S1 Block structure does not depend on a specific evolutionary model

The block structure that SBMClone relies upon does not require a particular evolutionary model, it simply reflects the presence of groups of cells that share groups of mutations. For example, see clone tree T in Fig. S1. Each vertex in T corresponds to a clone in the block matrix X, and each edge in T corresponds to a mutation cluster in X. While this clone tree has a deletion of mutation cluster B_1 and homoplasy of B_4 (i.e., the mutations in B_4 were gained twice by two distinct clones), both of which violate the infinite sites assumption (i.e., perfect phylogeny model), the block matrix X still conforms to our model in that rows and columns can be organized into blocks of 1-entries (gray) and blocks of 0-entries (white).

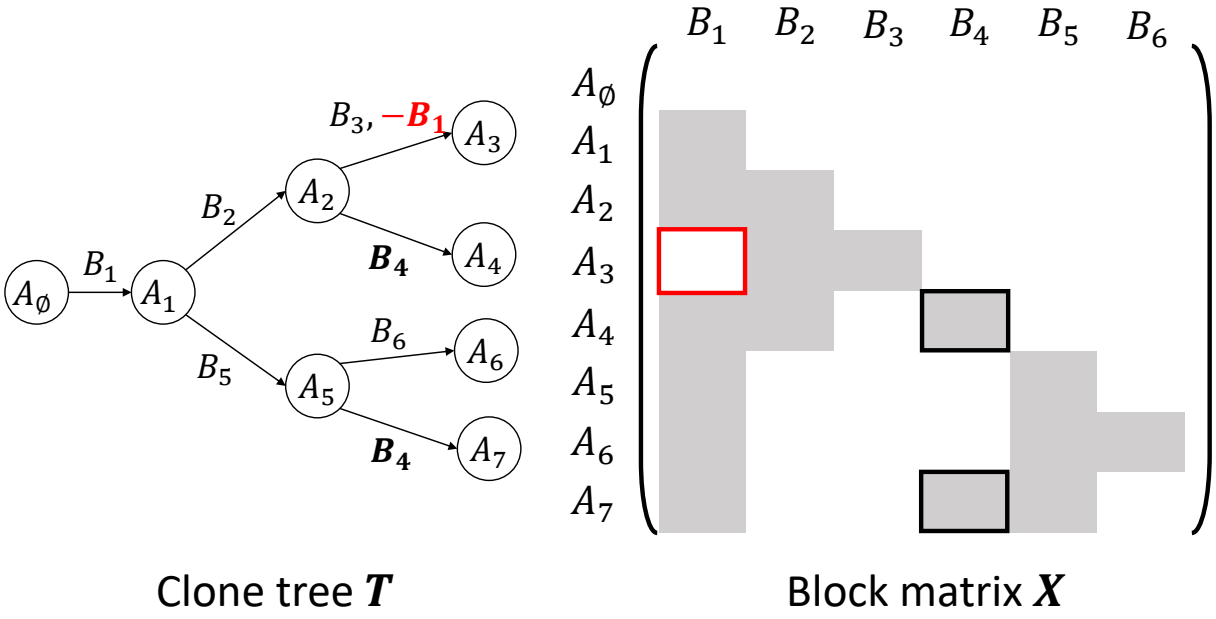

Figure S1: **Block structure does not depend on a specific evolutionary model. (Left)** A clone tree T containing a deletion of the mutations in cluster B_1 (red) and homoplasy (independent occurrence of same mutations in different cells) of the mutations in cluster B_4 (bold), both of which violate the infinite sites assumption. **(Right)** The block matrix X where each vertex in T corresponds to a row block in X and each edge in T corresponds to a column block in X. 1-entries are indicated in gray and 0-entries are indicated in white. The deletion of mutations in B_1 in clone A_3 is indicated by a red box, and the homoplasy of mutations in B_4 is indicated by black boxes.

S2 Relationship between p and sequencing coverage

While the per-cell coverage statistic (normally computed as the total number of bases sequenced from a cell divided by the length of the normal human genome) is in principle proportional to the block probability p, there are a few key issues that complicate this relationship:

1. Coverage is computed across all loci in the genome, but the number n of mutations being considered may be a subset of loci. For example, standard variant calling pipelines would reduce n to a subset of loci with higher coverage.

2. Coverage counts total reads, while the block probability p is more closely related to variant reads. In the context of diploid cells and heterozygous SNVs, this would imply that p would be smaller than the coverage by about a factor of two – in the presence of more widespread aneuploidy such as whole-genome duplication, however, this may be a factor of four or more. Furthermore, copy number aberrations in cancer genomes are normally non-uniform, meaning that the average number of reads required to observe a mutation (and thus the proportionality between the coverage and p) varies across the genome.

