Bioinformatics, 35, 2019, i520–i529
doi: 10.1093/bioinformatics/btz327
ISMB/ECCB 2019

PRISM: methylation pattern-based, reference-free inference of subclonal makeup

Dohoon Lee[1], Sangseon Lee[2] and Sun Kim[1,2,3,*]

[1]Interdisciplinary Program in Bioinformatics, [2]Department of Computer Science and Engineering and [3]Bioinformatics Institute, Seoul National University, Seoul 08826, Korea

*To whom correspondence should be addressed.

Abstract

Motivation: Characterizing cancer subclones is crucial for the ultimate conquest of cancer. Thus, a number of bioinformatic tools have been developed to infer heterogeneous tumor populations based on genomic signatures such as mutations and copy number variations. Despite accumulating evidence for the significance of global DNA methylation reprogramming in certain cancer types including myeloid malignancies, none of the bioinformatic tools are designed to exploit subclonally reprogrammed methylation patterns to reveal constituent populations of a tumor. In accordance with the notion of global methylation reprogramming, our preliminary observations on acute myeloid leukemia (AML) samples implied the existence of subclonally occurring focal methylation aberrance throughout the genome.

Results: We present PRISM, a tool for inferring the composition of epigenetically distinct subclones of a tumor solely from methylation patterns obtained by reduced representation bisulfite sequencing. PRISM adopts DNA methyltransferase 1-like hidden Markov model-based *in silico* proofreading for the correction of erroneous methylation patterns. With error-corrected methylation patterns, PRISM focuses on a short individual genomic region harboring dichotomous patterns that can be split into fully methylated and unmethylated patterns. Frequencies of such two patterns form a sufficient statistic for subclonal abundance. A set of statistics collected from each genomic region is modeled with a beta-binomial mixture. Fitting the mixture with expectation-maximization algorithm finally provides inferred composition of subclones. Applying PRISM for two AML samples, we demonstrate that PRISM could infer the evolutionary history of malignant samples from an epigenetic point of view.

Availability and implementation: PRISM is freely available on GitHub (https://github.com/dohlee/prism).

Contact: sunkim.bioinfo@snu.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The concept of clonal evolution in cancer (Nowell, 1976) has revolutionized our understanding of cancer biology throughout various subjects including progression of cancer (Merlo *et al.*, 2006), recurrence (Yates *et al.*, 2017), metastasis (Gundem *et al.*, 2015; Turajlic and Swanton, 2016; Yates *et al.*, 2017) and treatment response (Almendro *et al.*, 2014; Kreso *et al.*, 2013). Evolution of the cancer subclones often results in a tumor composed of several genetically or epigenetically distinct subclones. Therefore, the subclonal diversity arising from clonal evolution has long been acknowledged as one of the prominent causes of intratumor heterogeneity (ITH). ITH intuitively reflects the adaptive capacity of a tumor to survive changing conditions. Thus, its utility as a biomarker predicting the aggressiveness of a tumor has been widely studied.

Next-generation sequencing (NGS) has offered us an excellent opportunity to interrogate ITH at an unprecedented resolution. Accordingly, there have been several important approaches to define ITH measures in various omics level, including genomic (Mroz and Rocco, 2013), transcriptomic (Park *et al.*, 2016) and methylomic (Landau *et al.*, 2014) level. These ITH measures were proven to have remarkable clinical potentials. However, the best precision and utilization of ITH can be achieved through the direct characterization of cancer subclone itself. Formally, this problem of characterizing constituent subclones of a bulk tumor only from its molecular

signature is often referred to as 'subclonal inference'. A critical application of subclonal inference is the reconstruction of the evolutionary history of a tumor (Gerlinger *et al.*, 2012), which harbors great potential for the precision medicine when exploited (Hiley *et al.*, 2014). Most existing methods for subclonal inference such as ABSOLUTE (Carter *et al.*, 2012), THetA2 (Oesper *et al.*, 2014), SciClone (Miller *et al.*, 2014) or PyClone (Roth *et al.*, 2014) utilize prevalence of a subclonal genomic variation including somatic copy number alteration (CNA) or single nucleotide variant (SNV) as a proxy of subclonal abundance.

Meanwhile, clonal evolution accompanied by the genome-wide, dynamic reprogramming of DNA methylation has received increasing attention in recent years (Brocks *et al.*, 2014; Ferrando and López-Otín, 2017; Li *et al.*, 2016; Mazor *et al.*, 2016), inspiring researchers to develop methods to inspect the methylomic evolution in cancer (Barrett *et al.*, 2017; Li *et al.*, 2014). However, applications of those methods are limited to detecting methylation patterns that are likely to have undergone subclonal expansion, and none of them directly aims to uncover the subclonal population structure of a tumor sample, thereby enabling the inference of methylome-based evolutionary tree of the subclones. Therefore, we present a bioinformatic tool named PRISM as a solution for the *epigenetic* subclonal inference problem. Motivated by the strategy taken by mutation-based subclonal inference algorithms, which takes advantage of the dichotomousness of variant and reference alleles, PRISM tackles the problem by focusing on the particular genomic region harboring dichotomous groups of methylation patterns, namely fully methylated and unmethylated patterns.

2 Problem formulation and approach

2.1 Terminologies used in this research

In this section, we define terminologies used throughout the research. To help understand the terminologies, a schematic illustration is in Figure 1A. We define an *epigenetic subclone* as a clonal population of cells harboring distinct regional methylation patterns that exclusively belong to that subclone. Also, the exclusive subclonal methylation pattern will be referred to as *fingerprint methylation pattern* or *fingerprint pattern* for convenience. We defined an *epilocus* as a short genomic region (typically ~100 bp) at which a group of reduced representation bisulfite sequencing (RRBS) reads was mapped, and subsequently a *fingerprint epilocus* is defined as an epilocus harboring fingerprint pattern. From each fingerprint epilocus, a fraction of sequencing reads supporting fingerprint pattern can be calculated and this fraction will be denoted as *fraction of fingerprint pattern* (FF) (Fig. 1C). This value will serve as an estimator of subclonal abundance harboring the particular fingerprint patterns, and will be used for the intuitive illustration of the core algorithm of PRISM.

2.2 Overview of algorithm

The main approaches taken by PRISM are based on the following notions. Each of them is separately discussed in the upcoming sections.

1. The applicability of PRISM heavily depends on the existence of fingerprint epiloci, which indeed is not a well-established conception. A growing body of evidence for the role of genome-wide DNA methylation reprograming in cancer and our empirical observations on fingerprint epiloci justify the existence of fingerprint epiloci. Consequently, epigenetic subclones, if they exist, can be traced with a sufficient number of fingerprint methylation patterns.
2. Before the PRISM analysis, methylation patterns undergo *in silico* proofreading which corrects for the erroneous methylation states. It is based on DNA methyltransferase 1 (DNMT1)-like hidden Markov model (HMM) that is designed to mimic the DNA methylation maintenance process of DNMT1. Notably, *in silico* proofreading serves as an effective calibrator of estimated subclonal abundance, and it also increases the number of fingerprint epiloci enough for inferring epigenetic subclones as shown in Section 4.1.
3. The estimate of subclonal abundance is drawn from individual 'one-versus-all the other' binary pattern decomposition problem for each fingerprint epilocus. If there are k epigenetic subclones, these subclonal abundance estimates will be clustered around k genuine subclonal abundances. Thus, the problem can be viewed as a k-mixture decomposition problem, which is modeled as a beta-binomial mixture and solved by the expectation-maximization (EM) algorithm.

