# SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution

Christopher A. Miller[1,†],Brian S. White[1,2,†], Nathan D. Dees[1], Malachi Griffith[1], John S. Welch[2,3], Obi L. Griffith[1,2], Ravi Vij[2,3], Michael H. Tomasson[2,3], Timothy A. Graubert[2,3], Matthew J. Walter[2,3], Matthew J. Ellis[2,3], William Schierding[1], John F. DiPersio[2,3], Timothy J. Ley[1,2,3], Elaine R. Mardis[1,3,4], Richard K. Wilson[1,3,4], and Li Ding[1,2,3,4,∗]

**1 The Genome Institute, Washington University, St. Louis, Missouri, USA**

**2 Department of Internal Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, Missouri, USA**

**3 Siteman Cancer Center, Barnes-Jewish Hospital, Washington University School of Medicine, St. Louis, Missouri, USA**

**4 Department of Genetics, Washington University, St. Louis, Missouri, USA**

**† These authors contributed equally to this work**

**∗ E-mail: lding@genome.wustl.edu**

# Abstract

The sensitivity of massively-parallel sequencing has confirmed that most cancers are oligoclonal, with subpopulations of neoplastic cells harboring distinct mutations. A fine resolution view of this clonal architecture provides insight into tumor heterogeneity, evolution, and treatment response, all of which may have clinical implications. Single tumor analysis already contributes to understanding these phenomena. However, cryptic subclones are frequently revealed by additional patient samples (e.g., collected at relapse or following treatment), indicating that accurately characterizing a tumor requires analyzing multiple samples from the same patient. To address this need, we present SciClone, a computational method that identifies the number and genetic composition of subclones by analyzing the variant allele frequencies of somatic mutations. We use it to detect subclones in acute myeloid leukemia and breast cancer samples that, though present at disease onset, are not evident from a single primary tumor sample and, by doing so, track tumor evolution and identify the spatial origins of cells resisting therapy.

# Author Summary

Sequencing the genomic DNA of cancers has revealed that tumors are not homogeneous. As a tumor grows, new mutations accumulate in individual cells, and as these cells replicate, the mutations are passed on to their offspring, which comprise only a portion of the tumor when it is sampled. We present a method for identifying the fraction of cells containing specific mutations, clustering them into subclonal populations, and tracking the changes in these subclones. This allows us to follow the clonal evolution of cancers as they respond to chemotherapy or develop therapy resistance, processes which may radically alter the subclonal composition of a tumor. It also gives us insight into the spatial organization of tumors, and we show that multiple biopsies from a single breast cancer may harbor different subclones that respond differently to treatment. Finally, we show that sequencing multiple samples from a patient's tumor is often critical, as it reveals cryptic subclones that cannot be discerned from only one sample. This is the first tool that can efficiently leverage multiple samples to identify these as distinct subpopulations of cells, thus contributing to understanding the biology of the tumor and influencing clinical decisions about therapy.

# Introduction

Cancer is a disease largely driven by accumulated somatic mutations. Many of these are clonal mutations and occur in the founding cell to initiate disease. These become uniformly present in the tumor by propagating to that cell's progeny during clonal expansion. Others are subclonal events, which occur in an existing neoplastic cell and are then passed on only to the subpopulation of cells derived from it. The result of this accumulation of mutations is that tumors are composed of a heterogeneous mixture of cells. These subpopulations compete and evolve, [1–3] and the mutations "captured" [4] in subsets of cells during this evolution serve as a genetic signature of the resulting (sub)clones.

Recently, high resolution glimpses of this clonal heterogeneity have been provided by next-generation sequencing, [4–14] SNP array, [6, 10, 13, 15] and array comparative genomic hybridization [16, 17] platforms. Single-cell sequencing [16, 18, 19] may eventually address this heterogeneity directly without the confounding effects of mixing cell types, but technical challenges, such as allele dropout, [20] remain. There are also pragmatic concerns about the large number of cells that must be sequenced to establish the heterogeneity of a given sample. The emerging picture from these studies, across a diversity of solid [6, 8, 9, 11, 13, 16] and hematological [4, 5, 7, 10, 12, 14, 15, 17] disorders, is that tumors are both spatially [9, 13, 16] and temporally [4–17] heterogeneous and are frequently comprised of a single founding clone and several subclones.

Increasing evidence suggests that intra-tumor heterogeneity and clonal architecture have clinical implications [3, 21, 22] and contribute to therapy resistance. [23] Several studies have linked the presence of subclones to poor clinical outcome, as in chronic lymphocytic leukemia (CLL), [10] or to increased risk of progression to malignancy, as in Barrett's esophagus [24] and multiple myeloma (MM). [17] Subclonal mutations can drive resistance as well, as shown in *EGFR*-mutant non-small cell lung cancers. [25, 26] Studies in chronic myeloid leukemia have also demonstrated that drug-resistant subclones may harbor aggressive mutations that are restrained by more fit, but indolent clones. [2, 3] In these cases, therapeutic application of imatinib leads to competitive release of BCR-ABL mutant subclones, which renders the therapy ineffective. [3, 27] Thus, designing effective second line therapies requires a deep understanding of both a cancer's underlying mutations and how it's clonal structure evolves in response to treatment.

Existing methods [6, 10, 11, 28, 29] have been useful for inferring clonal architecture and its consequences, e.g., that putative driver mutations in *SF3B1* and *TP53* in CLL [10] and in *PIK3CA* and *PTEN* in triple-negative breast cancer [11] may arise during late-emerging, subclonal diversification [6] rather than as founding lesions. Recent results suggest that accurately describing the subclonal composition and evolution of tumors requires sampling cancer cells across multiple time points or spatially-distinct regions, [3, 21, 23] with current studies of distant metastases and of spatial heterogeneity collecting as many as six to twenty samples. [9, 13, 16, 18] The scale of these ambitious studies will challenge existing methods. For example, histogram-based approaches to representing clonal markers [10, 28] are attractive in avoiding model assumptions in low dimensions (i.e., with few samples), but with many samples will suffer from exponential computational complexity. Several approaches [6, 10, 11, 28] leverage Markov chain Monte Carlo (MCMC) techniques, but these, too, are computationally demanding and rely on assumptions about chain convergence.

Most existing methods [6, 10, 28] inferring clonality from copy number alterations (CNAs) avoid additional computational overhead by making the simplifying assumption that the tumor sample is "monogenomic" [28] and does not harbor subclonal copy-number events. Contrary to their assumptions, these methods have detected such subclonal events in CLL, [10] though without being able to correct for them. Similar subclonal events have been detected in MM (Ref. 30 and B.S.W., R.V., and M.H.T, data not shown). These methods introduce uncertainty, through the probabilistic inference of allele-specific copy numbers, and error, by ignoring subclonal CNAs. A recent approach [29] does generalize to subclonal CNAs, but also suffers from computational inefficiencies when extended beyond the simple case of clonal CNAs. A method which could avoid the uncertainty of deconvolving subclonal CNAs and operated with significantly lower computational demands would benefit studies aiming to understand the evolution of cancer.

To address these needs, we introduce SciClone, a method for estimating the number and content of subclones across one or many samples. It focuses primarily on variants in copy-number neutral, loss of heterozygosity (LOH)-free portions of the genome, which allows for the highest-confidence quantification of variant allele frequencies (VAF) and inference of clonality. SNVs of sufficient depth may be provided by whole-exome sequencing (WES) or first discovered by whole-genome sequencing (WGS) and subsequently deep sequenced in a targeted fashion. The approach is not limited to SNVs, but is amenable to any event that can be described as a frequency. In particular, we demonstrate the integration of copy number events and discuss how copy-altered VAFs could be accommodated.

Computational efficiency is achieved by clustering the VAFs using a variational Bayesian mixture model [31] (VBMM), which differs substantially from the Dirichlet process models previously used to infer subclones. [6, 10, 11, 28] VBMMs similarly automatically infer the number of clusters and provide a probabilistic interpretation of clustering, but their deterministic nature allows them to scale to high dimension, where they enjoy efficiency advantages [31] over stochastic MCMC techniques employed by existing clonality detection methods. [6,10,11,28] Further, the variational Bayesian approach provides a computational termination condition more straightforward than monitoring techniques [32] required of MCMC. Though VBMMs are heuristic, and their approximations occasionally result in sub-optimal solutions, [33] we demonstrate their effectiveness here through simulation and application to several real tumor data sets. In particular, SciClone advances our preliminary [14] variational Bayesian beta mixture modeling approach for clustering VAFs in a single sample by: (1) applying the standard technique of factorizing the density over samples [34,35] to extend applicability to an arbitrary number of samples, (2) replacing our previous *ad hoc* notion of cluster overlap with a quantitative measure, [36, 37] (3) leveraging the probabilistic nature of VBMMs to quantify a variant's likelihood of belonging to a cluster via a $p$-value, and (4) incorporating a newly-derived variational Bayesian binomial mixture model as well as a published [31, 38] variational Bayesian Gaussian mixture model.