Ultimately the density of the mutation matrix D depends strongly on the preprocessing steps used to construct it: quality thresholds for read alignment, variant calling, and any other filtering steps will all have an effect.

S3 Parameters for SBMClone , BnpC [1], and SCG [2]

The results shown in the main text marked as SBMClone apply the SBM inference algorithm implemented in [8] specifying a minimum of 4 blocks to infer. Those marked as H were run in the same way with the "nested" flag set to True, which corresponds to applying the hierarchical SBM likelihood function. SBMClone took no more than 30 minutes to complete on any instance.

BnpC [1] was run with default parameters except for the running time limit which was set to 4 hours (originally there is no limit on running time, which resulted in a running time of 22-48 hours for mutation matrices with 4000 rows and 5000 columns). For several mutation matrices with empirical densities (Fig S4A) and the real data (Fig 4B), BnpC exceeded its allocated memory of 64 GB or crashed with a floating point underflow error.

SCG was run using the doublet model and the same parameters as those used for the example data. Runtime was normally under 30 minutes, but occasionally exceeded 20 hours or consistently crashed on a dataset.

S4 Basic two-population simulations

Given block probability matrix P, block matrix X, clone assignments $\boldsymbol{v}$, and mutation group assignments $\boldsymbol{e}$, we observe each entry $x_{i,j}$ independently at random with probability p_{v_i,e_j}.

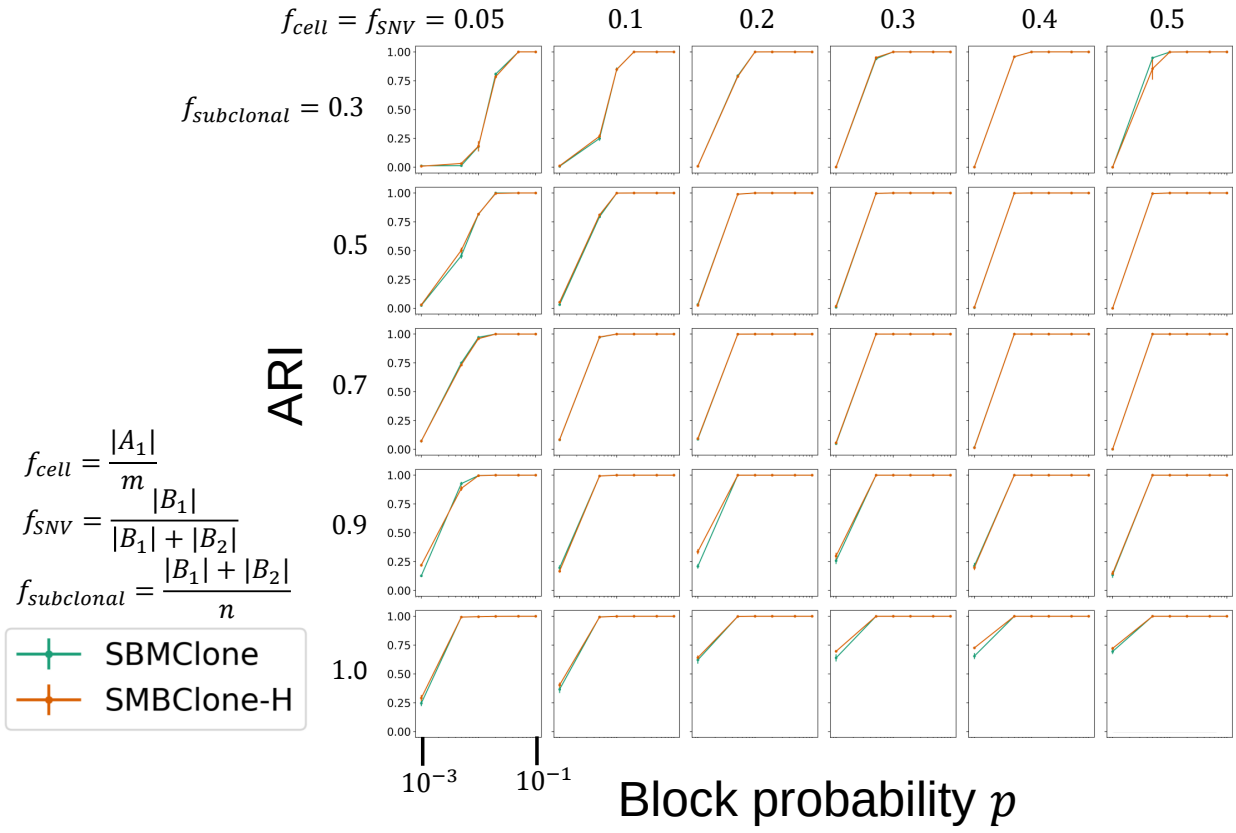

Figure S2: Applying SBMClone to basic two-population simulated data with varying block sizes.