2.2.1 Existence of fingerprint epiloci

Using the prevalence of subclonal variant as an estimate of subclonal abundance has been a successful strategy for mutation-based subclonal inference (Miller *et al.*, 2014; Roth *et al.*, 2014). PRISM adopts a similar approach, by considering methylation fingerprint patterns as subclonal variants. Therefore, justifying the existence of fingerprint epiloci is crucial for the feasibility of PRISM.

Our point of view on the epigenetic subclonal inference problem is shown in Figure 1A. Aberration of methylation in cancer is characterized by genome-wide hypomethylation, as well as focal hypermethylation at regions including CpG islands (Baylin and Jones, 2011). Also, the global alterations of DNA methylation have been extensively studied in cancer, especially in leukemias (Heller *et al.*, 2016; Oakes *et al.*, 2016). Although defective epigenetic regulators are thought to affect DNA methylation reprogramming, the precise mechanism of the phenomenon still needs to be clarified. Nevertheless, accumulating evidence suggests that the methylation landscape of cancer evolves under selective pressure (Mazor *et al.*, 2016), implying that epigenetic subclones expand due to the increased fitness conferred by reprogrammed methylomic profile. Given the existence of epigenetic subclones, we postulated that distinct methylation patterns (e.g. regional hypermethylation) that uniquely define the subclone could be found by scrutinizing their methylation profiles. Therefore, the region with those distinct methylation patterns (i.e. fingerprint epilocus) became the principal unit of analysis in PRISM workflow because it could be treated as evidence of epigenetic subclone.

Furthermore, we have encountered some empirical observations on fingerprint epiloci in several, if not all, clinical samples from acute myeloid leukemia (AML) patients. Our preliminary analysis revealed that the fingerprint epiloci are uniformly distributed throughout the genome (Supplementary Fig. S1), thereby supporting our supposition that fingerprint epiloci arise from global DNA methylation reprogramming. More concrete examples of these findings are going to be discussed in Section 4.3.

2.2.2 *In silico* proofreading of methylation pattern based on DNMT1-like HMM

DNA methylation data sequenced by RRBS protocol suffers from diverse sources of errors such as the relatively high error rate of DNMT1 that is responsible for DNA methylation maintenance, or

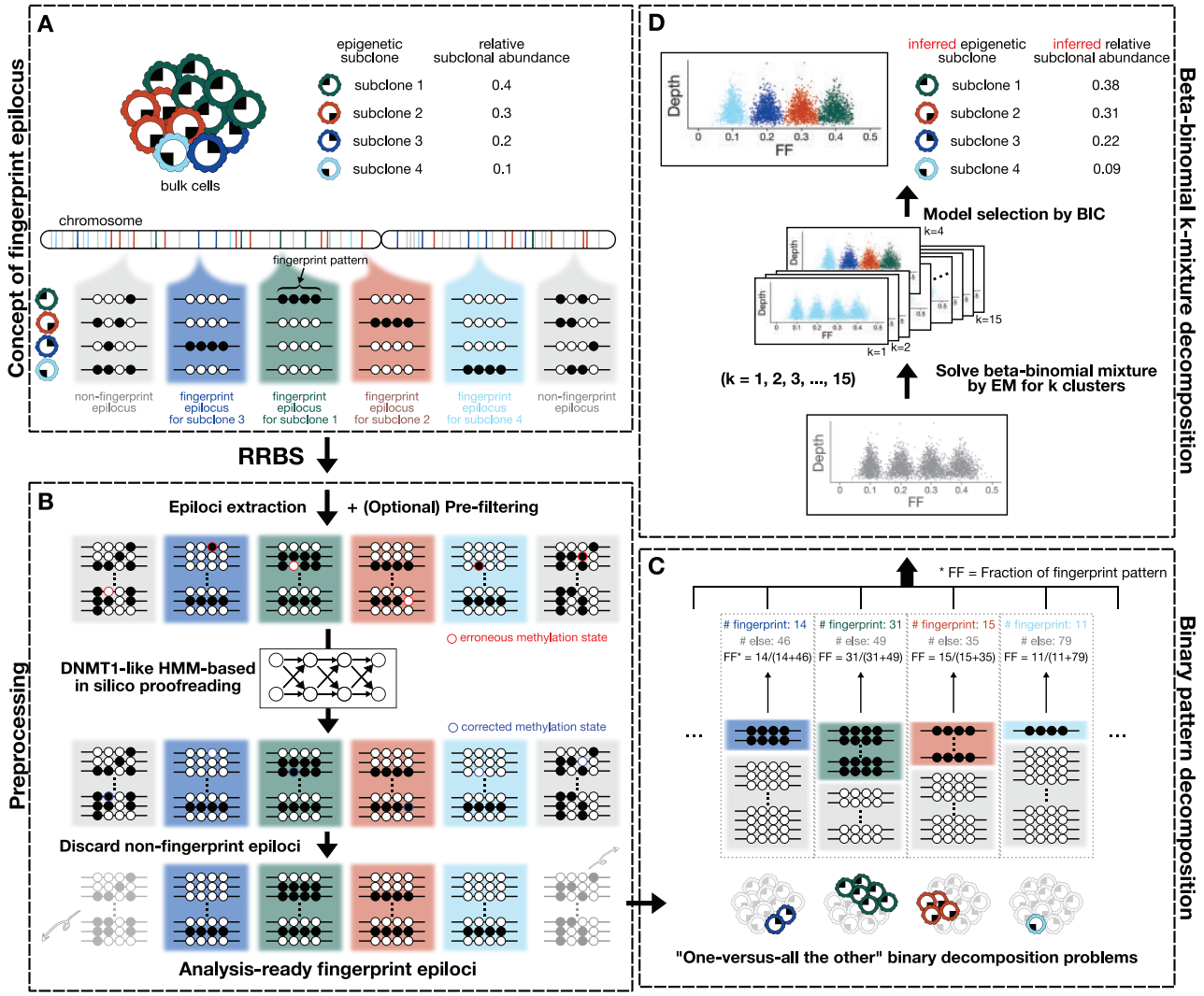

Fig. 1. Workflow of PRISM. (**A**) The concept of fingerprint epilocus. White and black circles denote unmethylated and methylated CpGs, respectively. Assume a bulk tumor comprising four epigenetic subclones (denoted by green, red, blue, sky blue cells) with relative abundance as shown in the figure. We expect fingerprint pattern (illustrated as four consecutive black circles) of the four subclones to be found throughout the genome. Inside the chromosome ideogram, fingerprint epiloci are shown as vertical bars with corresponding colors. Shown below are some detailed examples of fingerprint epiloci with the status of methylation patterns. For simplicity, only fully methylated fingerprint patterns are considered as fingerprint pattern in the figure. However, fully unmethylated fingerprint patterns will also be taken into account in the *post hoc* processing step of actual analyses. (**B**) Preprocessing step. Initially, RRBS mapping result is used for epiloci extraction, and optional pre-filtering is done to facilitate downstream steps. Methylation patterns obtained by RRBS are susceptible to error (red circles) for various reasons. Therefore, DNMT1-like HMM-based *in silico* proofreading attempts to correct for these errors. Obvious non-fingerprint epiloci will be subsequently discarded. Finally, we obtain analysis-ready fingerprint epiloci for the main analysis of PRISM. (**C**) Binary pattern decomposition problem. Analysis-ready fingerprint epiloci consist of nearly dichotomous groups of methylation patterns, namely, fully methylated and unmethylated patterns. Separate counting of fingerprint and non-fingerprint patterns gives estimates for the relative abundance of subclones. (**D**) Beta-binomial mixture model fitting and model selection. All the solutions of binary pattern decomposition problems are merged into a single beta-binomial mixture problem. Each k-cluster model is fit by EM algorithm, where $k = 1, 2, \ldots, 15$ by default. Among the model fits, the model that best explains the data are chosen by selecting the model with the minimum BIC. Finally, the subclonal inference result is obtained, and used for further analyses such as functional annotation

incomplete bisulfite conversion. Meanwhile, PRISM utilizes the exact count of specific methylation patterns, which is vulnerable to even small amounts of biased errors. Thus, techniques used for subclonal inference at mutation level (Miller *et al.*, 2014; Roth *et al.*, 2014) are not directly applicable to PRISM. Motivated by the successful application of *in silico* error correction algorithm of DNA sequencing data for fragment assembly (Pevzner *et al.*, 2001), we developed a novel *in silico* proofreading algorithm for methylation patterns to overcome this difficulty.