We demonstrate SciClone by inferring low-frequency subclones from a single MM sample and by quantitatively assessing the clonality of driver mutations in (potentially noisy) WES-derived data. We extend this approach to accommodate multiple samples and apply it to track clonal evolution of an acute myeloid leukemia (AML) tumor to relapse in response to therapy. As a further example of SciClone's scalability and utility in correlating mutations across samples, we examine spatial heterogeneity and aromatase-inhibitor resistance within three samples from a single breast cancer patient. In both the AML and breast cancer data sets, our analysis reveals subclones present in the primary tumor but not discernible from a single primary tumor sample. This reinforces the necessity of analyzing multiple patient-derived tumor samples to elucidate the full complexity of a cancer's heterogeneity.

The SciClone package is available at http://github.com/genome/sciclone.

# Results

## Mixture modeling objectively identifies (low-frequency) subclones.

Many tumors are highly heterogeneous and visualizations of somatic VAFs reveal high-density aggregations that correspond to specific subpopulations of cells (Fig. 1). To test our ability to detect and segregate these clusters, we used 2,018 validated, deep-sequenced (median depth 188x), genome-wide somatic SNVs from a primary MM tumor (M.H.T., et al., in preparation). These formed a high density region near 50% VAF, as expected of heterozygous SNVs in the founding clone of a nearly pure tumor sample (Figs. 1a and c). The actual median VAF of this founding clone is slightly less (near 46%), reflecting a small amount of normal cell admixture. Lower VAFs correspond to subclone-specific mutations that arose later in the tumor's expansion; a cluster of such VAFs thus represents a subclonal population, whose cells contain all of the founding clone mutations, as well as these subclonal mutations.

This tumor is hyperdiploid with characteristic trisomies of chromosomes 3, 5, 7, 9, 11, 15, 19, and 21. Founding clone mutations in these (and other) copy number altered regions are shifted in a predictable pattern, with a doubling of VAFs in regions of single-copy deletion where only the variant allele remains (Fig. 1b). In regions of single-copy amplification, mutations group near the expected frequency of ~31% where the wild-type allele is amplified or ~62% when the mutant allele is amplified (Fig. 1d). In this patient, these wild-type amplified SNVs occur at similar VAFs to subclonal events that are copy number neutral (red circles in Fig 1c). Disambiguating the two would require inference of allele- (and subclone-)specific copy number profiles, with its attendant uncertainty. To confirm that the two cases are in fact distinct and not an artifact of inaccurate copy number calls, we verified that none of the subclonal event VAFs in *putative* copy-number neutral regions reside instead on trisomic chromosomes. Further, they are not restricted to one or a few chromosomes, whose amplification or deletion might not have been detected, thus leading to apparent subclonal events.

To obtain a more objective view of this sample's clonal architecture, we identified low-frequency subclones by clustering the VAFs from copy number neutral, LOH-free, non-repetitive regions of the genome using SciClone, which uses an approach based on variational Bayesian beta mixture modeling [34, 35, 39]. The method automatically infers the optimal number of clusters, based on an initial overestimation of their expected number (ten, unless specified otherwise). In this MM sample, SciClone detected three clusters (Fig. 1): the first with average VAF of 46.1%, and two lower-frequency clusters with average VAFs of 32.0% and 11.9% (Fig. 1). Each cluster is represented by a posterior predictive density (see Materials and Methods), which provides the probability of a VAF given the observed data (and subsequent model fit). These densities probabilistically define boundaries between clusters, including the visually ambiguous separation between the highest-frequency cluster and the cluster with average VAF of 32.0%.

As a comparison for our results, we applied an MCMC method, PyClone [11], to the same variants in copy-number neutral, LOH-free regions. PyClone recapitulated the presence of the minor clusters, though with increased computational demands (Table S1). We tested PyClone's ability to integrate copy-number altered SNVs expected to be in the founding clone and found that it often assigned these sites to independent clusters.

As a second point of comparison, we applied THetA [29], a method that infers clonal architecture based on copy number events alone. For reasons of computational complexity, we applied it to a limited number of segments with representative copy number states. THetA detected clonal amplifications that occurred in 89.8% of the cells (44.9% VAF) and a subclonal deletion that occurred in 54.1% of the cells (27.0% VAF). We integrated these data and SciClone clustering of all CNA and SNV VAFs revealed that the THetA-inferred CNA VAFs support the presence of subclones originally inferred from SNVs alone (Fig. S1).

In some samples, ordering of subclonal VAFs may reveal the clonal phylogeny of the tumor. [40] However, in this sample, the data are insufficient to distinguish between a branched phylogeny, in which the two subclones arose from independent cells within the founding clone, and a linear phylogeny, where the lower VAF subclone is descended from the higher VAF subclone. The latter case implies that all mutations in the higher VAF clone are also present in the lower VAF clone, as are all founding clone mutations.

## Bayesian modeling quantifies the (un)certainty of mutation clonality.

Using WES to both discover variants and obtain deep read counts for defining VAFs may be an attractive, direct approach to clonality analysis, [10] as it avoids the additional time and expense of WGS followed by targeted sequencing. However, while WES data captures the coding variants likely driving disease progression, their number may be insufficient to reliably infer clonal architecture, particularly for cancers with relatively low somatic mutation rates. To begin to address these considerations, we applied SciClone to whole-exome sequenced breast [41] and endometrioid carcinoma [42] cancer samples from The Cancer Genome Atlas (TCGA) project. To obtain a sufficient number of variants, we relaxed the minimum depth of coverage requirement to 50x, resulting in 29 copy-number neutral variants from the breast sample and 53 from the endometrial sample. The breast tumor has a high-VAF cluster corresponding to its founding clone as well as a subclonal cluster, with most variants occurring in the latter (Fig. 2a). The endometrial sample is more complex, with both a high-VAF cluster and three tightly-grouped and poorly-separated subclonal clusters (Fig. 2b).

Drawing inferences about mutation clonality (e.g., assessing whether mutations generally occur in the founding clone and hence are likely to be early, disease-initiating events [14] or attempting to correlate subclonal mutations with clinical outcome [10]) requires accurately and confidently assigning individual VAFs to clusters. Our variational Bayesian approach does so via "fuzzy" cluster assignments, which describe the (conditional, posterior) probability that a VAF belongs to a particular cluster (given that it belongs to one of them). In particular, the likely driver *PIK3CA* mutations in the endometrial sample are assigned confidently to the highest-frequency cluster one, with probabilities of 93.1% for the lower VAF variant and 97.1% for the higher. In contrast, the potential driver *ATM* mutation is nearly as likely to belong to cluster one (42.1% probability) as to the lower VAF cluster two (57.8% probability) to which it was "assigned" (i.e., that which maximized its posterior probability). Given the relatively few SNVs, this ambiguous assignment indicates that the data are insufficient to accurately define the clonal structure and that the separation between cluster one and cluster two may be an artifact of sparse data. This uncertainty might be resolved by increased depth of sequencing or by additional clonal markers (e.g., as discovered by WGS). Nevertheless, the strong assignments of the *PIK3CA* mutations to a cluster with average VAF near 50% suggest that, despite the relatively high level of noise in the data, they belong to the founding clone.

## Longitudinal studies refine subclonal architecture and reveal mechanisms of resistance.

Tumors evolve in response to treatment, both through loss of specific mutations and acquisition of new ones. Understanding this process in the context of a tumor's clonal architecture is critically important in defining mechanisms of resistance and in informing therapeutic decisions. To better understand mechanisms of therapy resistance, we extended our method to accommodate multiple samples and applied it (Fig. 3) to samples from a primary AML tumor and post-treatment relapse occurring 26 months after chemotherapy. [5] These primary and relapse tumors were initially sequenced to depths of 26.7x and 31.5x, respectively, with subsequent capture validation providing deep read counts for all discovered variants (median depth: 753x). All variants were analyzed, as no CNAs are present in either sample.

Analysis of the primary tumor sample in isolation (Fig. 3c) suggests a simple organization consisting of a single subclone and a founding clone containing an *IDH2* R140L mutation. Mutations in this residue may play a role in oncogenesis, given their recurrence in AML [43] and resulting neomorphic enzymatic activity. [44] Hence, this clonal mutation is an attractive target for small molecule inhibitors, such as those reactive against *IDH2* R140Q. [45] However, simultaneous analysis of the relapse genome further dissects the apparently homogeneous highest-frequency cluster harboring *IDH2* R140L into three distinct subpopulations of cells (Fig. 3a): one that is effectively eliminated by chemotherapy (cluster three, average relapse VAF $< 0.5\%$), a second diminished by treatment (cluster two, average relapse VAF 11.6%), and a third largely unaffected by treatment (cluster one, average relapse VAF 41.3%). As further evidence of their high degree of overlap in the tumor sample, their respective average VAFs in the tumor are 42.7%, 43.1%, and 44.9%. The additional resolution provided by the relapse sample distinguishes these subpopulations to expose a more complex clonal architecture (Fig. 3d) and indicates that the *IDH2* R140L mutation in cluster two is subclonal. Thus, targeting it therapeutically would be unlikely to eradicate the founding clone. We do observe that the subclonal mutations in cluster five were eliminated by treatment, suggesting that it carried a lower proclivity for resistance than the surviving clones.