For the basic simulated data, X had 2 clones and 3 mutation groups (see description and Fig. 2 in the main text).

The number m of cells was varied from 100 to 4000, and the number n of mutation was varied from 100 to 10000. The fraction f_{cell} of cells in clone 1 (i.e., A_1/m) and the fraction $f_{mutation}$ of subclonal mutation in mutation group 2 (i.e., $|B_2|/(|B_2| + |B_3|)$) were varied independently from 0.01 to 0.6. The fraction $f_{subclonal}$ of mutations that were not shared between clones (i.e., $(|B_2| + |B_3|)/n$) was varied from 0.3 to 1.0. Across all 128 888 combinations of parameters with 10 trials each, SBMClone-H yielded a median ARI of 0.49, while k-means yielded a median ARI of 0.04. For the figure shown in the main text (Fig. 2B), $f_{cell} = f_{mutation} = 0.4$ and $f_{subclonal} = 0.3$. For $m = 4000$ and $n = 5000$, SBMClone is able to consistently recover the two clones with p as low as 0.01, and often as low as 0.005 (Fig. SS2).

S5 Additional considerations for basic simulations

S5.1 Normal cells

To investigate the performance of SBMClone with varying tumor purity (i.e., the proportion

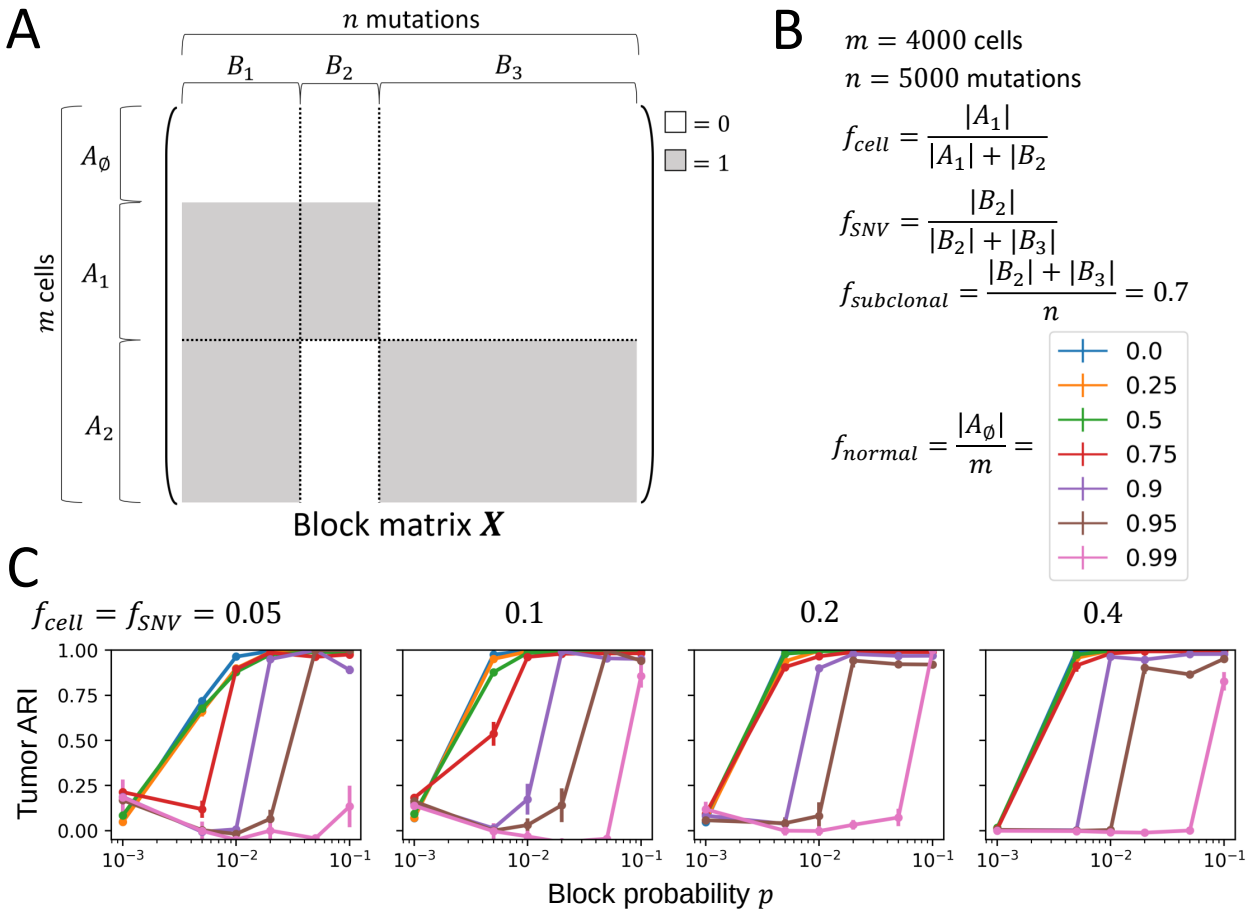

Figure S3: **SBMClone accurately recovers tumor cells on simulated block matrices with varying tumor purity.** **(A)** The block matrix X used to generate simulated data. Clone $A_\emptyset$ corresponds to normal cells with no mutations, and clones A_1 and A_2 are tumor clones that share mutation cluster B_1 and each have an exclusive mutation cluster. **(B)** Parameters used to generate simulated data and their associated values. **(C)** ARI computed on cells in A_1 and A_2 ("Tumor ARI," y-axis) measures the performance of SBMClone in recovering the two tumor clones A_1 and A_2 from simulated data with varying block probability p (x-axis). Each line corresponds to a different fraction f_{normal} of normal cells.

of tumor cells in a sample), we applied SBMClone to simulated block mutation matrices X in which we included a population $A_\emptyset$ of cells that contain no mutations (Fig. S3). The fraction $f_{normal}\ A_\emptyset/m$ of normal cells was varied from 0 to 0.99, the fractions f_{cell} and f_{SNV} were varied from 0.05 to 0.4, and the block probability p was varied from 10^{-3} to 10^{-1} (Fig. S3B-C). We