Computational modeling of DNMT1 enzymology is at the core of our methylation error correction algorithm (Figs 1B and 2A). We called this model as DNMT1-like HMM. It is indeed a generative model, which produces approximate copies of template methylation pattern that may contain some errors in them (Fig. 2B). Having established the generative model for DNA methylation maintenance, conversely, we can infer the template methylation pattern from the observed erroneous methylation patterns (Fig. 2B). This inference constitutes the essence of our error correction algorithm which is referred to as *in silico* proofreading of methylation patterns based on DNMT1-like HMM. Please refer to Section 3.1.2 for the full description of DNMT1-like HMM-based *in sillico* proofreading of methylation patterns.

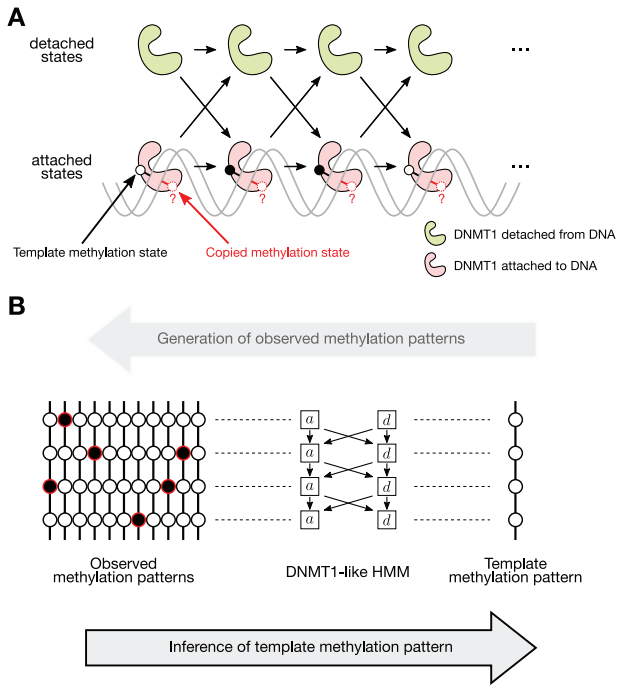

Fig. 2. Description of our DNMT1-like HMM. (**A**) Schematic diagram of DNA methylation maintenance. The diagram shows the template methylation states (black-outlined circles) being copied. Copied methylation states (red-outlined circles) are yet unknown. We can think of DNMT1 in two states: detached from DNA (green) and attached to DNA (red). The state of DNMT1 changes as it travels along the DNA. In particular, the transition from attached state to attached state denotes 'processive maintenance', and the transition from detached state to attached state denotes 'recruitment' of DNMT1. As it is structured in the diagram, we argue that this process can be properly modeled with HMM. Empty circles and filled circles denote unmethylated and methylated cytosine, respectively. (**B**) Properties of DNMT1-like HMM. By modeling the generative process of DNA methylation maintenance, our model clearly possesses the property of generative model (leftwards arrow). It means that the probability of an observed methylation pattern is defined by our model, given a template methylation pattern. We exploit those probabilities to infer template methylation pattern from observed methylation patterns (rightwards arrow). Squares with *a* and *d* represent 'attached' and 'detached' hidden states, respectively

2.2.3 *k*-mixture decomposition for epigenetic subclonal inference

The core analysis of PRISM can be viewed as a *k*-mixture decomposition problem after the HMM-based proofreading of methylation patterns. In this section, we give brief illustration of *k*-mixture decomposition problem. Algorithmic details of *k*-mixture decomposition are described in Section 3.2.

Subclonal abundance estimates drawn from binary pattern decomposition problems.

Decomposing methylation patterns for each of fingerprint epiloci generates an estimate of subclonal abundance, even though we do not know from which subclone the fingerprint pattern is originated.

Specifically, 'dichotomous' epiloci will only be considered as fingerprint epiloci where most of its methylation patterns are fully methylated or unmethylated, and this binarization significantly facilitates the whole workflow. Therefore, for each fingerprint epilocus, we can reduce the problem into a 'binary' pattern decomposition problem (one subclone versus all the other subclones) (Fig. 1C). It is because there are only two types of methylation patterns, i.e. fingerprint pattern from one subclone and non-fingerprint pattern from all the other subclones. To simplify the approach, we regard fully methylated patterns as fingerprint patterns and fully unmethylated patterns as non-fingerprint patterns. With that simplification, finally, a binary pattern decomposition results in two values for each fingerprint epilocus *i*: the number of fingerprint patterns (m_i) and non-fingerprint patterns (u_i). These two values (m_i, u_i) together constitute a sufficient statistic for underlying subclonal abundance, where the maximum likelihood estimator of subclonal abundance is $m_i/(m_i+u_i)$ which is equivalent to FF. Determining the subclone from which (m_i, u_i) is originated is actually a *k*-mixture decomposition problem.

Beta-binomial k-mixture decomposition for subclonal inference.

Given a set of statistics from binary decompositions of E fingerprint epiloci, say $\{(m_1,u_1),\ldots,(m_E,u_E)\}$, the problem is to determine which subclone is responsible for (m_i, u_i) among k subclones for each fingerprint epilocus. Conceptually, we can expect subclonal abundance estimate $m_i/(m_i+u_i)$ (FF) to be distributed around the genuine subclonal abundances, and these true subclonal abundances are likely to be detected by solving a *k*-mixture problem (Fig. 1D). In practice, we solve the mixture problem not with subclonal abundance estimate $m_i/(m_i+u_i)$ but with its sufficient statistics (m_i, u_i). Value of m_i can be intuitively viewed as a binomial random variable parameterized by the number of trial m_i+u_i and subclonal abundance as its probability. However, a common binomial model often does not account for the overdispersion of NGS (Heinrich *et al.*, 2012), and thus underestimates its variance. Therefore, we introduced the beta-binomial model instead of the binomial model and attempted to solve the beta-binomial mixture model for the exact counts of fingerprint and non-fingerprint methylation patterns.

3 Methods

3.1 Preprocessing of RRBS mapping result

3.1.1 Extraction of epiloci and methylation patterns from RRBS data

Epiloci and methylation patterns assigned to them were extracted from RRBS mapping data. Mapped reads harboring the same set of CpGs are grouped, and the read groups carrying at least d reads with at least c CpGs were retained for further analysis. By default, d was set to 20, and c was set to 4. Since subsequent *in silico* proofreading is a resource-intensive step, we allowed an optional pre-filtering before *in silico* proofreading step if a sufficient amount of data is given. A read group was pre-filtered out if it did not meet any of the following criteria, which is rather lenient: (i) The two most frequent patterns together should be fully methylated or unmethylated. (ii) The two most frequent patterns should account for >50% (by default) of reads mapped at the epilocus.