Remarkably, there is a second, independent *IDH2* mutation (R140W) in the relapse sample. But, as above, defining its clonality from this sample alone (Fig. 3b) is confounded by an inability to associate its VAF (32.8%) unambiguously with either the founding clone or a subclone. This uncertainty is resolved through multidimensional analysis that incorporates the tumor sample and places the mutation in cluster four. Mutations within this cluster, including *IDH2* (R140W), were either present in the primary tumor below the level of detection or are new mutations, possibly induced by cytotoxic chemotherapy. [5] In either case, they are potential drivers of disease progression.

Given the clonal complexity of this sample, we next asked how many variants were required to capture this complexity and whether we were likely to have missed additional complexity. To address these concerns, we randomly selected a subset of the original variants and performed clustering. The number of clusters inferred as a function of the number of variants analyzed is fairly constant for $> 200$ variants (Fig. 4a), whereas it drops precipitously for $< 100$ variants. As sequencing detected a total number of variants within the flat regime of this curve, we can be confident that no subclones with a higher VAF than the most infrequent cluster identified (average VAF $\sim$12%) were missed. Further, this suggests that $\sim$200 variants would have been sufficient to reveal this sample's clonal architecture. To assess the sensitivity of our approach in inferring the separation of clusters, we performed one-dimensional analysis of cluster one and two (Fig. 3) relapse sample VAFs after varying their inter-cluster separation (Fig. 4b). While the results are sample-dependent, they indicate that clusters can be reliably distinguished if they lie greater than $\sim 7\%$ away from one another.

To ensure that the inferred number and composition of clones were not overly sensitive to our computational method, we varied both the number of initial clusters and the clustering approach itself. Consensus clustering indicated that the (subjectively) correct number of clusters (five) was inferred by the variational Bayesian beta mixture modeling for a range of initial number of clusters from six to 15 (data not shown). We next used SciClone to cluster using a variational Bayesian binomial mixture model and a previously-published [31, 38] variational Bayesian Gaussian mixture model (see Supporting Text). Consensus clustering indicates that the results are stable for the majority of variants as both the number of initial clusters and the method (beta, binomial, or Gaussian) are varied (Fig. 4c). The few variants that clustered differently between methods (Fig. S2) were situated near cluster boundaries or between clusters. A similar effect was seen when clustering the data with PyClone (Table S1), though in this case variants along cluster boundaries tended to coalesce into independent clusters (clusters six and eight in Fig. S3): PyClone's default hyperparameter settings lead it to overdissect the founding clone. After increasing the number of iterations from 10,000 (with a burn-in of 1,000 iterations) to 100,000 iterations (with a burn-in of 10,000 iterations), PyClone results were even more similar to those obtained with SciClone, but the former still split the highest-VAF cluster into two (data not shown). Despite these differences, the results are largely consistent between SciClone and PyClone and we have increased confidence in variants that are similarly assigned by both approaches.

## Multiple biopsies reveal intratumoral heterogeneity and impact of treatment.

Spatial heterogeneity complicates the analysis of solid tumors, as distinct regions of a tumor may harbor different subclonal populations. [9, 13, 16] Assaying multiple regions of heterogeneous tumors should assist in uncovering the full spectrum of mutations and subclones present in a tumor and help identify the spatial origins of subclones that give rise to therapy resistance. To investigate this effect, we analyzed two pre-treatment biopsies from the same breast tumor and added a temporal dimension by examining a single sample from the tumor collected 16 weeks after aromatase-inhibitor (AI) therapy. Mutations in the three samples had median coverage of 130.5x from deep capture sequencing.

Three-dimensional clustering with SciClone revealed five distinct groups of mutations and a fairly low purity, resulting in reduced VAFs in all samples (Fig. 5, Video S1). Differences between the two pre-treatment biopsies were captured in clusters four and five, containing region-specific mutations. Cluster two cannot be identified from pre-treatment tumor 1 alone, but the second biopsy reveals it as a distinct subpopulation of cells with higher VAF in the first biopsy (36.03% vs. 8.13%). The effect of AI therapy is revealed by inclusion of the post- treatment sample, in which clusters two and four are eliminated. These likely represent AI-responsive subpopulations of cells, though additional spatial heterogeneity leading to their apparent removal cannot be discounted. Cluster five contains mutations specific to the second biopsy; while some of the cells harboring them expanded in the post-treatment sample, others appear to have been eradicated completely. The heterogeneity in response observed in this cluster suggests that it actually encapsulates several distinct, but overlapping subclonal populations that occur at similar VAFs in the pre-treatment biopsy and are difficult to separate without additional data.

Application of PyClone to these data (Table S1, Fig. S4) reveals several significant differences. While it infers two distinct clusters from the heterogeneous cluster five, it also partitions variants in the founding clone into two clusters. This separation is likely a clustering artifact, since (1) the two clusters are merged when all of the data (in copy-altered and -neutral regions) are clustered using 34,000 iterations (data not shown) and (2) the presence of two large, *independent* clusters comprising ~70% of the cellular population each is biologically unreasonable. The discordance between methods suggests that the limited number of variants affected require special attention.

## Discussion

Clonal heterogeneity complicates both our understanding of the biology of tumors and the design of effective treatment strategies. While individual tumor samples provide a glimpse of this complexity, additional temporally or spatially distinct samples allow higher resolution mapping of subclonal architecture, including the isolation of drug-sensitive clones and small subpopulations driving relapse. To leverage the increasingly commonplace and cost-effective opportunities to sequence multiple samples from an individual, we developed SciClone, which scalably analyzes large numbers of samples to provide an unbiased, probabilistic dissection of a cancer's clonal landscape. To do so, SciClone employs variational Bayesian mixture modeling of beta, binomial, and Gaussian distributions. Each of these may have advantages (see Supporting Text) in certain situations, though our tests suggest that the beta mixture model works best in practice. We have previously used related techniques in analyzing FACS data [46] and expect them to be of general interest to those requiring methods that automatically and efficiently infer the number of clusters from high-dimensional biological data.

Application of SciClone to primary and relapse AML tumors identified subclonal populations with dramatically divergent response to conventional therapy. Such analyses are the first step towards inferring driver mutations responsible for both resistance to therapy and clonal expansion following treatment. Insight into the spatial origins of treatment response was provided by analysis of three samples from a breast tumor, two of which were obtained from distinct regions of a single tumor at the same time point.

The AML and breast cancer cases highlight an inherent limitation of bulk sequencing of tumor cells: subclonal populations cannot be distinguished if they occur at similar frequencies. Single-cell sequencing may eventually offer a solution, but will require dramatic improvements in fidelity and throughput. Using currently available data, we demonstrated that temporally or spatially distinct samples from the same tumor can be used to tease apart these overlapping subclones. This is demonstrated in AML, where the apparent founding clone in the primary tumor is dissected into two additional subclones by incorporating the relapse sample. The breast cancer samples exhibit two-fold complexity. As in the AML primary tumor, the apparent founding clone in one of the pre-treatment breast tumors is revealed to consist of an additional subclone by the second spatially-isolated biopsy. Additionally, each pre-treatment sample exhibits a clone not detected in the other. This suggests that manipulation of the patient's tumor-derived cells (e.g., passage within culture or as mouse xenografts) may be a viable method for identifying additional subclones and predicting those with differential responses to therapy.

Our analysis of exome-sequenced cases showed that SciClone can be useful on samples with as few as 29 SNVs (Fig. 2a), but our simulations (Fig. 4a) showed that in more complex cases, such as the AML tumor/relapse pair, establishing subclonal boundaries may require two hundred or more variants and that subclones be separated by VAFs of ~7% or more (Fig. 4b). This downsampling approach may be applied to any data set to establish a baseline sensitivity. Cases with poorly defined cluster boundaries (e.g., due to a paucity of mutations), such as the endometrial case (Fig. 2b), benefit from SciClone's probabilistic formalism. In particular, by assigning an *ATM* mutation similar probability of belonging to the founding clone and a subclone, SciClone reflected the lack of certainty inherent in the data and indicated that their sparsity may poorly characterize the tumor's clonal diversity. The sensitivity of any clustering method in dissecting clonal boundaries is dependent on cluster overlap, which we have characterized via the "uncertainty" of their probabilistic assignments (Refs.36 and37 and Materials and Methods). An additional, qualitative means of detecting high-confidence variant/cluster assignments involves taking the consensus, or intersection, across clustering methods (Fig. 4c). Confidence in detecting all major subclones increases with the number of variants, including passenger mutations more likely to be missed by exome sequencing. Thus, WGS followed by deep validation sequencing is most likely to capture the full spectrum of mutations and yield high-quality characterization of subclonal entities.