evaluated the performance of SBMClone using tumor ARI, i.e., ARI computed only for cells in A_1 and A_2. We ran SBMClone using the hierarchical model (i.e., SBMClone-H) and set the minimum total number $k + \ell$ of blocks to 4. We found that SBMClone was able to recover (tumor ARI > 0.95) clones representing as few as 5% of the tumor cells with as much as 75% normal cells and with a block probability $p \geq 0.02$ (Fig. S3C)

S5.2 Whole-genome duplication

Whole genome duplications (WGD) also affect inference of clones using SNVs. Since WGDs duplicate the number of copies of every genomic region, somatic mutations that occur after WGDs are generally present in fewer copies of a genomic region. For example, a mutation that occurs in a genomic region not affected by further CNAs would be present in only one copy of the four total copies of that region, while, in the absence of WGD, a mutation is generally present in one copy of the 2 total copies. Therefore, the density of these mutations in the matrix would be lower. For example, if a block probability $p = 0.01$ corresponds to a simulated dataset with a sequencing coverage of $0.2\times$ in the absence of WGD, then $p = 0.02$ would generally result in a dataset with the same coverage but with the presence of a WGD.

S6 Generating ternary simulated data for SCITE and BnpC

SCITE and BnpC take as input a ternary matrix $D \in \{0, 1, ?\}$ in which a 1-entry represents a cell containing a mutation, a 0-entry indicates that a cell does not contain a mutation, and ? indicates missing entries. In order to adapt our simulation to this ternary matrix setting, we modify it as follows. Intuitively, our normal simulation framework corresponds to obtaining $p * n$ variant reads from each cell. Assuming that SNVs are in one copy of a diploid genome in every cell, one would also obtain $p * n$ reference reads from each cell. Thus, rather than simply sampling the 1-entries of $\hat{D}$, we also observed the 0-entries in $\hat{D}$ uniformly at random with probability p. Note that obtaining $p \cdot n$ reference reads from each cell is generous, because a) each cell might have more than 2 reference copies of the mutated locus and b) we assume that we do not obtain any false negatives, i.e., reference reads from cells with a mutation at that locus.