3.1.2 DNMT1-like generative HMM-based *in silico* proofreading of methylation patterns

We suggest that the enzymology of DNMT1 can be elegantly modeled by HMM, as demonstrated in Figure 2A. Suppose a situation in which DNMT1 tries to maintain methylation patterns on hemimethylated DNA duplex. There are two possible states of DNMT1 enzyme with respect to the target DNA: DNMT1 attached to DNA and detached from DNA. Accordingly, our DNMT1-like HMM consists of two hidden states, attached (a) and detached (d) states. To account for the processive methylation of DNMT1, we introduced two parameters, processivity of DNMT1 (p) and recruitment efficiency of DNMT1 to DNA (q) (Supplementary Fig. S2A and B). For example, the probability of transiting from state a to a is p, and from state d to d is $1-q$. Then we modeled DNMT1 copying the methylation status from existing methylation pattern on the

template strand (template pattern) to the daughter strand by unrolling the HMM configuration by the length of template pattern (Supplementary Fig. S2C). For each CpG site, emission probability of observed methylation status depends on both corresponding hidden state and methylation state of template pattern. In attached state, probabilities of emitting observed methylation status (o) from template methylation status (t) are defined with the error rate of DNMT1 (ϵ_a) as follows:

$$e_{a,t}(o) = \begin{cases} 1-\epsilon_a & \text{if } t = o \\ \epsilon_a & \text{if } t \neq o \end{cases} \quad (1)$$

For example, the probability that methylated cytosine (m) is emitted from attached state (a) and unmethylated template cytosine (u) is represented as $e_{a,u}(m)$ and its value is ϵ_a. Emission probabilities from the detached state are defined as follows:

$$e_{d,m}(o) = \begin{cases} 1-\epsilon_d & \text{if } o = u \\ \epsilon_d & \text{if } o = m \end{cases}$$
$$e_{d,u}(o) = \begin{cases} 1-\epsilon_b & \text{if } o = u \\ \epsilon_b & \text{if } o = m \end{cases} \quad (2)$$

here ϵ_d and ϵ_b denote the overall sequencing error rate and bisulfite conversion error rate, respectively.

Given the template methylation pattern, we could compute the probability of any methylation pattern using forward algorithm. These DNMT1-like HMM-based pattern probabilities were used in the hard-EM algorithm for subsequent template inference, which finally results in the set of the most likely template methylation patterns. The details of the forward algorithm and the hard-EM algorithm used for DNMT1-like HMM-based *in silico* proofreading is provided in Supplementary Information.

3.1.3 Identification of fingerprint epiloci

After correcting for the errors, we discarded an epilocus from further analysis if it did not meet any of the following criteria, which is stricter than pre-filtering criteria: (i) The two most frequent patterns should be fully methylated or unmethylated. (ii) The two most frequent patterns should account for >80% of reads mapped at the epilocus.

3.2 PRISM model

3.2.1 Establishment of PRISM model

We let m_i and u_i denote the number of reads at fingerprint epilocus i which supports methylated patterns and unmethylated patterns, respectively (Supplementary Fig. S3). We modeled two values, $m_i + u_i$ and m_i, with beta-binomial distribution parameterized by α and β. The model for fingerprint epilocus i became as follows:

$$m_i \sim BetaBinomial(m_i + u_i, \alpha, \beta) \quad (3)$$

We attempted to solve beta-binomial mixture model and derive K clusters from E fingerprint epiloci in total, where K is defined *a priori*. More specifically, for a cluster k, independent parameters α_k and β_k were introduced, and the optimal values of α_k and β_k were determined by EM algorithm.

For the following descriptions, we let superscripts in parenthesis denote the iteration number of EM algorithm. All of the cluster weight $\pi_k^{(1)}$ were initialized with $1/K$. For reasonable initialization of $\alpha_k^{(1)}$ and $\beta_k^{(1)}$, we fit Gaussian mixture model with values of $m_i/(m_i + u_i)$, or FFs. Fitting Gaussian mixture model gives mean and variance for each cluster k, denoted by μ_k and σ_k^2, respectively. We could initialize $\alpha_k^{(1)}$ and $\beta_k^{(1)}$ using the result of Gaussian mixture fit as follows:

$$\alpha_k^{(1)} = \left(\frac{1-\mu_k}{\sigma_k^2} - \frac{1}{\mu_k}\right)\mu_k^2$$
$$\beta_k^{(1)} = \alpha_k^{(1)}\left(\frac{1}{\mu_k} - 1\right) \quad (4)$$

In the E-step, $L_{ik}^{(n)}$, the likelihood of $\alpha_k^{(n)}$ and $\beta_k^{(n)}$ with regard to epilocus i was computed with beta-binomial probability mass function f. Assume we are at iteration n of EM algorithm.

$$L_{ik}^{(n)} = f(m_i | m_i + u_i, \alpha_k^{(n)}, \beta_k^{(n)})$$
$$= \frac{\Gamma(m_i + u_i + 1)}{\Gamma(m_i + 1)\Gamma(u_i + 1)} \frac{\Gamma(m_i + \alpha_k^{(n)})\Gamma(u_i + \beta_k^{(n)})}{\Gamma(m_i + u_i + \alpha_k^{(n)} + \beta_k^{(n)})} \frac{\Gamma(\alpha_k^{(n)} + \beta_k^{(n)})}{\Gamma(\alpha_k^{(n)})\Gamma(\beta_k^{(n)})} \quad (5)$$

Accordingly, the posterior probability of epilocus i being assigned to cluster k ($p_{ik}^{(n)}$) can be computed as follows:

$$p_{ik}^{(n)} = \frac{\pi_k^{(n)} L_{ik}^{(n)}}{\sum_{j=1}^{K} \pi_j^{(n)} L_{ij}^{(n)}} \quad (6)$$

In the M-step, $\pi_k^{(n+1)}$, $\alpha_k^{(n+1)}$ and $\beta_k^{(n+1)}$ are computed. The maximum likelihood estimation of $\pi_k^{(n+1)}$ is straightforward:

$$\pi_k^{(n+1)} = \frac{\sum_{i=1}^{E} p_{ik}^{(n)}}{E} \quad (7)$$

However, the maximum likelihood estimation of $\alpha_k^{(n+1)}$, and $\beta_k^{(n+1)}$ given m, u, and $p^{(n)}$, which is not trivial, is done through Newton's iteration (Minka, 2000). A more detailed explanation of the maximum likelihood estimation of $\alpha_k^{(n+1)}$ and $\beta_k^{(n+1)}$ is demonstrated in Supplementary Information.

The termination condition of the EM iteration is described as follows. For each iteration, we computed log likelihood $l^{(n)}$:

$$l^{(n)} = \sum_{i=1}^{E} \log \sum_{k=1}^{K} p_{ik}^{(n)} \quad (8)$$

If $|l^{(n)} - l^{(n-1)}| < 0.001$, the EM iteration was terminated.

3.2.2 Model selection

We used Bayesian information criterion (BIC) (Schwarz *et al.*, 1978) to select the model with optimal number of clusters. BIC of a model with K clusters (M_K) is defined as

$$\mathrm{BIC}(M_K) = (3K - 1)\log E - 2\sum_{i=1}^{E} \log \sum_{k=1}^{K} p_{ik} \quad (9)$$

Since there are $2K$ free parameters for α and β, and $K - 1$ free parameters for π, M_K has $3K - 1$ free parameters in total. In practice, PRISM selects the model with minimum BIC among M_1 through M_{15} by default.

