moore,L., Cagan,A., Coorens,T.H.H., Neville,M.D.C., Sanghvi,R., Sanders,M.A., Oliver,T.R.W., Leongamornlert,D., Ellis,P., Noorani,A., *et al.* (2021) The mutational landscape of human somatic and germline cells. *Nature*, **597**, 381–386.
15. Dang,H.X., White,B.S., Foltz,S.M., Miller,C.A., Luo,J., Fields,R.C. and Maher,C.A. (2017) ClonEvol: clonal ordering and visualization in cancer sequencing. *Ann. Oncol.*, **28**, 3076–3082.
16. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn Res.*, **12**, 2825–2830.
17. Griffiths,D.A. (1973) Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, **29**, 637–648.
18. Stoler,N. and Nekrutenko,A. (2021) Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.*, **3**, lqab019.
19. Beck,T.F., Mullikin,J.C., Program,N.C.S. and Biesecker,L.G. (2016) Systematic evaluation of Sanger validation of next-generation sequencing variants. *Clin. Chem.*, **62**, 647–654.
20. Welch,B.L. (1947) The generalisation of student's problems when several different population variances are involved. *Biometrika*, **34**, 28–35.
21. Tibshirani,R., Walther,G. and Hastie,T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Stat. Soc. B*, **63**, 411–423.
22. Van der Auwera,G.A., Carneiro,M.O., Hartl,C., Poplin,R., Del Angel,G., Levy-Moonshine,A., Jordan,T., Shakir,K., Roazen,D., Thibault,J., *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, **43**, 11.10.1–11.10.33.
23. Chacon,J.E. (2020) A close-up comparison of the misclassification error distance and the adjusted Rand index for external clustering evaluation. *Br. J. Math. Stat. Psychol.*, **74**, 203–231.

24. Diaz-Gay,M., Vangara,R., Barnes,M., Wang,X., Islam,S.M.A., Vermes,I., Narasimman,N.B., Yang,T., Jiang,Z., Moody,S., *et al.* (2023) Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. bioRxiv doi: https://doi.org/10.1101/2023.07.10.548264, 11 July 2023, preprint: not peer reviewed.
25. Kim,S.N., Viswanadham,V.V., Doan,R.N., Dou,Y., Bizzotto,S., Khoshkhoo,S., Huang,A.Y., Yeh,R., Chhouk,B., Truong,A., *et al.* (2023) Cell lineage analysis with somatic mutations reveals late divergence of neuronal cell types and cortical areas in human cerebral cortex. bioRxiv doi: https://doi.org/10.1101/2023.11.06.565899, 07 November 2023, preprint: not peer reviewed.
26. Bae,T., Tomasini,L., Mariani,J., Zhou,B., Roychowdhury,T., Franjic,D., Pletikos,M., Pattni,R., Chen,B.J., Venturini,E., *et al.* (2018) Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science*, **359**, 550–555.
27. Gottschau,M.S.u.e.E.d.N.b.S. (1883) Archiv fur Anatomie und Physiologie.
28. Freedman,B.D., Kempna,P.B., Carlone,D.L., Shah,M., Guagliardo,N.A., Barrett,P.Q., Gomez-Sanchez,C.E., Majzoub,J.A. and Breault,D.T. (2013) Adrenocortical zonation results from lineage conversion of differentiated zona glomerulosa cells. *Dev. Cell*, **26**, 666–673.
29. Walczak,E.M. and Hammer,G.D. (2015) Regulation of the adrenocortical stem cell niche: implications for disease. *Nat. Rev. Endocrinol.*, **11**, 14–28.
30. Moore,L., Leongamornlert,D., Coorens,T.H.H., Sanders,M.A., Ellis,P., Dentro,S.C., Dawson,K.J., Butler,T., Rahbari,R., Mitchell,T.J., *et al.* (2020) The mutational landscape of normal human endometrial epithelium. *Nature*, **580**, 640–646.
31. Martincorena,I., Fowler,J.C., Wabik,A., Lawson,A.R.J., Abascal,F., Hall,M.W.J., Cagan,A., Murai,K., Mahbubani,K., Stratton,M.R., *et al.* (2018) Somatic mutant clones colonize the human esophagus with age. *Science*, **362**, 911–917.
32. Hasaart,K.A.L., Manders,F., van der Hoorn,M.L., Verheul,M., Poplonski,T., Kuijk,E., de Sousa Lopes,S.M.C. and van Boxtel,R. (2020) Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. *Sci. Rep.*, **10**, 12991.
33. Li,R., Di,L., Li,J., Fan,W., Liu,Y., Guo,W., Liu,W., Liu,L., Li,Q., Chen,L., *et al.* (2021) A body map of somatic mutagenesis in morphologically normal human tissues. *Nature*, **597**, 398–403.
34. Bizzotto,S., Dou,Y., Ganz,J., Doan,R.N., Kwon,M., Bohrson,C.L., Kim,S.N., Bae,T., Abyzov,A., Network,N.B.S.M., *et al.* (2021) Landmarks of human embryonic development inscribed in somatic mutations. *Science*, **371**, 1249–1253.
35. Miller,M.B., Huang,A.Y., Kim,J., Zhou,Z., Kirkham,S.L., Maury,E.A., Ziegenfuss,J.S., Reed,H.C., Neil,J.E., Rento,L., *et al.* (2022) Somatic genomic changes in single Alzheimer's disease neurons. *Nature*, **604**, 714–722.
36. Striano,P. and Nobile,C. (2018) Whole-exome sequencing to disentangle the complex genetics of hippocampal sclerosis-temporal lobe epilepsy. *Neurol. Genet.*, **4**, e241.
37. Skene,N.G., Bryois,J., Bakken,T.E., Breen,G., Crowley,J.J., Gaspar,H.A., Giusti-Rodriguez,P., Hodge,R.D., Miller,J.A., Munoz-Manchado,A.B., *et al.* (2018) Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.*, **50**, 825–833.
38. Abascal,F., Harvey,L.M.R., Mitchell,E., Lawson,A.R.J., Lensing,S.V., Ellis,P., Russell,A.J.C., Alcantara,R.E., Baez-Ortega,A., Wang,Y., *et al.* (2021) Somatic mutation landscapes at single-molecule resolution. *Nature*, **593**, 405–410.
39. Wardell,C.P., Ashby,C. and Bauer,M.A. (2021) FiNGS: high quality somatic mutations using filters for next generation sequencing. *BMC Bioinf.*, **22**, 77.

Received: June 7, 2023. **Revised:** May 27, 2024. **Editorial Decision:** May 31, 2024. **Accepted:** June 12, 2024