S7 Applying SCITE to example simulated data

We applied SCITE to 3 example simulated mutation matrices: a basic simulation with $f_{clonal} = 0.3$, balanced populations and $p = 0.1$ ("easy" case); a basic simulation with

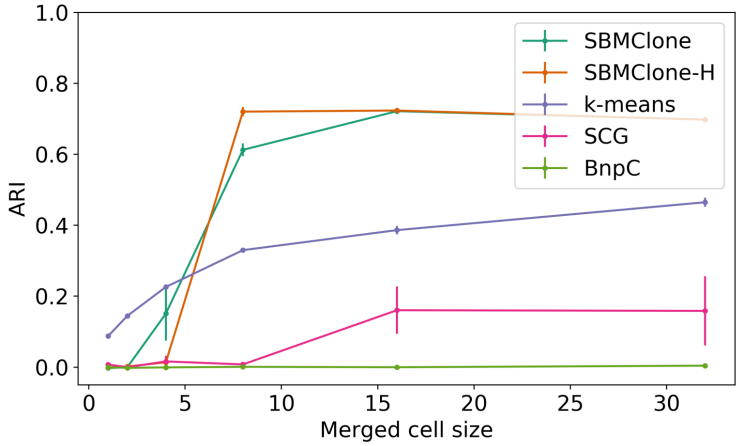

Figure S4: ARI (y-axis) measures the performance of the 5 methods on data simulated with empirical block probabilities (Figure 3C) when merging mutations across varying numbers of cells (x-axis).

balanced populations and $p = 0.01$ ("hard") case, and a tree-structured simulation with $p = 0.1$ ("tree" case). SCITE produced uninterpretable results on the easy case, returning a tree with 5000 vertices and attaching each cell to an average of 5.27 vertices. SCITE was unable to complete successfully on the hard or tree cases, each crashing multiple times after running for several days per attempt.

S8 Empirical probability matrix $\hat{P}$

S8.1 Constructing the empirical probability matrix $\hat{P}$

Using previous analysis [10] of a 10X Chromium single-cell dataset, we constructed an empirical probability matrix and used it to generate realistic simulated datasets (Fig. 3C) . The previous analysis of this dataset by Zaccaria and Raphael [10] inferred a phylogenetic tree from copy-number aberrations (CNAs) which divided 4 085 tumor cells into 8 clones. The authors identified 10 556 somatic mutations to corroborate this tree which were divided into 15 mutation clusters corresponding to the edges of the tree (Fig. 4A). Given the 4 085 $\times$ 10 556 mutation matrix $\widetilde{D} = [\widetilde{d}_{i,j}]$ corresponding to the analyzed cells and mutations, as well as the assignments of cells to the 8 clones and mutation to the 15 mutation clusters from this previous analysis, we computed each entry $\hat{p}_{r,s}$ of the empirical probability matrix $\hat{P}$ as follows:

$$\hat{p}_{r,s} := \frac{\sum_{i \in A_r} \sum_{j \in B_s} \widetilde{d}_{i,j}}{|A_r||B_s|}$$

S8.2 Evidence of errors in $\hat{P}$

We computed the empirical probability matrix $\hat{P}$ (Fig. 3C) from a 10X Chromium single-cell dataset to generate realistic simulated datasets. In particular, we analyzed the 10 556 somatic mutations that were identified in a previous study of this dataset by Zaccaria & Raphael [10] and that were assigned to 8 distinct tumor clones (Fig. 4A). To identify mutations present in small subpopulations of cells (< 50), the study also included mutations with low numbers of variant reads. Since errors are also generally characterized by a low number of variant reads, distinguishing such mutations from errors is very challenging. Therefore, many of these 10 556 previously identified mutations are likely to be false positives, as evidenced by the presence of mutation clusters with block probabilities lower than expected (Fig. 3C). For example, while one would expect that the blocks corresponding to the mutation cluster shared across all clones (Fig. 3C, first column) should have the high block probabilities, the block probabilities of these blocks are low (< 0.0023). This occurs because each mutation was assigned to the "latest" edge of the tree such that all of the cells with the mutation are within clones that descend from that edge. Therefore, false positive mutation calls may be in arbitrary cells that belong to clones in distinct phylogenetic branches, hence corresponding mutations are assigned to "early" edges of the tree even if they have a low number of variant reads. For example, a mutation that is shared between clones `J-I` and `J-II` but is falsely detected in clone `J-VII` would be placed in this mutation cluster even though it is not present in any clones on the right half of the tree.

S9 Merging cells to increase coverage *in silico*

On simulated data, we can evaluate the sensitivity of SBMClone by modifying the block probability p. On real data we cannot multiply the coverage, but we can merge information across similar cells to synthesize a higher-coverage dataset with fewer cells. Specifically, we randomly cluster cells within each population/clone into clusters of size c. Then, for each cluster, we merge the cells into a single row representing a higher-coverage pseudo-cell: this row harbors the union of the individual cells' mutations/1-entries. This approach offers similar performance on empirical densities to directly modifying the coverage (compare Figure 3D to Figure S4).