Next-generation sequencing of variants within copy-number neutral regions of autosomal chromosomes leads to a straightforward interpretation of the inferred VAFs as half the cellular frequency harboring the corresponding variant. Because of the widespread availability of variants to serve as clonal markers and the relative reliability of their bioinformatic analysis and quantification, our initial clonality analyses have focused on SNVs. Nevertheless, other genomic events have been used to identify clonal dynamics. For example, the alternate "waxing" and "waning" of subclonal CNAs has been observed in multiple myeloma [17, 30]. However, the analysis and discovery of CNAs pose several challenges for clonality: (1) Cancer types may be described hierarchically in terms of their propensity to elicit either mutations or copy number changes [47]. For mutation-dominated cancer types, such as the cytogenetically-normal AML analyzed here, few CNA events may be available for analysis. The converse does not apply: since SNVs accumulate with age [4], an abundance of SNV clonal markers are expected in all malignant, as well as in non-malignant, tissues. Given the density of SNVs, clonality analyses that rely solely on them may well capture the full clonal architecture, while missing specific (copy number) events of pathogenic interest; clonality analyses relying solely on copy number events are likely to miss both. (2) There is no digital readout of CNAs, rather observed copy number reflects the admixture of subclonal populations and is a (linear) combination of the copy number state of each subclone, weighted by the fractional subpopulation of the subclone. In principle, the correctness of the analysis requires the simultaneous inference of this admixture and the number of copies (0, 1, 2, 3, or greater) of each chromosomal segment in each subclone. Though such an analysis would infer the clonal hierarchy directly, rather than the clusters of variants that serve to identify them as in a SNV-based analysis, inference in the latter case is simplified since there are fewer mutational states (presence or absence, at least of the vast majority of variants, which are heterozygous) and the correctness of inferring one cluster is independent of a second cluster.

For these reasons, we prefer to overlay CNA events on the higher confidence copy-number neutral SNV VAFs. Incorporating such events may be done: (1) by determining the fractional population harboring the event (as in Fig. S1) or (2) by adjusting a SNV's VAF based on it's inferred copy number states across subclonal populations. One approach to the latter involves inferring copy number states from the B-allele frequencies of germline SNPs (e.g., using ASCAT [48] or APOLLOH [49]) and phasing these to somatic variants (e.g., by detecting a single sequencing read spanning both) to impute subclone-specific copy numbers to each variant [6]. After adjusting the SNV VAFs, they could be clustered by SciClone in a manner completely analogous to the analysis of unadjusted VAFs (using the beta or Gaussian mixture model approaches). We are currently pursuing this approach.

MCMC techniques, such as PyClone [11], offer an alternative approach to clustering variants. However, our comparisons of SciClone and PyClone (Table S1) reinforce the computational inefficiencies of MCMC approaches relative to variational Bayesian techniques [31] and show that Sciclone is between one and two orders of magnitude faster. SciClone inherits the simple variational Bayesian (computational) convergence condition of monitoring monotonic changes in a lower bound (see Supporting Text). While this approach may converge to a local extremum, more subtlety is required to ensure the (theoretical) asymptotic convergence to the global extremum guaranteed by MCMC, e.g., monitoring variance within a Markov chain relative to variance between independent Markov chains [32]. PyClone provides no direct facilities to monitor convergence. Regardless, the theoretical convergence properties of MCMC seem unjustified given the involved computational overhead for a clustering application, such as clonality detection, where error estimates of the parameters are of marginal interest.

SciClone has already contributed to the understanding of biological mechanisms underlying cancer and has the potential for increased utility with the advent of clinical sequencing. Towards this end, we are developing methods that cross-reference the clonal status of specific mutations with databases of targeted therapeutics. As an example, the Drug-Gene Interaction Database [50] identifies three genes in the AML sample as as potentially druggable: (*DRD2*, *KCNQ2*, and *P2RY2*). The fact that each of these mutations lies in a subclone complicates their interpretation, and suggests that careful study is needed to understand how specific subclonal populations respond to different therapeutics. While clinical decisions of which (sub)clones to target and how remain controversial, it is clear that making these decisions will require accurate assessment of clonal architecture using tools such as SciClone.

# Materials and Methods

## Variational Bayesian mixture modeling of beta distributions.

A VAF $f$ is defined with respect to the number of reads, $x^{\mathrm{var}}$, supporting the variant allele and the number of reads, $x^{\mathrm{ref}}$, supporting the reference (or non-major-variant, in the case of multiple variants) allele: $f = \frac{x^{\mathrm{var}}}{x^{\mathrm{var}}+x^{\mathrm{ref}}}$. Our previous method [14] considered variants in a *single* sample and modeled the probability of a VAF $f$ belonging to cluster $k$ as

$$\mathrm{Beta}(f;u_k,v_k) = \frac{\Gamma(u_k+v_k)}{\Gamma(u_k)\Gamma(v_k)} f^{u_k-1}\left(1-f\right)^{v_k-1}$$

where $\Gamma(\cdot)$ is the gamma function. Here, we extend this to the case with $S \geq 1$ samples by defining the $S$-vector $\mathbf{f} \equiv (f_1, f_2, \ldots, f_S)$, whose $s^{\mathrm{th}}$ component, $f_s$, is the VAF of that variant in the $s^{\mathrm{th}}$ sample. We make the assumption that, *within a cluster*, the VAFs are independent across samples, so that the cluster may simply be modeled as

$$p(\mathbf{f}|\mathbf{u}_k,\mathbf{v}_k) = \mathbf{Beta}(\mathbf{f};\mathbf{u}_k,\mathbf{v}_k) = \prod_{s=1}^{S} \mathrm{Beta}(f_s;u_{ks},v_{ks}) \ \ , \tag{1}$$

where $\mathbf{u}_k$ and $\mathbf{v}_k$ are the $S$-vectors whose $s^{\mathrm{th}}$ components are $u_{ks}$ and $v_{ks}$, respectively. This implies only that *within* a cluster there is no correlation between samples. The validity of this assumption is indicated by the visually-observed orthogonality of the VAF principal component axes to the coordinate (i.e., sample) axes. We have rarely, if ever, seen evidence for such intra-cluster correlation. Nevertheless, this assumption may be relaxed through use of a mixture of multivariate Gaussian distributions (see Supporting Text), each of which has a full-rank covariance matrix.

In considering a mixture of $K$ (multi-dimensional) beta components (Eq. (1)), we introduce a $K$-dimensional latent (or unobserved) binary random variable $\mathbf{z}_n$ indicating whether VAF $\mathbf{f}_n$ does ($z_{nk} = 1$) or does not ($z_{nk} = 0$) belong to component $k$ and satisfying a 1-of-$K$ representation in which $\sum_{k=1}^{K} z_{nk} = 1$. The marginal probability $p(z_{nk} = 1)$ that a VAF belongs to component $k$ is given by its mixing coefficient $\pi_k$,

$$p(z_{nk} = 1) = \pi_k \ ,$$

subject to the probabilistic constraints

$$0 \leq \pi_k \leq 1 \ ,$$
$$\sum_{k=1}^{K} \pi_k = 1 \ .$$

Given the 1-of-$K$ representation of $\mathbf{z}_n$, this may be written as

$$p(\mathbf{z}_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{z_{nk}} \ . \tag{2}$$

Similarly, the conditional distribution $p(\mathbf{f}_n|\mathbf{z}_n, \mathcal{U}, \mathcal{V})$ that a VAF $\mathbf{f}_n$ arises from the mixture may be written

$$p(\mathbf{f}_n|\mathbf{z}_n, \mathcal{U}, \mathcal{V}) = \prod_{k=1}^{K} \mathbf{Beta}(\mathbf{f}_n;\mathbf{u}_k,\mathbf{v}_k)^{z_{nk}} \tag{3}$$

in terms of the shape parameter vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ of the $k^{\text{th}}$ beta component, with aggregate parameters $\mathcal{U} \equiv \{\mathbf{u}_k\}$ and $\mathcal{V} \equiv \{\mathbf{u}_k\}$.

To accomodate binomial and Gaussian mixture models in addition to the beta mixture model, we introduce abstract notation used below to define quantities (e.g., $p$-values) independently of the concrete representation of likelihoods and posterior distributions. We begin by defining abstract parameters $\Phi$, which differ according to the model, i.e., beta, binomial, or Gaussian. For example, $p(\mathbf{f}_n|\mathbf{z}_n, \Phi^{\text{beta}}) \equiv p(\mathbf{f}_n|\mathbf{z}_n, \mathcal{U}, \mathcal{V})$, with $\Phi^{\text{beta}} \equiv \{\mathcal{U}, \mathcal{V}\}$. Further, while the Gaussian mixture model is also a function of VAFs $\mathbf{f}$, the binomial mixture model is defined with respect to the variant and reference count vectors, $\mathbf{x}^{\text{var}}$ and $\mathbf{x}^{\text{ref}}$, respectively. To abstract away these details, we use the notation $\boldsymbol{\chi}$ to denote the VAFs $\mathbf{f}$ of a beta or Gaussian mixture model or the counts $\mathbf{x}^{\text{var}}$ and $\mathbf{x}^{\text{ref}}$ of a binomial mixture model, when convenient. In particular, $p(\boldsymbol{\chi}_n|\mathbf{z}_n, \Phi^{\text{beta}}) \equiv p(\mathbf{f}_n|\mathbf{z}_n, \Phi^{\text{beta}}) \equiv p(\mathbf{f}_n|\mathbf{z}_n, \mathcal{U}, \mathcal{V})$, while $p(\boldsymbol{\chi}_n|\Phi_k^{\text{beta}}) \equiv p(\mathbf{f}_n|\mathbf{u}_k, \mathbf{v}_k)$, with $\Phi_k^{\text{beta}} \equiv \{\mathbf{u}_k, \mathbf{v}_k\}$.