3.2.3 Joint analysis of multiple samples from single tumor

Subclonal inference generally benefits from joint analysis of multiple sequencing data from the same tumor because different samples often have different mixing proportion of subclones which increases the chance of separation for two subclones that could not be separated by investigating a single sample due to their similar relative abundance. Thus many of existing subclone detection tools are able to analyze several samples jointly to increase their resolution of subclone detection (Miller *et al.*, 2014; Roth *et al.*, 2014). PRISM can also be applied for two or more samples from a single tumor.

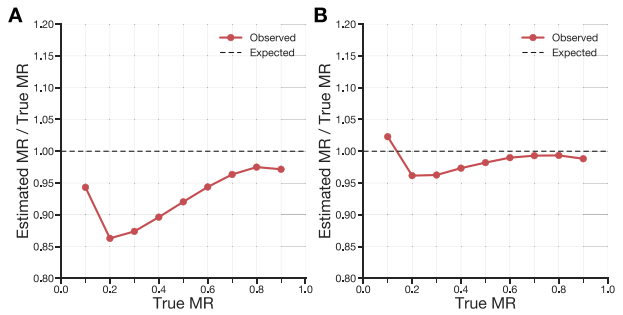

Fig. 3. Effect of *in silico* proofreading. Two raw RRBS data from fully methylated and unmethylated cell lines were mixed computationally. We let PRISM estimate the true MR of the mixtures (**A**) without or (**B**) with *in silico* proofreading. The accuracy of MR estimate was measured as its fraction with respect to the true MR. The expected values for perfect estimations are represented as dotted lines. MR, mixing ratio

For the detailed explanation of the multi-sample joint analysis, please refer to Supplementary Information.

3.2.4 *Post hoc* processing of clusters and copy number-aware analysis

After selecting the optimal model, PRISM excludes uninformative clusters that seem to have arisen from outliers. It also checks whether it should treat several clusters as single subclones by taking account of the unmethylated fingerprint, by merging 'reflected' clusters. Also, PRISM can utilize CNA information to obtain corrected methylation pattern counts which reflect methylation patterns originating from copy number-gained segments. For further details of *post hoc* processing step and copy number-aware analysis, please see Supplementary Information.

3.3 Data retrieval and processing

Methylated and unmethylated cell line RRBS data (Barrett *et al.*, 2017) were downloaded from European Nucleotide Archive under accession number PRJEB21102. MCF10A-Er-Src cell line, GM06990 B-lymphocyte cell line and T-47D cell line RRBS data were downloaded from Sequence Read Archive under run accession SRR222454, SRR531452 and SRR222532, respectively. We also downloaded public AML RRBS data along with corresponding whole exome sequencing (WES) data from dbGaP under accession phs001027.v2.p1. For the full description of the data processing step, please refer to Supplementary Information.

4 Results and discussion

4.1 Impact of *in silico* proofreading on PRISM analysis

We assessed the impact of *in silico* proofreading on the accuracy of the estimated size of subclones (Fig. 3). For that, we obtained two raw RRBS data representing methylation states of the fully methylated cell line and fully unmethylated cell line. We mixed two raw RRBS data to simulate a mixture of epigenetically homogeneous cells. Each of the two raw data were subsampled with 10%, 20%, …, 90% of reads to generate benchmark mixtures of the two cell lines. We then concatenated corresponding pairs of subsampled raw data such that their mixing ratio (MR) would sum up to 100%. For example, we joined 30%-subsampled fully methylated cell line RRBS data and 70%-subsampled fully unmethylated cell line RRBS data together. This entire step was repeated for 10 times. We then examined the accuracy of MR estimates given by PRISM, with or without *in silico* proofreading.

Running PRISM without *in silico* proofreading resulted in markedly biased estimations, which consistently underestimated MR (Fig. 3A). The worst estimation was for the MR of 20%, where the ratio between estimated MR and true MR was 0.86. The estimations after *in silico* proofreading were more 'calibrated' to correct estimations (Fig. 3B), and the ratio between the worst estimation of MR and true MR was 0.96 for MR of 20%. *In silico* proofreading seems to correct for the biased error rate of methylation patterns, which is significantly higher for fully methylated patterns than for fully unmethylated patterns (Supplementary Fig. S4). Meanwhile, we were also possible to show that *in silico* proofreading facilitated PRISM analysis by considerably increasing the number of fingerprint epiloci for the inference of epigenetic subclones with marginal introduction of artificial fingerprint epiloci (Supplementary Fig. S5).

4.2 Simulated mixture of tissue cell lines

We generated more realistic benchmark mixtures of cells by mixing cell line RRBS data established from various tissues to evaluate the performance of PRISM in practical situations. In particular, three cell line RRBS data were carefully chosen from ENCODE project (Varley *et al.*, 2013): (i) MCF10A-Er-Src cell line which is derived from non-tumorigenic epithelial cells of the mammary gland, (ii) GM06990 B-lymphocyte cell line from lymphoblastoid and (iii) T-47D cell line established from mammary ductal carcinoma. Epigenomic reprogramming plays a crucial role in development, shaping distinct methylation landscape for each cell type from different cell lineage. Therefore, PRISM should be able to detect global DNA methylation reprogramming event in order to distinguish a lymphocyte cell line (GM06990) from two epithelial cell lines (MCF10A-Er-Src and T-47D). Moreover, we asked whether PRISM could accurately separate non-cancerous (MCF10A-Er-Src) and cancerous (T-47D) cell lines, where both of them were derived from the mammary gland.

In this experiment, it should be noted that merely mixing the raw data will result in an undesired result because the sequencing libraries were prepared separately, so the cleavage site of restriction enzyme may differ between samples. Furthermore, the sequencing depth of the same epilocus will be strikingly different, which may affect the proportion of reads severely when a mixture is generated. Therefore, we mixed them with a deliberate approach as follows. The three raw RRBS data were independently processed and mapped to the reference genome. Then the epiloci which appear in all of three alignment results and have 20 or more mapped reads were retained for the mixing procedure. For each epilocus, simulated sequencing depth d was sampled from $NegBin(5, 0.03)$ with constraint $d \geq 20$. We randomly sampled MRs (P_1, P_2, P_3) from $Dirichlet(3, 3, 3)$, and for each epilocus, $\lfloor P_i d \rfloor$ reads were sampled from each of the three data. The entire mixing step was repeated two times to generate two independent mixtures (Supplementary Table S1).

We supposed each cell line as a putative subclone in the mixture, and let PRISM estimate the number and abundance of the subclones only from their mixed methylation patterns in the two mixtures. PRISM identified four subclones (Fig. 4A). Regarding the average FF of each cluster as MR estimate, we observed that the resulting MR estimates of subclones 1, 2 and 3 reasonably represented the true MRs (Fig. 4B and Supplementary Table S2). For subclone 4, which was unexpected, we could not draw a concrete conclusion on whether it was an artifact of sequencing procedures, or it was a