S10 Assessing SBMClone model selection

To assess how well SBMClone inferred the number k of clones (i.e., model selection), we computed the difference $\hat{k} - k$ between the number $\hat{k}$ of inferred clones and the number k of true clones across the simulated and real data matrices included in the paper. We observed that SBMClone converges to the correct number of clones as the block probability p or merged cell size (x-axis) increases, and generally infers the correct number of clones when the ARI is > 0.5 (Fig. S5).

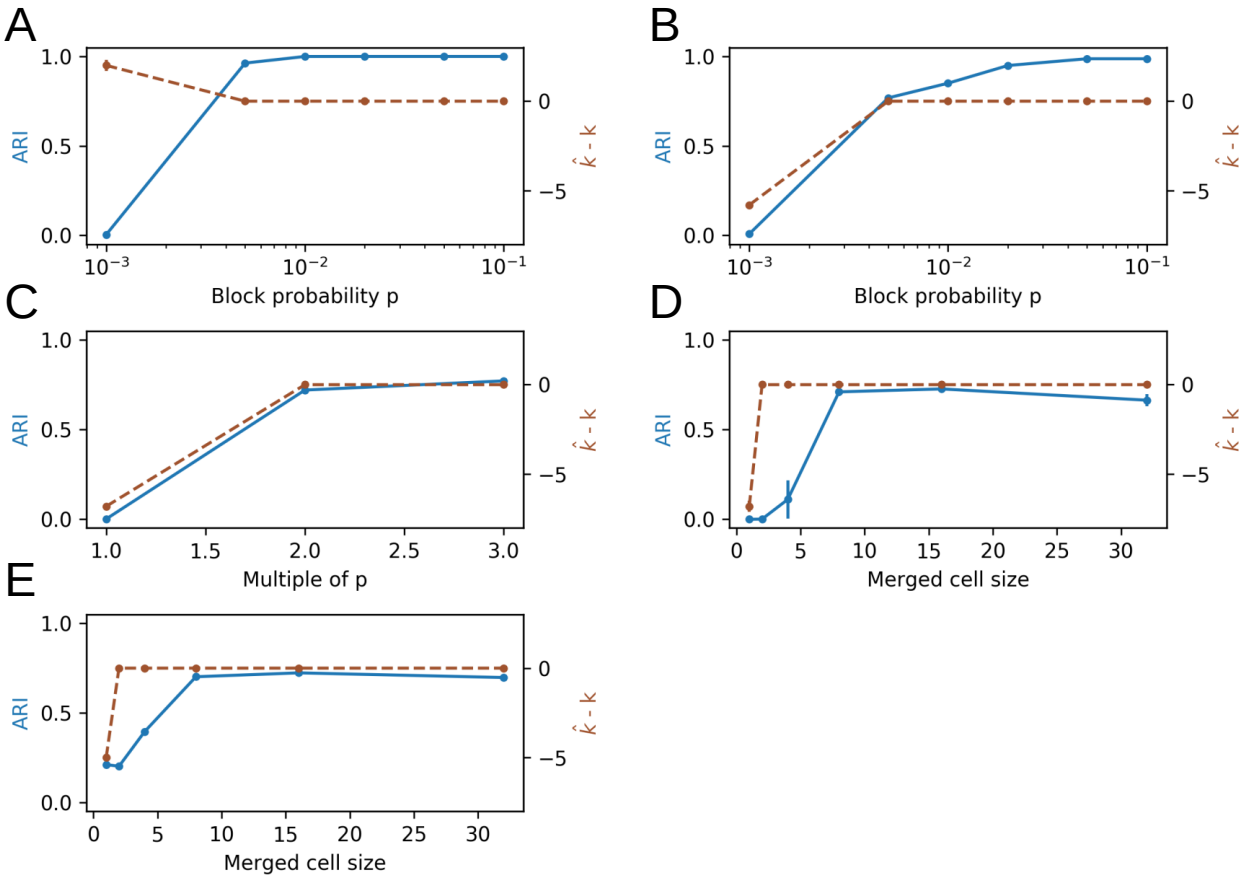

Figure S5: **SBMClone accurately infers the number of distinct clones.** Each plot matches a plot in the main text or supplement, showing the ARI (left y-axis, blue line) and the difference $\hat{k} - k$ between the number $\hat{k}$ of inferred clones and the number k of true clones (right y-axis, dotted brown line) of SBMClone-H when applied to **(A)** basic 2-population simulated data (matching Fig. 2B), **(B)** tree-structured simulated data with uniform block probability p (matching Fig. 3B), **(C)** tree-structured simulated data with empirical block probabilities $P = [p_{r,s}]$ scaled by a multiple (matching Fig. 3D), **(D)** tree-structured simulated data with empirical block probabilities and cell merging (matching Fig. S4), and **(E)** 10X breast cancer data with cell merging (matching Fig. 4B). Each point shows the mean and standard error across 5 random instances (where the randomness in real data is in the cell merging process).