Eqns. (2) and (3) extended across the entire set $\mathcal{F} \equiv \{\mathbf{f}_n\}$ of VAFs (or, more abstractly, $X \equiv \{\boldsymbol{\chi}_n\}$ of data) and their associated latent variables $\mathcal{Z} \equiv \{\mathbf{z}_n\}$ are combined to express the complete-data (i.e., including the latent variables) likelihood

$$p(X, \mathcal{Z}|\boldsymbol{\pi}, \Phi^{\text{beta}}) \equiv (\mathcal{F}, \mathcal{Z}|\boldsymbol{\pi}, \mathcal{U}, \mathcal{V}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathbf{Beta}(\mathbf{f}_n|\mathbf{u}_k, \mathbf{v}_k)\right]^{z_{nk}} \ , \tag{4}$$

which may be summed over $\mathbf{z}_n$ to give the incomplete likelihood

$$p(\mathcal{F}|\boldsymbol{\pi}, \mathcal{U}, \mathcal{V}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathbf{Beta}(\mathbf{f}_n|\mathbf{u}_k, \mathbf{v}_k) \ .$$

These equations could be used to fit the beta parameters using expectation maximization (EM) or Markov chain Monte Carlo (MCMC) techniques.

We instead use a previously described [34,35] variational Bayesian approach [31,38] to modeling a mixture of beta distributions. Self-contained presentations of the general theory of variational Bayesian inference and its application to Gaussian mixture models [31, 38] and to binomial mixture models are provided in the Supporting Text. Here, we summarize its application to beta mixture models to provide sufficient context for its use in and extension for clonality analysis. For full details of this more involved derivation, the reader is referred to the original references. [34, 35]

Variational Bayesian beta mixture modeling approximates the posterior distribution, $p(\mathcal{Z}, \boldsymbol{\pi}, \Phi^{\text{beta}}|X) \equiv p(\mathcal{Z}, \boldsymbol{\pi}, \mathcal{U}, \mathcal{V}|\mathcal{F})$, over the model parameters $\boldsymbol{\pi}$, $\mathcal{U}$, and $\mathcal{V}$ and the latent variables $\mathcal{Z}$ with a distribution $q(\mathcal{Z}, \boldsymbol{\pi}, \Phi^{\text{beta}}) \equiv q(\mathcal{Z}, \boldsymbol{\pi}, \mathcal{U}, \mathcal{V})$. The form of this approximate distribution is a consequence of choice of prior distribution, whose product with the likelihood [Eq. (4)] defines the posterior $p(\mathcal{Z}, \boldsymbol{\pi}, \mathcal{U}, \mathcal{V}|\mathcal{F})$ according to Bayes' theorem, and of the mild and standard [31] assumption that the latent variables $\mathcal{Z}$ factorize from the model parameters, i.e., $q(\mathcal{Z}, \boldsymbol{\pi}, \mathcal{U}, \mathcal{V}) = q(\mathcal{Z})q(\boldsymbol{\pi}, \mathcal{U}, \mathcal{V})$. This further simplifies, without assumption, to $q(\mathcal{Z}, \boldsymbol{\pi}, \Phi^{\text{beta}}) \equiv q(\mathcal{Z}, \boldsymbol{\pi}, \mathcal{U}, \mathcal{V}) = q(\mathcal{Z})q(\boldsymbol{\pi})q(\mathcal{U}, \mathcal{V}) \equiv q(\mathcal{Z})q(\boldsymbol{\pi})q(\Phi^{\text{beta}})$. Finally, the authors assume the $\mathcal{U}$ and $\mathcal{V}$ variables are independent and factorize to ultimately give $q(\mathcal{Z}, \boldsymbol{\pi}, \mathcal{U}, \mathcal{V}) = q(\mathcal{Z})q(\boldsymbol{\pi}) \prod_{k,s} q(u_{ks})q(v_{ks})$.

Ma and Leijon used four synthetic one-dimensional data sets (Fig. 4 of Ref. 34), including two with highly overlapping beta distributions, to demonstrate the high accuracy of this method despite its assumption that the parameters of the beta distribution are independent. Fan et al. [35] additionally analyzed six three-dimensional data sets and similarly found that accuracy was not adversely effected by this factorization approximation (Table I of Ref. 35). Our own extensive simulation results further support these findings. We generated data sets by sampling a mixture of beta distributions in one, two, or three dimensions and having between two and five clusters (100 data sets for each dimensionality/number of clusters pair). Fig. S5 shows the concordance (i.e., fraction of correctly assigned items) between the clustered and known results for each simulated data set. The average concordance is $0.861$, $0.985$, and $0.999$ in one, two, and three dimensions, respectively. Performance increases with dimensionality as the clusters become increasingly separated. This sparsity may be quantified by the minimum cluster self-overlap (see below). Data sets having a relatively small minimum cluster self-overlap have a relatively large overlap *between* clusters, which leads to uncertainty and degrading performance.

The prior distributions are generally selected to be conjugate to the likelihood for analytic convenience (e.g., see the derivations of the variational Bayesian Gaussian and binomial mixture models in Supporting Text). While a conjugate prior to the beta likelihood is formally available, its use would lead to an analytically intractable integration. [34] Therefore, Ma and Leijon [34] instead propose use of the following prior distribution

$$\begin{aligned} p(u_{ks}) &= \mathcal{Gam}(u_{ks}; \mu^0_{ks}, \alpha^0_{ks}) \\ p(v_{ks}) &= \mathcal{Gam}(v_{ks}; \nu^0_{ks}, \beta^0_{ks}) \\ p(\boldsymbol{\pi}) &= \mathcal{D}(\boldsymbol{\pi}; \boldsymbol{c}^0) \end{aligned} \tag{5}$$

where $\mathcal{Gam}(u_{ks}; \mu^0_{ks}, \alpha^0_{ks})$ and $\mathcal{Gam}(v_{ks}; \nu^0_{ks}, \beta^0_{ks})$ are gamma distributions

$$\mathcal{Gam}(u; \mu, \alpha) = \frac{\alpha^\mu}{\Gamma(\mu)} u^{\mu-1} e^{-\alpha u} \quad \mu, \alpha \in \mathbb{R}^+ \ ,$$

and $\mathcal{D}(\boldsymbol{\pi}; \boldsymbol{c}^0)$ is the Dirichlet distribution

$$\mathcal{D}(\boldsymbol{\pi}; \boldsymbol{c}) = C(\boldsymbol{c}) \prod_{k=1}^{K} {\pi_k}^{c_k - 1} \tag{6}$$

over proportions $\boldsymbol{\pi}$, with the normalizing constant $C(\boldsymbol{c})$

$$C(\boldsymbol{c}) = \frac{\Gamma(\hat{c})}{\prod_k \Gamma(c_k)}$$

and

$$\hat{c} \equiv \sum_k c_k \ .$$

The parameters of the approximate posterior distribution are now determined by iteratively minimizing the Kullback-Leibler divergence, a measure of the difference, between it and the posterior distribution, following the general prescription of variational Bayesian inference (see Supporting Text). The authors make a non-linear approximation to an expectation value arising during the iterative procedure so that the resulting, approximate posterior distribution has the form of a gamma distribution, despite the fact that the above gamma prior distribution is not conjugate to the beta likelihood. Significantly, the authors show that this additional approximation can be used to minimize the original, desired Kullback-Leibler divergence between the posterior distribution and the approximate, non-gamma posterior distribution. This results in the approximate posterior distribution

$$q(\boldsymbol{\pi}, \Phi^{\text{beta}}) \equiv q(\boldsymbol{\pi}) \prod_{k=1}^{K} \prod_{s=1}^{S} q(u_{ks}) q(v_{ks}) = \mathcal{D}(\boldsymbol{\pi}; \boldsymbol{c}) \prod_{k=1}^{K} \prod_{s=1}^{S} \mathcal{Gam}(u_{ks}; \mu_{ks}, \alpha_{ks}) \mathcal{Gam}(v_{ks}; \nu_{ks}, \beta_{ks}) \ , \tag{7}$$

where $c_k$, $\mu_{ks}$, $\alpha_{ks}$, $\nu_{ks}$, and $\beta_{ks}$ are defined in Eqns. 47-51, respectively, of Ref. 34. These parameters are updated from the corresponding initial hyperparameter values $c_k^0$, $\mu_{ks}^0$, $\alpha_{ks}^0$, $\nu_{ks}^0$, and $\beta_{ks}^0$ as in a traditional EM iterative algorithm. It will also be convenient to define the posterior density with respect to the $k^{\mathrm{th}}$ component

$$q(\Phi_k^{\mathrm{beta}}) \equiv \prod_s q(u_{ks})q(v_{ks}) \; . \tag{8}$$

## Probabilistic and hard cluster assignments

Variational Bayesian mixture models provide probabilistic assignments of variant $\boldsymbol{\chi}_n$ (i.e., a VAF $\mathbf{f}_n$ for beta or Gaussian mixture models or variant counts $\mathbf{x}_n^{\mathrm{var}}$ for a binomial mixture model) to cluster $k$ according to the posterior probabilities $p(z_{nk} = 1|\boldsymbol{\chi}_n) \equiv r_{nk}$. The $r_{nk}$ act as "responsibilities" and satisfy $\sum_k r_{nk} = 1$. In the case of the beta mixture model, the $r_{nk}$ are defined by Eqns. 31 and 32 of Ref. 34. A more general derivation is provided in the Supporting Text, along with specific calculations for binomial and Gaussian mixture models.