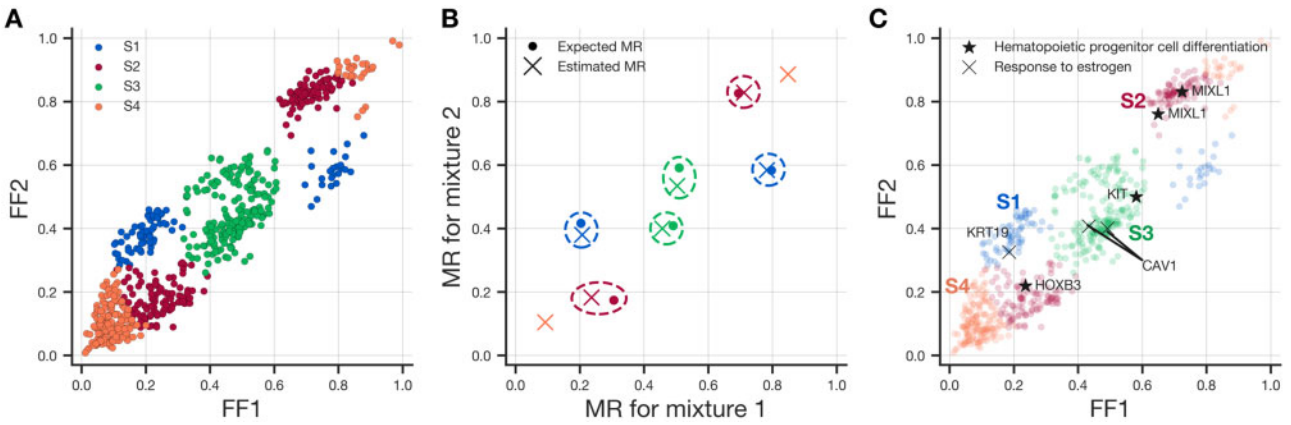

Fig. 4. PRISM results for simulated mixtures of three tissue cell lines. Epigenetic subclones were jointly inferred with two mixtures. (**A**) In two-sample inference, PRISM detected four subclones. (**B**) The accuracy of estimated mixing proportions. True underlying MRs are shown as X's. Expected MR and its 'reflection' (1 − MR), which accounts for the unmethylated fingerprint patterns, are shown as colored dots. Three genuine subclones (blue, red and green) were captured by PRISM. Dotted circles group expected MRs and their closest estimates. (**C**) Functional annotation of detected subclones. Two sets of marker genes for two biological processes, 'hematopoietic progenitor cell differentiation' and 'response to estrogen', were used. S1 through S4, subclone 1 through subclone 4; FF, fraction of fingerprint pattern; MR, mixing ratio

genuine subclone originated from clonal evolution of the cell line. By comparing MR estimates with true MRs, we postulated that subclones 1, 2 and 3 represent MCF10A-Er-Src, GM06990 and T-47D cell line, respectively.

Epiloci assigned to infer subclones were then functionally annotated by seeking for the epiloci overlapping predefined sets of marker genes. Two sets of marker genes that may help distinguish each cell type were used, which were annotated to GO terms 'hematopoietic progenitor cell differentiation' (GO: 0002244) and 'response to estrogen' (GO: 0043627). We annotated an epilocus with a gene if it had overlap with either of the gene body region or the upstream 3 000 bp region from TSS which accounts for the promoter region. Subclone 2 exhibited two epiloci associated with the differentiation of hematopoietic progenitor cells. Each of the epiloci was annotated with *HOXB3* and *MIXL1*. *HOXB3* has been implicated in regeneration of hematopoietic stem cell (Björnsson *et al.*, 2003) and early hematopoiesis (Sauvageau *et al.*, 1997). Meanwhile, *MIXL1* is shown to be required to determine the fate of cells from the primitive streak to blood (Ng *et al.*, 2005), and it has been used as a marker of primitive hematopoietic stem cell (Davis *et al.*, 2008). The evidence collectively shows subclone 2 belongs to hematopoietic lineage. Similarly, subclone 3 carried epiloci for *CAV1*. The methylation status of CpG shore in its promoter has been reported to be associated with aggressiveness of breast cancer (Rao *et al.*, 2013). Therefore, we showed that PRISM successfully reconstructed the underlying cellular composition.

4.3 Detection of epigenetic subclones in AML dataset

We next applied PRISM for AML data to test if PRISM can draw clinically meaningful observations. For each subject, a couple of samples were taken at two time points (diagnosis and relapse) and sequenced by RRBS. Two-sample joint analysis were done, which resulted in 3.13 inferred subclones on average (Supplementary Fig. S6). In this section, we describe our explanatory analyses for subject AML-105 and AML-109, since both of them had five inferred subclones, which seemed to be sufficient to reveal intriguing clonal dynamics. Microscopic inspection of data providers revealed that both of the samples had relatively normal cytogenetic properties, except for AML-105 relapse sample, which had a small fraction (∼10%) of cells harboring genomic deletion in q-arm of chromosome 7. Moreover, no significant CNA was detected from WES data of those samples (Supplementary Fig. S7). Therefore, we concluded that the CNA of the samples would not affect our analysis.

For AML-105, initially, diagnosis and relapse samples were analyzed separately (Fig. 5A). Four and three putative epigenetic subclones were found, respectively. However, the two-sample joint analysis identified five epigenetic subclones (Fig. 5B), emphasizing the necessity of multi-sample joint analysis to achieve a reasonable resolution of subclonal inference. Interestingly, independent analysis of variant allele frequency (VAF) from WES data (Supplementary Fig. S8A) revealed that the subclonal abundance inferred by VAFs of heterozygous somatic mutations within isocitrate dehydrogenase 2 (*IDH2*) and DNA methyltransferase 3 alpha (*DNMT3A*) was concordant with the subclonal abundance estimates of subclone 2 (0.61–0.84) identified by PRISM. This indirectly suggests that somatic mutations in *IDH2* and *DNMT3A* may have served as driver events that accelerate subsequent epigenomic evolution, given the role of *IDH2* (Kernytsky *et al.*, 2015) and *DNMT3A* (Yang *et al.*, 2015) as epigenetic regulators (Fig. 5E). Additionally, possible epigenetic drivers of subclonal expansion were identified by functionally annotating identified subclones (Fig. 5C). We used a set of 68 recurrently mutated genes in AML (Metzeler *et al.*, 2016) (Supplementary Table S3) since we assumed that the epigenetic aberration of these genes would also be critical for the progression of AML. Subclone 3 harbored intensively altered methylation (hereafter referred to as *epi-mutation*) at Wilms' tumor 1 (*WT1*), whose overexpression and mutation have significant implications in AML (Menssen *et al.*, 1995; Miwa *et al.*, 1992). What is most striking is that dysregulated *WT1* has recently been known to alter the methylation landscape of cells by impeding the activity of TET2 (Rampal and Figueroa, 2016). Epi-mutations in *DNMT3A* characterized subclones 1 and 3, while subclone 1 was annotated with additional gene, GATA binding protein 2 (*GATA2*). One potential pitfall of PRISM is that it cannot avoid reporting a cluster of epiloci arisen from genomic imprinting. However, it can simply overcome by annotating subclones with known imprinted genes and excluding the subclone showing a considerably high proportion of imprinted epiloci (Fig. 5D). A curated list of imprinted genes was obtained from Geneimprint (Falls *et al.*, 1999). Therefore, subclone 4 was