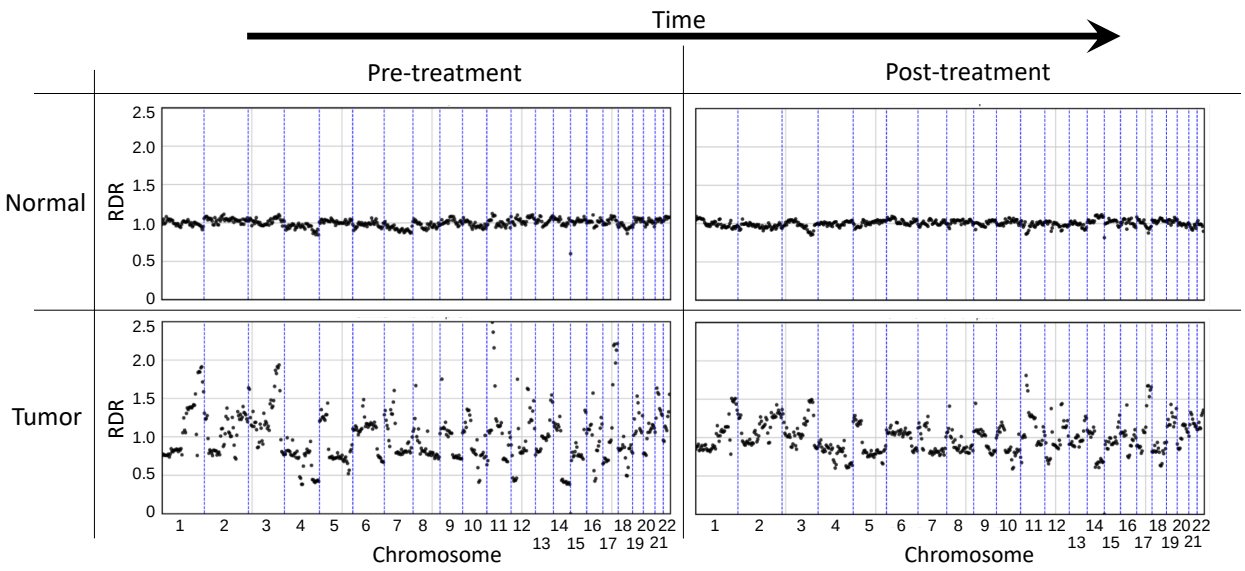

Figure S6: **Read-depth profiles support SBMClone's classification of tumor and normal cells in both pre- and post-treatment DOP-PCR sequencing data of breast cancer patient [3].** Read-depth profiles are computed by pooling all the sequencing reads from all pre- (left) or post-treatment (right) cells that are classified as either normal (top) or tumor (bottom) cells by SBMClone.

S11 DOP-PCR data analysis

After de-duplicating reads using SAMtools[7], we used Bowtie2 [6] to align DNA sequencing reads using the same procedure and reference genome hg19 as described in the published analysis [3]. We called mutations using BCFtools [7] and removed germline variants using dbDNP release 150 [9]. Then, since we did not have a matched normal sample, we formed a pseudo matched-normal sample by pooling together the sequencing reads from diploid cells as in previous studies [11, 10, 5], and removed mutations that were present in more than 2 cells in the matched normal sample as putative germline variants. Finally, we restricted our analysis to the 51 511 mutations with at least 10 total reads and a variant allele frequency (VAF) below 0.8 to avoid sequencing errors and germline homozygous variants, respectively.

To validate SBMClone's classification of tumor and normal cells in the DOP-PCR data [3], we computed the read-depth ratio (RDR) of four different groups of cells – normal and tumor cells identified in the pre-treatment and post-treatment samples – as follows. For each group of cells, we pooled the reads across all cells in the same group, partitioned the reference genome into 5 Mb genomic bins, and computed the RDR of each bin as the ratio between the total number of reads in the bin and the number of reads in the same bin from a pseudo matched-normal sample described above. While the normal cells had constant read depth across the genome in both pre- and post-treatment samples, the tumor cells in both pre- and post-treatment samples exhibited highly variable read-depth profiles across large portions of the genome. These read-depth profiles are consistent with the large CNAs identified in tumor cells in the published analysis of this dataset [3] (Fig. S6). This supports the novel finding of tumor cells in the post-treatment sample.