For visualization purposes, for example, we occasionally transform these probabilistic assignments into hard assignments, which assign $\boldsymbol{\chi}_n$ to one and only one cluster $k$ according to

$$k = \arg\max_{k'} p(z_{nk'} = 1|\boldsymbol{\chi}_n) \; .$$

## Posterior predictive density

The posterior predictive density gives the probability of a new (i.e., unobserved) variant, $\hat{\chi}$, given the observed data $X$

$$p(\hat{\chi}|X) = \sum_{\hat{\mathbf{z}}} \int p(\hat{\chi}|\hat{\mathbf{z}}, \Phi)p(\hat{\mathbf{z}}|\boldsymbol{\pi})p(\boldsymbol{\pi}, \Phi|X)d\boldsymbol{\pi}d\Phi$$

and all possible assignments $\hat{\mathbf{z}}$ of that variant to a cluster. Evaluating the sum over $\hat{\mathbf{z}}$, making use of Eq. (2), gives

$$p(\hat{\chi}|X) = \sum_k \int \pi_k p(\hat{\chi}|\Phi_k)p(\boldsymbol{\pi}, \Phi|X)d\boldsymbol{\pi}d\Phi \; .$$

We next approximate the true posterior distribution, $p(\boldsymbol{\pi}, \Phi|X)$, with the variational approximation $q(\boldsymbol{\pi}, \Phi)$ to give

$$p(\hat{\chi}|X) \approx \sum_k \int \pi_k p(\hat{\chi}|\Phi_k)q(\boldsymbol{\pi}, \Phi)d\boldsymbol{\pi}d\Phi \; .$$

Since for all mixture models considered in this manuscript $q(\boldsymbol{\pi}, \Phi) = q(\boldsymbol{\pi})q(\Phi)$, with $q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \boldsymbol{c})$ and

$$\mathrm{E}_{\boldsymbol{\pi}}[\pi_k] = \int \pi_k q(\boldsymbol{\pi})d\boldsymbol{\pi} = \frac{c_k}{\hat{c}} \; ,$$

this evaluates to

$$p(\hat{\chi}|X) \approx \sum_k \frac{c_k}{\hat{c}} \int p(\hat{\chi}|\Phi_k)q(\Phi)d\Phi \; . \tag{9}$$

Ma and Leijon [34] assumed that $q(\Phi) \approx \delta(\Phi - \Phi^*)$, where $\delta(\cdot)$ is the Dirac delta function and $\Phi^*$ are the converged parameter values, i.e., that the posterior distribution has negligible probability when any of the parameters differ from their converged values. In this case,

$$p(\hat{\chi}|X) \approx \sum_k \frac{c_k}{\hat{c}} p(\hat{\chi}|\Phi_k^*) \; ,$$

which may be efficiently evaluated. We instead use Eq. (9), which avoids any assumption on the approximate posterior distribution. In the case of binomial and Gaussian mixture models, Eq. (9) may be evaluated analytically. In the case of a beta mixture model, for which $p(\hat{\chi}|\Phi_k)$ is given by Eq. (1) and $q(\Phi)$ is given by Eq. (8), we instead resort to numerical integration, evaluating data sampled from Eq. (8) via Eq. (1).

## Prior initialization

We choose hyperparameters resulting in prior distributions sufficiently broad to ensure that the number of clusters and their posterior parameterization are determined primarily from the data rather than from prior assumptions. In particular, following Ma and Leijon [34], we choose $c_k^0 = 0.001$ for all $k$. We also choose $\alpha_{ks}^0 = \beta_{ks}^0 = 0.005$ and $\mu_{ks}^0 = \nu_{ks}^0 = 1$ for all $k$. Given the latter choice, the gamma distributions $\mathcal{Gam}(u_{ks}; \mu_{ks}^0 = 1, \alpha_{ks}^0)$ and $\mathcal{Gam}(v_{ks}; \nu_{ks}^0 = 1, \beta_{ks}^0)$ collapse to exponential distributions. The resulting variances of these distributions, e.g., $\mathrm{Var}[u_{ks}] = \left(1/\alpha_{ks}^0\right)^2$, are large given our choice of hyperparameters and hence provide a broad prior.

We initialize the $r_{nk}$ according to the hard assignments computed by $k$-means (provided in the R stats package and using default parameters, except with $nstart = 1000$ and $centers = 10$). We initialize the parameters $\mu_{ks}$, $\nu_{ks}$, and $\beta_{ks}$ to their respective hyperparameter values $\mu_{ks}^0$, $\nu_{ks}^0$, and $\beta_{ks}^0$. Finally, we initialize the $\alpha_{ks}$ such that the expected means of the cluster centers, $\bar{u}_{ks}/(\bar{u}_{ks} + \bar{v}_{ks})$, with $\bar{u}_{ks} = \mathrm{E}[u_{ks}] = \mu_{ks}^0/\alpha_{ks}^0$ and $\bar{v}_{ks} = \mathrm{E}[v_{ks}] = \nu_{ks}^0/\beta_{ks}^0$, are set to the values returned by $k$-means. We then perform the variational E step (i.e., calculate the expectations immediately following Eq. 51 of Ref. 34) *without* updating the $r_{nk}$, followed by the variational M step to update the parameters $\mu_{ks}$, $\nu_{ks}$, $\nu_{ks}$, $\beta_{ks}$, and $c_k$ (via Eqns. 47-51 of Ref. 34). For the AML28 data set, this initialization results in the clusters shown in Fig. S6A. Initialization is followed by iteratively applying the variational E step (including updating the $r_{nk}$) and M step. To avoid undefined behavior in evaluating the beta distribution, we shift VAFs at zero or one by $\delta$ or $-\delta$, respectively, with $\delta$ equal to machine precision.

## Cluster pruning and outlier detection

Variational Bayesian mixture modeling generally discards clusters that do not contribute to the model, as determined by the data and strength of the prior distribution. Specifically, following convergence of the variational iteration and hard assignments of variants to clusters, we remove any clusters having less than the larger of three variants or 0.5% of $N$, the total number of variants assigned to them, a condition similar to our earlier approach. [14] If clusters are removed, the algorithm is again executed until convergence. For the beta mixture model, convergence is achieved when the absolute difference between all $\pi_k$ across consecutive iterations is less than $10^{-4}$. This condition differs slightly for binomial and Gaussian mixture models (see Supporting Text). The minimum cluster membership is motivated by the requirement of needing at least two proportions to fix the two degrees of freedom, $u_k$ and $v_k$, of a beta distribution. More intuitively, clustering is effectively a separation of intra- and inter-cluster distances. Defining an intra-cluster distance requires at least two items be assigned to that cluster.

To be conservative in our assessment of subclonality, we require clusters be well separated. Previously, [14] we used a condition on overlapping cluster standard error of the means to detect and remove overlapping clusters. Here, we instead adopt a quantitative notion of cluster overlap, [36, 37] in which overlap between clusters $k$ and $k'$ results in uncertain assignments of some variants, causing them to have appreciable $r_{nk}$ *and* $r_{nk'}$. This in reflected in a large relative (to the "size", $\sum_n r_{nk}$, of cluster $k$) cluster overlap

$$\Upsilon_{k,k'} \equiv \frac{\sum_n r_{nk} r_{nk'}}{\sum_n r_{nk}} \ .$$

Minimizing this quantity for all $k' \neq k$ is equivalent to maximizing the "self-overlap" of cluster $k$ with itself,

$$\bar{\Upsilon}_k \equiv 1 - \sum_{k' \neq k} \Upsilon_{k,k'} = \frac{\sum_n r_{nk} r_{nk}}{\sum_n r_{nk}} \ ,$$

which satisfies, $N^{-1} \leq \bar{\Upsilon}_k \leq 1$. Hence, we remove any cluster having a $\bar{\Upsilon}_k$ less than a threshold $\Upsilon$. Overlap between (independent) clusters will be more likely in lower dimensional problems; hence, to determine a dimensionality-dependent $\Upsilon$ we clustered simulated data sets by sampling a mixture of beta distributions in one, two, or three dimensions and having between two and five clusters. Average concordance (across data sets of a given dimensionality) between the clustered and known results (in terms of fraction of correctly assigned items) was stable for a wide range of $\Upsilon$ within each dimension: $\Upsilon$ in the range of $0.5$ to $0.8$ achieved the maximal concordance (of $0.86$) in one dimension, $\Upsilon$ in the range of $0.83$ to $0.96$ achieved a concordance of $0.96 - 0.97$ in two dimensions, and $\Upsilon$ in the range of $0.84$ to $0.99$ achieved a concordance greater than or equal to $0.97$ in three dimensions. Intuitively, we anticipate that the probability of clusters overlapping scales inversely with the number of dimensions. Hence, we define $\Upsilon_S$ for an $S$-dimensional problem as $\Upsilon_S \equiv \Upsilon_1^{1/S}$, where $\Upsilon_1 = 0.70$ was selected so that $\Upsilon_S$ passes through the above optimal regions defined by the simulation. Namely, $\Upsilon_2 \approx 0.84$ and $\Upsilon_3 \approx 0.89$. Results for these settings of $\Upsilon_S$ across all simulated data sets are shown in Fig. S5.