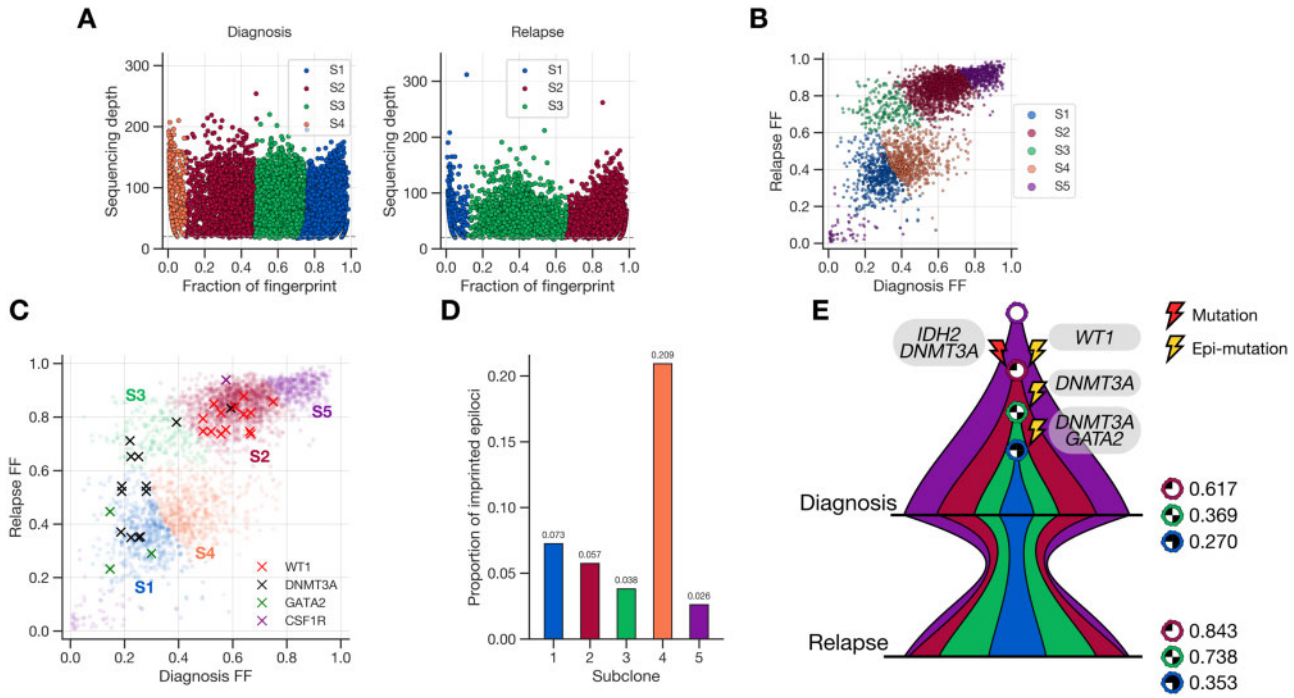

Fig. 5. PRISM results for AML-105. (**A**) Separate single-sample analyses of biopsies taken at the time of diagnosis and relapse. PRISM detected four and three candidate epigenetic subclones for diagnosis and relapse sample, respectively. (**B**) Two-sample joint analysis. PRISM reported five epigenetic subclones. (**C**) Functional annotation of the joint analysis result. Recurrently mutated genes in AML are used for annotation. Each mark denotes epilocus which overlaps with the corresponding gene or its promoter. (**D**) The proportion of imprinted epiloci for each of the putative epigenetic subclones. Notably, 20.9% of epiloci assigned to subclone 4 were annotated to known imprinted genes. Therefore, we excluded subclone 4 from further analyses. (**E**) Schematic diagram of inferred evolutionary history. One of the possible evolutionary histories of identified epigenetic subclones is shown. Mutations and epi-mutations characterizing each subclone are represented. The horizontal black line represents the time point at which the two biopsies were taken. Considerably, subclone 3 (green) underwent rapid clonal expansion (from relative abundance 0.369 to 0.738) after chemotherapy. FF, fraction of fingerprint pattern; S1 through S5, subclone 1 through subclone 5

excluded from further analyses since 20.9% of its epiloci were known to be imprinted. Finally, we could infer one of the possible evolutionary histories of subclones regarding both genomic and epigenomic events that the subclones underwent (Fig. 5E).

PRISM revealed much intriguing clonal dynamics for AML-109 (Fig. 6). Three subclones were found in each of the separate analysis of diagnosis and relapse sample (Fig. 6A), while joint analysis of the two longitudinal samples revealed five subclones (Fig. 6B). Similarly, independent whole exome analysis revealed a novel subclone at relapse that was not detectable at diagnosis (Supplementary Fig. S8B). The novel subclone carried somatic mutations in four AML-related genes, including *WT1* and *IDH2*. Interestingly, functional annotation again discovered altered methylation of *WT1* (Fig. 6C). Regarding subclone 4 as a representative of the whole clonal population, *WT1* methylation alteration was clonal rather than subclonal in this case; in other words, it was deemed to have occurred at the onset of the malignancy. Furthermore, subclone carrying epiloci at *DNMT3A* (subclone 3) showed dramatic increase in its relative abundance (from 0.062 to 0.921). It implies that subclone 3 had a remarkable evolutionary advantage over other subclones in the tumor niche established by anti-cancer drugs, presumably due to the combined effect of genomic and epigenomic variations. Subclone 1 was excluded from the inference of evolutionary history because of its significantly high proportion of imprinted epiloci (Fig. 6D). The reconstructed subclonal evolutionary trajectory (Fig. 6E) verified *WT1* epi-mutation as an early event in clonal evolution, followed by *MPL*, *JAK2* and also *DNMT3A* epi-mutation accompanied by several mutations.

5 Conclusion

While it has long been considered that the evolution of the cancer genome is the primary factor constituting inherent heterogeneity of a tumor, the evolution of epigenome brings up another dimension to defining the heterogeneity. Whether the genomic and epigenomic evolution occur co-ordinatively or independently is still obscure, and even seem to be case-dependent (Li *et al.*, 2016), which is reproduced by comparing PRISM and two SNV-based subclonal inference methods (Supplementary Fig. S9). Nevertheless, investigating the epigenomic evolutionary history of a tumor at the resolution of cancer subclone provides valuable insight into the epigenetic mechanism of the progression of the malignancy. It also offers novel implications for the history of clonal evolution and helps design the therapeutic strategy.

Despite the limitation of our research that we could not provide experimental evidence of the existence of the subclones, we showed that the inference of epigenetic subclonal population structure was possible by focusing on the fingerprint epiloci that seem to have arisen from global DNA methylation reprogramming. Analyzing clonal dynamics of the two AML samples implied the significance of interplay between epigenetic regulators such as *WT1*, *IDH2* and *DNMT3A* in clonal evolution. Moreover, by combining information of genomic variation, we could gain valuable insight into the simultaneous evolution of the genomic and epigenomic landscapes. Consequently, we carefully suggest that the development of epigenomic subclonal inference algorithm brings us one step closer to the multi-omics level characterization