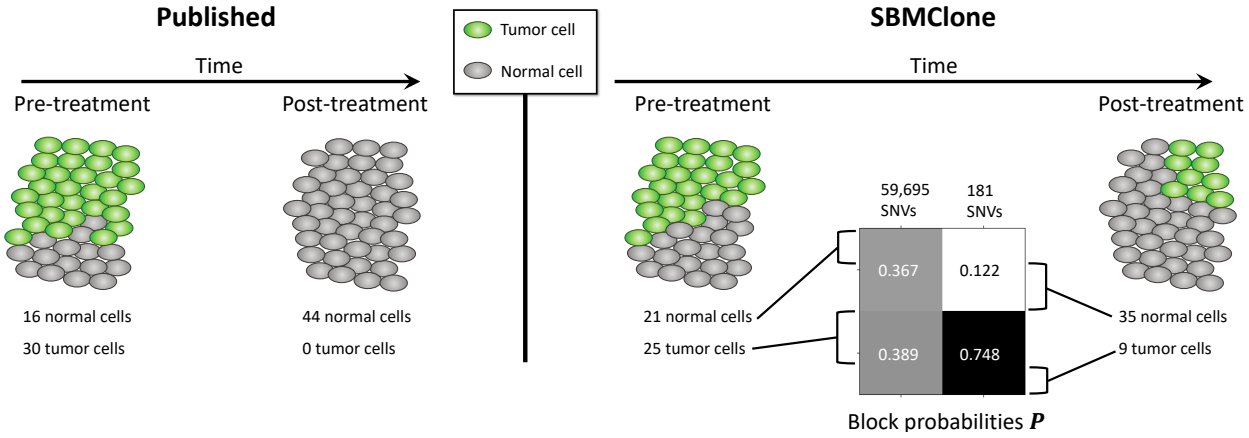

Figure S7: **Secondary analysis by SBMClone corroborates the previously uncharacterized presence of both pre- and post-treatment tumor cells in contrast to the published analysis of 90 breast tumor cells.** (Left) Published copy-number analysis of 90 cells from a breast cancer patient P2 [3] identified 16 tumor cells (green) exclusively among cells that were obtained pre-treatment, and no tumor cells were those that were acquired post-treatment. (Right) SBMClone analyzed 59 876 mutations from the 90 cells and identified tumor cells (green) both across pre-treatment (25 tumor cells) and post-treatment (9 tumor cells) cells. SBMClone's results are well supported by the identification of 181 SNVs that separate tumor from normal cells: the 181 SNVs have a high block probability in the tumor cells (0.784) but a very low block probability in the remaining normal cells (0.122).

We also analyzed these data with less attention to germline variants: we identified mutations using Varscan2 [4], then restricted our analysis to the set of 59 876 mutations that had at least 5 variant reads, at least 50 total reads, and a VAF below 0.8. While this $90 \times 59\,876$ mutation matrix was very dense (over 37% of the entries in the matrix were 1, as opposed to $< 10\%$ for most simulated data, $< 1\%$ for the 10X data, and about 6% for the other set of mutations from the same sequencing data), indicating that many germline variants may have been retained, we obtained a very similar bipartition of tumor and normal cells to the result in the main text (Fig. S7).

References

[1] Nico Borgsmueller, Jose Bonet, Francesco Marass, Abel Gonzalez-Perez, Nuria Lopez-Bigas, and Niko Beerenwinkel. Bayesian non-parametric clustering of single-cell mutation profiles. *bioRxiv*, 2020.

[2] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.

[3] Charissa Kim, Ruli Gao, Emi Sei, Rachel Brandt, Johan Hartman, Thomas Hatschek, Nicola Crosetto, Theodoros Foukakis, and Nicholas E Navin. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell*, 173(4):879–893, 2018.

[4] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[5] Emma Laks, Andrew McPherson, Hans Zahn, Daniel Lai, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Jerome Ting, et al. Clonal decomposition and dna replication states defined by scaled single-cell genome sequencing. *Cell*, 179(5):1207–1221, 2019.

[6] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

[7] Heng Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[8] Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014.

[9] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[10] Simone Zaccaria and Benjamin J Raphael. Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with chisel. *bioRxiv* 837195 [Preprint], November 2019.

[11] Hans Zahn, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature methods*, 14(2):167, 2017.