We detect outliers using a more formal approach than our previous method, [14] by calculating the $p$-value of a variant with the respect to the cluster to which it has been assigned (via a hard assignment). If the probability of the variant belonging to that cluster is less than $p^{\min}$, the variant is removed from the analysis. The default used in this manuscript is $p^{\min} = 10^{-2}$ (which is *not* corrected for multiple testing). The $p$-value of a variant $\boldsymbol{\chi}_n$ is calculated with respect to the predictive posterior distribution [Eq. (9)] as

$$\int \theta \left[ p(\boldsymbol{\chi}_n | X) - p(\boldsymbol{\chi} | X) \right] p(\boldsymbol{\chi} | X) d\boldsymbol{\chi}$$

where $\theta(x)$ is the Heaviside step function with $\theta(x) = 1$ for $x > 0$ and $\theta(x) = 0$ for $x < 0$. In the case of beta mixture models, this integral is evaluated numerically by sampling from the predictive posterior distribution and then evaluating sampled variants with that distribution, which again involves numerical integration. For computational efficiency, we only calculate this integral for variants likely to be outliers, which we heuristically define as variants whose VAF $f$ in each sample $s$ lies outside of the narrowest interval containing $\mathrm{erf}(0.75/\sqrt{2}) \approx 0.55$ of the fluctuation in the mean of cluster $k$. This interval $(lo_{ks}, hi_{ks})$ is determined as the narrowest such interval satisfying

$$0.55 = \int_{f=lo_{ks}}^{hi_{ks}} \int_{u=0}^{\infty} \int_{v=0}^{\infty} \delta\left( f - \frac{u}{u+v} \right) \mathcal{G}am(u; \mu_{ks}, \alpha_{ks}) \mathcal{G}am(v; \nu_{ks}, \beta_{ks}) df\, du\, dv$$

and involves integrating the mean, $u/(u+v)$, with respect to the posterior distribution.

Several iterations of the AML28 data set following the $k$-means initialization (above and Fig. S6A) are shown in Figs S6B and C, with the complete run shown in Video S2.

### Variant Detection and Copy Number calling

Sequencing, alignment, and variant calling were performed as previously described [5]. Somatic copy number events were detected using copyCat (http://github.com/chrisamiller/copycat/).

### PyClone

PyClone version 0.12.3 was downloaded from http://compbio.bccrc.ca/software/pyclone/. VarScan 2 [51] was used to detect regions of LOH (which were excluded from further analysis). Copy number events detected by copyCat were quantized and passed to PyClone as `major_cn`, with `minor_cn` set to zero; additionally, PyClone was run with `--var_prior total_copy_number`, since allele-specific copy number calls were not provided. PyClone clustering used the beta-binomial mixture model. Initially, we attempted to cluster using 10,000 iterations and 1,000 burn-in iterations, as suggested by the authors (https://bitbucket.org/aroth85/pyclone/wiki/Tutorial). However, these parameters yielded discordant clusterings across three runs. Therefore, we varied the number of total iterations (and additionally varied the number of burn-in iterations to be 10% of the total iterations) and for each configuration assessed concordance across three independent runs. We choose the number of iterations at which the concordance across the three runs stabilized. These are given in Table S1. Concordance was evaluated for each of the three pairs and was calculated as the maximal fraction of items assigned to the same cluster across permutations of the cluster labels of one of the two runs being compared.

### THetA

THetA version 0.51 was downloaded from http://compbio.cs.brown.edu/projects/theta/. After failing to successfully run the program on the complete set of copy number events in MMY, we selected seven copy number regions, representing neutral, amplified, and subclonally deleted chromosomes, and ran THetA as described in the manual (parameters: `-n 3 -k 4 -m 0.10 --NUM_PROCESSES 2`). The resulting population frequencies and copy number assignments were used to infer the VAF at which a SNV in that region would appear. These sites were added to the list of SNV inputs to SciClone and clustered with default parameters.

## Acknowledgments

This work was supported by the National Human Genome Research Institute grants U01HG006517 to L.D. and U54HG003079 to R.K.W and by an International Myeloma Foundation Brian D. Novis research grant and Barnes-Jewish Hospital Foundation Award 7603-55 to B.S.W. We thank Joshua McMichael for help with figure preparation and David Spencer and Michael Wendl for critical review of the manuscript.

# References

1. Nowell PC (1976) The clonal evolution of tumor cell populations. Science 194: 23-8.

2. Greaves M, Maley CC (2012) Clonal evolution in cancer. Nature 481: 306-13.

3. Gerlinger M, Swanton C (2010) How darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. Br J Cancer 103: 1139-43.

4. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, et al. (2012) The origin and evolution of mutations in acute myeloid leukemia. Cell 150: 264-78.

5. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, et al. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 481: 506-10.

6. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, et al. (2012) The life history of 21 breast cancers. Cell 149: 994-1007.

7. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, et al. (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A 105: 13081-6.

8. Ding L, Ellis MJ, Li S, Larson DE, Chen K, et al. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 464: 999-1005.

9. Yachida S, Jones S, Bozic I, Antal T, Leary R, et al. (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467: 1114-7.

10. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, et al. (2013) Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 152: 714-26.

11. Shah SP, Roth A, Goya R, Oloumi A, Ha G, et al. (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 486: 395-9.

12. Egan JB, Shi CX, Tembe W, Christoforides A, Kurdoglu A, et al. (2012) Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. Blood 120: 1060-6.

13. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 366: 883-92.

14. Walter MJ, Shen D, Shao J, Ding L, White BS, et al. (2013) Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. Leukemia .

15. Anderson K, Lutz C, van Delft FW, Bateman CM, Guo Y, et al. (2011) Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature 469: 356-61.

16. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, et al. (2010) Inferring tumor progression from genomic heterogeneity. Genome Res 20: 68-80.

17. Keats JJ, Chesi M, Egan JB, Garbitt VM, Palmer SE, et al. (2012) Clonal competition with alternating dominance in multiple myeloma. Blood 120: 1067-76.

18. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90-4.

19. Xu X, Hou Y, Yin X, Bao L, Tang A, et al. (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 148: 886-95.

20. Hou Y, Song L, Zhu P, Zhang B, Tao Y, et al. (2012) Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. Cell 148: 873-85.

21. Fisher R, Pusztai L, Swanton C (2013) Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 108: 479-85.

22. Ma QC, Ennis CA, Aparicio S (2012) Opening pandora's box–the new biology of driver mutations and clonal evolution in cancer as revealed by next generation sequencing. Curr Opin Genet Dev 22: 3-9.

23. Yap TA, Gerlinger M, Futreal PA, Pusztai L, Swanton C (2012) Intratumor heterogeneity: seeing the wood for the trees. Sci Transl Med 4: 127ps10.

24. Merlo LM, Shah NA, Li X, Blount PL, Vaughan TL, et al. (2010) A comprehensive survey of clonal diversity measures in barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. Cancer Prev Res (Phila) 3: 1388-97.

25. Turke AB, Zejnullahu K, Wu YL, Song Y, Dias-Santagata D, et al. (2010) Preexistence and clonal selection of met amplification in egfr mutant nsclc. Cancer Cell 17: 77-88.

26. Su KY, Chen HY, Li KC, Kuo ML, Yang JC, et al. (2012) Pretreatment epidermal growth factor receptor (egfr) t790m mutation predicts shorter egfr tyrosine kinase inhibitor response duration in patients with non-small-cell lung cancer. J Clin Oncol 30: 433-40.

27. Roche-Lestienne C, Soenen-Cornu V, Grardel-Duflos N, Lai JL, Philippe N, et al. (2002) Several types of mutations of the abl gene can be found in chronic myeloid leukemia patients resistant to sti571, and they can pre-exist to the onset of treatment. Blood 100: 1014-8.

28. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, et al. (2012) Absolute quantification of somatic dna alterations in human cancer. Nat Biotechnol 30: 413-21.

29. Oesper L, Mahmoody A, Raphael BJ (2013) Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. Genome Biology 14: R80.