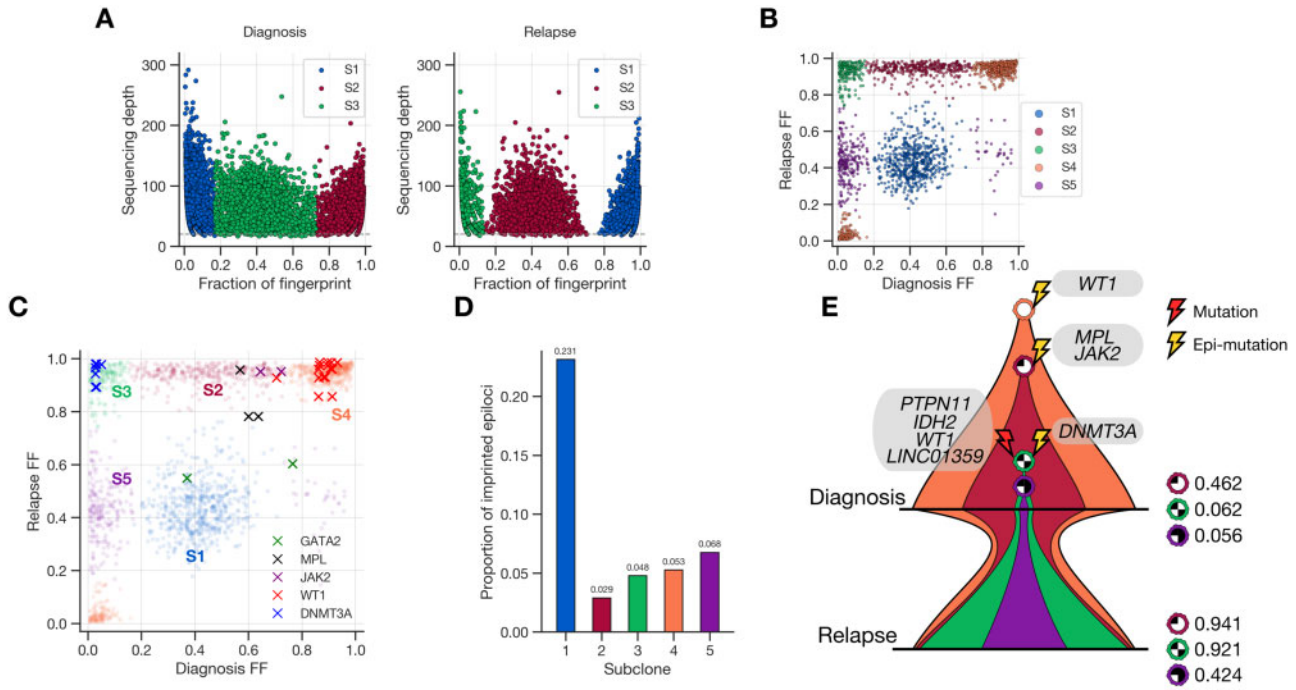

Fig. 6. PRISM results for AML-109. (**A**) Separate single-sample analyses of biopsies taken at the time of diagnosis and relapse. PRISM detected three candidate epigenetic subclones for each sample. (**B**) Two-sample joint analysis. PRISM reported six epigenetic subclones. (**C**) Functional annotation of the joint analysis result. Total five genes that are recurrently mutated in AML were detected. (**D**) The proportion of imprinted epiloci for each of the epigenetic subclones. We excluded subclone 1 from further analyses since 23.1% of epiloci assigned to subclone 1 were imprinted. (**E**) Schematic diagram of inferred evolutionary history. One of the possible evolutionary histories of identified epigenetic subclones is shown. Mutations and epi-mutations characterizing each subclone are shown. The horizontal black line represents the time point at which the two biopsies were taken. Considerably, subclone 3 (green) underwent rapid clonal expansion (from relative abundance 0.062 to 0.921) after chemotherapy. FF, fraction of fingerprint pattern; S1 through S5, subclone 1 through subclone 5

of cancer subclones, which is one of the ultimate goals of molecular oncology.

Funding

This research is supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT(No.NRF-2017M3C4A7065887), the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (No.NRF2014M3C9A3063541), and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI15C3224).

Conflict of Interest: none declared.

References

Almendro,V. *et al.* (2014) Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep.*, **6**, 514–527.

Barrett,J.E. *et al.* (2017) Quantification of tumour evolution and heterogeneity via Bayesian epiallele detection. *BMC Bioinform.*, **18**, 354.

Baylin,S.B. and Jones,P.A. (2011) A decade of exploring the cancer epigenome-biological and translational implications. *Nat. Rev. Cancer*, **11**, 726.

Björnsson,J.M. *et al.* (2003) Reduced proliferative capacity of hematopoietic stem cells deficient in Hoxb3 and Hoxb4. *Mol. Cell. Biol.*, **23**, 3872–3883.

Brocks,D. *et al.* (2014) Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep.*, **8**, 798–806.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413.

Davis,R.P. *et al.* (2008) Targeting a gfp reporter gene to the mixl1 locus of human embryonic stem cells identifies human primitive streak-like cells and enables isolation of primitive hematopoietic precursors. *Blood*, **111**, 1876–1884.

Falls,J.G. *et al.* (1999) Genomic imprinting: implications for human disease. *Am. J. Pathol.*, **154**, 635–647.

Ferrando,A.A. and López-Otín,C. (2017) Clonal evolution in leukemia. *Nat. Med.*, **23**, 1135.

Gerlinger,M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N E J. Med.*, **366**, 883–892.

Gundem,G. *et al.* (2015) The evolutionary history of lethal metastatic prostate cancer. *Nature*, **520**, 353.

Heinrich,V. *et al.* (2012) The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res.*, **40**, 2426–2431.

Heller,G. *et al.* (2016) Next-generation sequencing identifies major DNA methylation changes during progression of Ph+ chronic myeloid leukemia. *Leukemia*, **30**, 1861.

Hiley,C. *et al.* (2014) Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine. *Genome Biol.*, **15**, 453.

Kernytsky,A. *et al.* (2015) IDH2 mutation-induced histone and DNA hypermethylation is progressively reversed by small-molecule inhibition. *Blood*, **125**, 296–303.

Kreso,A. *et al.* (2013) Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science*, **339**, 543–548.

Landau,D.A. *et al.* (2014) Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*, **26**, 813–825.

Li,S. *et al.* (2014) Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biol.*, **15**, 472.

Li,S. *et al.* (2016) Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.*, **22**, 792.

Mazor,T. *et al*. (2016) Intratumoral heterogeneity of the epigenome. *Cancer Cell*, **29**, 440–451.

Menssen,H. *et al*. (1995) Presence of Wilms' tumor gene (WT1) transcripts and the WT1 nuclear protein in the majority of human acute leukemias. *Leukemia*, **9**, 1060–1067.

Merlo,L.M. *et al*. (2006) Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6**, 924.

Metzeler,K.H. *et al*. (2016). Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood*, **128**, 686–698.

Miller,C.A. *et al*. (2014) Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol*., **10**, e1003665.

Minka,T. (2000). Estimating a Dirichlet distribution. *Technical report, MIT*, **1**, 4.

Miwa,H. *et al*. (1992) Expression of the Wilms' tumor gene (WT1) in human leukemias. *Leukemia*, **6**, 405–409.

Mroz,E.A. and Rocco,J.W. (2013) Math, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol*., **49**, 211–215.

Ng,E.S. *et al*. (2005) The primitive streak gene Mixl1 is required for efficient haematopoiesis and BMP4-induced ventral mesoderm patterning in differentiating ES cells. *Development*, **132**, 873–884.

Nowell,P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–28.

Oakes,C.C. *et al*. (2016) DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet*., **48**, 253.

Oesper,L. *et al*. (2014) Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, **30**, 3532–3540.

Park,Y. *et al*. (2016) Measuring intratumor heterogeneity by network entropy using RNA-seq data. *Sci. Rep*., **6**, 37767.

Pevzner,P.A. *et al*. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA*, **98**, 9748–9753.

Rampal,R. and Figueroa,M.E. (2016) Wilms tumor 1 mutations in the pathogenesis of acute myeloid leukemia. *Haematologica*, **101**, 672–679.

Rao,X. *et al*. (2013) Cpg island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*, **32**, 4519.

Roth,A. *et al*. (2014) Pyclone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396.

Sauvageau,G. *et al*. (1997) Overexpression of hoxb3 in hematopoietic cells causes defective lymphoid development and progressive myeloproliferation. *Immunity*, **6**, 13–22.

Schwarz,G. *et al*. (1978) Estimating the dimension of a model. *Ann. Stat*., **6**, 461–464.

Turajlic,S. and Swanton,C. (2016) Metastasis as an evolutionary process. *Science*, **352**, 169–175.

Varley,K.E. *et al*. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*., **23**, 555–567.

Yang,L. *et al*. (2015) DNMT3A in haematological malignancies. *Nat. Rev. Cancer*, **15**, 152.

Yates,L.R. *et al*. (2017) Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell*, **32**, 169–184.