30. Magrangeas F, Avet-Loiseau H, Gouraud W, Lode L, Decaux O, et al. (2013) Minor clone provides a reservoir for relapse in multiple myeloma. Leukemia 27: 473-81.

31. Bishop CM (2006) Pattern recognition and machine learning. Information science and statistics. New York: Springer, 738 p. pp.

32. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. Boca Raton, Florida: Chapman & Hall/CRC.

33. Beal MJ (2003) Variational algorithms for approximate bayesian inference. Ph.D. thesis, University College London.

34. Ma Z, Leijon A (2011) Bayesian estimation of beta mixture models with variational inference. IEEE Trans Pattern Anal Mach Intell 33: 2160-73.

35. Fan W, Bouguila N, Ziou D (2012) Variational learning for finite dirichlet mixture models and applications. IEEE Trans Neural Netw Learn Syst 23: 762–774.

36. Korenblum D, Shalloway D (2003) Macrostate data clustering. Phys Rev E 67: 056704.

37. White BS, Shalloway D (2009) Efficient uncertainty minimization for fuzzy spectral clustering. Phys Rev E Stat Nonlin Soft Matter Phys 80: 056705.

38. Svensén M, Bishop CM (2005) Robust bayesian mixture modelling. Trends in Neurocomputing 64: 235–252.

39. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, et al. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 361: 1058-66.

40. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q (2013). Inferring clonal evolution of tumors from single nucleotide somatic mutations. URL `http://arxiv.org/abs/1210.3384`.

41. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, et al. (2012) Comphrehensive molecular portraits of human breast tumours. Nature 490: 61-70.

42. Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. (2013) Integrated genomic characterization of endometrial carcinoma. Nature 497: 67-73.

43. Paschka P, Schlenk RF, Gaidzik VI, Habdank M, Kronke J, et al. (2010) Idh1 and idh2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with npm1 mutation without flt3 internal tandem duplication. J Clin Oncol 28: 3636-43.

44. Ward PS, Patel J, Wise DR, Abdel-Wahab O, Bennett BD, et al. (2010) The common feature of leukemia-associated idh1 and idh2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. Cancer Cell 17: 225-34.

45. Wang F, Travins J, DeLaBarre B, Penard-Lacronique V, Schalm S, et al. (2013) Targeted inhibition of mutant idh2 in leukemia cells induces cellular differentiation. Science 340: 622-6.

46. Lee J, Hoi CS, Lilja KC, White BS, Lee SE, et al. (2013) Runx1 and p21 synergistically limit the extent of hair follicle stem cell quiescence in vivo. Proc Natl Acad Sci U S A 110: 4634-9.

47. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, et al. (2013) Emerging landscape of oncogenic signatures across human cancers. Nat Genet 45: 1127–1133.

48. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, et al. (2010) Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 107: 16910-5.

49. Ha G, Roth A, Lai D, Bashashati A, Ding J, et al. (2012) Integrative analysis of genome-wide loss of heterozygosity and mono-allelic expression at nucleotide resolution reveals disrupted pathways in triple negative breast cancer. Genome Research 22: 1995–2007.

50. Griffith M, Griffith O, Coffman A, Weible J, McMichael J, et al. (2013) Dgidb: mining the druggable genome. Nat Methods 10: 1209–1210.

51. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-76.

52. Boyd S, Vandenberghe L (2004) Convex Optimiziation. Cambridge University Press.

# Figure Legends

**Figure 1** Inferring subclonal architecture objectively in multiple myeloma. (a) Kernel density plots of VAFs across regions with copy number one, two, or three, posterior predictive densities summed over all clusters for copy number neutral variants, and posterior predictive densities for each cluster/component. (b-d) VAFs plotted versus read depth for each of the three copy number regions. Colors in (b) depict cluster assignments. (c) Three mutation clusters (green, dark orange, and blue) were detected using variants from copy number neutral segments. (d) Two clusters centered at VAF 31% and 62% were detected from variants in copy number three segments; they likely result from single-copy amplification of the wild-type or the mutant allele of mutations in the founding clone.

**Figure 2** Overcoming uncertainty in sparse exome-sequencing data to determine clonal structure and mutation clonality. (a) Breast cancer sample with well-defined clones. (b) Endometrial cancer sample with overlapping clusters. *PIK3CA* mutations are strongly associated with the dominant clone (posterior probabilities $>$93%), whereas the clonal context of an *ATM* mutation is more ambiguous (57.8%).

**Figure 3** Refining subclonal architecture from longitudinal analysis of tumor/relapse pair in acute myeloid leukemia (AML). Two-dimensional analysis of tumor/relapse sample (a) dissects clusters one and four, which overlap in the relapse sample (b), and one, two, and three, which overlap in the tumor sample (c). Single-sample analyses (b and c) show histogram (rectangles) with posterior predictive densities. Several genes recurrently mutated in AML [5] are highlighted. (d) Inferred schematic of clonal evolution from a single hematopoietic stem cell, showing percentage of cells belonging to each clone (i.e., twice VAF for this nearly pure sample). Broken vertical white lines correspond to primary tumor sample (before chemotherapy) and relapse subsequent to treatment.

**Figure 4** Determining stability of inferred subclones as a function (a) of number of variants, (b) of inter-cluster separation, and (c) of clustering method from AML sample. (a) A fraction of the ~800 variants from Fig. 3 were randomly sampled and the resulting number of clusters was inferred using beta mixture modeling. Error bars represent standard deviation ($n = 10$). (b) Mutations from clusters one and two from the AML relapse sample were used to assess the limits of cluster separability. As the distance between the two mutation groups was varied, the resulting clusters were assessed for overlap (the fraction of the data within a single standard deviation of both clusters) and accuracy (the fraction of items that were correctly assigned to a second cluster). (c) Consensus clustering of the AML data set (Fig. 3) for number of initial clusters varied from six to 15 and clustering method varied across beta, Gaussian, and binomial mixture models for a total of 30 runs. $N \times N$ consensus matrix holds all $N$ variants across both rows and columns and has been reordered so that variants belonging to the same cluster are adjacent to one another. Matrix entry $i, j$ is the fraction of runs in which variant $i$ and $j$ were co-clustered; entry $1, 1$ corresponds to the top-left of the matrix heat map. The narrowest neutral-colored band corresponds to a single variant alternatively classified by Gaussian mixture modeling (Fig. S2A). The larger neutral-colored band corresponds to variants alternatively classified as a sixth cluster by binomial mixture modeling (Fig. S2B).

**Figure 5** Assessing intratumor spatial heterogeneity and treatment response with multiple biopsies. Three breast tumor samples from a single individual were simultaneously analyzed: two spatially distinct samples from a primary tumor and one sample taken after aromatase-inhibitor treatment. (a-c) Two-dimensional slices and (d) still frame of the full three-dimensional interactive plot.

**Supplementary Figure 1** Integration of copy number-derived subclonal information from THetA. THetA was used to detect clonal and subclonal copy-number events in a multiple myeloma sample, then converted to pseudo-VAFs and co-clustered with SNV data using SciClone. CN-derived points are highlighted in yellow. The leftmost two CN events are single points and the rightmost point consists of six overlapping points.

**Supplementary Figure 2** Detecting ambiguous or low-confidence associations between a variant and clone from inconsistent assignments across clustering methods. Clonal dissection of AML sample (Fig. 3) based on (a) Gaussian or (b) binomial variational mixture modeling. Beta mixture modeling (Fig. 3) differs from Gaussian mixture modeling in the single variant highlighted by arrow in (a) and from binomial mixture modeling in the separation of cluster six from cluster one in (b).

**Supplementary Figure 3** Confirming subclonal AML populations using an independent method. PyClone largely recapitulates subclonal architecture inferred by SciClone (Fig. 3), though the parameter settings used here (default hyperparameters to beta-binomial mixture, with 10,000 iterations, and a burn-in of 1,000 iterations) overdissect the founding clone.

**Supplementary Figure 4** Confirming subclonal breast tumor populations using an independent method. PyClone clustering of variants in copy-number neutral regions is similar to that obtained by SciClone (Fig. 5), though the former partitions the variants spread along the pre-treatment tumor 2 axis (clusters 1 and 2), as well as those belonging to the founding clone (clusters 7 and 8).

**Supplementary Figure 5** Assessing concordance between known and clustered results. Beta mixtures having two to six components were sampled in (a) one, (b) two, or (c) dimensions and clustered. Concordance is the fraction of data points correctly clustered; the highest concordance resulting from a permutation of the cluster labels is reported. Reported self-overlap is the minimum reported over any cluster, i.e., $\min_k \bar{\Upsilon}_k$. Self-overlap is shifted by $0.1$ in the plots for visual purposes to avoid obscuring concordance.

**Supplementary Figure 6** Converging to clustering solution using variational Bayesian beta mixture model. $k$-means initialization (A) of AML sample (Fig. 3) and results following second (B) and fourth steps (of six) in iteration (C).

**Supplementary Video 1** Interactive, three-dimensional clustering of three breast tumor samples from a single individual (see Fig. 5d).

**Supplementary Video 2** Video of convergence of AML sample clustering (see Fig. 3 and Fig. S4).