MAGOS: Discovering Subclones in Tumors Sequenced at Standard Depths

Navid Ahmadinejad[1,2], Shayna Troftgruben[1], Carlo Maley[2], Junwen Wang[1,3], Li Liu[1,2,3*]

[1]College of Health Solutions, Arizona State University, Tempe, AZ, 85004, USA

[2]Biodesign Institute, Arizona State University, Tempe, AZ, 85281, USA

[3]Department of Health Sciences Research & Center for Individualized Medicine, Mayo Clinic Arizona, Scottsdale, AZ, 85259, USA

[*] To whom correspondence should be addressed. Tel: (01)480-727-9813; Fax: (01)480-727-6947; Email: liliu@asu.edu

ABSTRACT

Understanding intratumor heterogeneity is critical to designing personalized treatments and improving clinical outcomes of cancers. Such investigations require accurate delineation of the subclonal composition of a tumor, which to date can only be reliably inferred from deep-sequencing data (>300x depth). To enable accurate subclonal discovery in tumors sequenced at standard depths (30-50x), we develop a novel computational method that incorporates an adaptive error model into statistical decomposition of mixed populations, which corrects the mean-variance dependency of sequencing data at the subclonal level. Tested on extensive computer simulations and real-world data, this new method, named model-based adaptive grouping of subclones (MAGOS), consistently outperforms existing methods on minimum sequencing depth, decomposition accuracy and computation efficiency. MAGOS supports subclone analysis using single nucleotide variants and copy number variants from one or more samples of an individual tumor. Applications of MAGOS to whole-exome sequencing data of 331 liver cancer samples discovered a significant association between subclonal diversity and patient overall survival. MAGOS is freely available as an R package at github (https://github.com/liliulab/magos).

INTRODUCTION

The development of a tumor is an evolutionary process that typically initiates from a single clone and grows into a diverse population of cells via incessant mutations and selections (1-3). As a tumor progresses over time and space, different cell populations (i.e., subclones) emerge, expand and diminish, leading to a heterogeneous malignancy with multifarious clinical presentations. Understanding of this dynamic system provides valuable knowledge to facilitate early diagnosis, effective treatment and outcome monitoring of cancers (4-9).

Whole-exome or whole-genome sequencing is a common approach to studying intratumor heterogeneity (10-12). By tracking relative abundances of genomic variants in a collection of cancerous cells, scientists aim to quantify the genetic diversity of a tumor and to reconstruct the phylogeny of subclones. While single-cell sequencing is on the rise, bulk sequencing remains as the dominant technology that interrogates an amalgam of heterogeneous cells collectively and relies on *in silico* analysis to de-convolute the mixed populations. Several computational methods have been developed for this purpose, such as SciClone, PyClone and Expands (4, 13, 14). However, because these methods require a minimum sequencing depth of 100x, they are not suited for samples sequenced at a standard 30-50x depth (15). This precludes the overwhelming majority of samples sequenced to date, including those generated by collaborative consortia, such as the TCGA Pan-Cancer Atlas (average depth = 68x). Furthermore, an independent evaluation of SciClone showed that consistent subclonal characterizations can only be achieved when the depth exceeds 300x (16). New methods capable of identifying subclones reliably at a reduced sequencing depth will help discover valuable information hidden in the myriad existing tumor genomic data that currently remain unexploited.

At the algorithmic level, the constraints of sequencing depths in current methods are at least partially due to unexplained variances in bulk sequencing data. A key assumption taken by these methods is the correspondence between variant allele frequencies (VAFs, i.e., fraction of reads containing a specific mutant allele among total reads) and cellular prevalence (i.e., fraction of cells carrying this particular mutant among all cells). Subclone discovery is then translated into a task of clustering similar VAFs (13). However, VAFs is also influenced by technical factors, such as sequencing depth. Due to randomness in sequencing procedures, the same subclone may give rise to a dispersed cluster of VAFs when sequenced at a low depth, but a tight cluster of VAFs when sequenced at a high depth (16). The spread of VAFs also correlates with cellular prevalence. VAFs of variants in a common subclone are expected to scatter more broadly than those in a rare subclone (16). Without considering these confounders, subclones reported by current methods are inevitably adulterated, especially when the sequencing depth is not high enough to create strong contrasts between subclones with similar cellular prevalence.

To detangle these technical variabilities from biological variabilities, we have developed a new method, named model-based adaptive grouping of subclones (MAGOS) that explicitly models the impact of sequencing depth and cellular prevalence on the variance of VAFs in subclone decomposition. Through extensive tests using computer simulations and real-world data, we show that MAGOS can accurately delineate subclonal structures of tumor samples sequenced at depths as low as 30x. MAGOS is also the fastest program when compared to SciClone and PyClone, showing an acceleration of 3-20 folds. We implemented MAGOS as an R package that is freely available at github (https://github.com/liliulab/magos).

RESULTS:

The purpose of MAGOS, as well as SciClone and PyClone, is to group variants that emerge and evolve together into a cluster based on similarities of VAFs. Each cluster thus corresponds to a subclonal expansion. In this context, we use VAF cluster and subclone interchangeably.

MAGOS Algorithm:

MAGOS supports subclone analysis of a tumor containing single nucleotide variants (SNVs) and copy number variants (CNVs) obtained from one or more samples. To illustrate the algorithms of MAGOS, we start with simple scenarios and gradually introduce complexities into the data model.

The simplest scenario involves a single tumor sample that has only SNVs, no CNVs and no contamination of normal cells. The task is to find clusters of SNVs with similar VAFs. Given a variant i, we denote the total number of reads aligned to this position as its sequencing depth e_i, and denote the fraction of reads containing the mutant allele among all reads as its VAF $v_i \in (0,1)$. For a set of m variants belonging to the same subclone, we model their VAFs as random samples from a beta distribution $Beta(\alpha,\beta)$ and require the two shape parameters (α and β) to satisfy

$$\begin{cases} \alpha + \beta = \bar{e} \\ \frac{\alpha}{\alpha+\beta} = \bar{v} \end{cases} \tag{1}$$

where $\bar{e}$ is the mean sequencing depth and $\bar{v}$ is the mean VAF of these variants. This configuration has the desired property that the variance of the beta distribution is positively correlated to the mean VAF and negatively correlated to the mean sequencing depth. Through this setup, we links the variance of VAFs to cellular prevalence and sequencing depth.

When multiple subclones are present, the observed VAFs are a mixture of samples from multiple beta distributions, each defined by a set of shape parameters. Therefore, identification of subclones is equivalent to decomposing mixed beta distributions (**Fig. 1A**). We solve this problem with a two-phase algorithm that performs agglomerative hierarchical clustering and adaptive partitioning.

In the first phase, we organize variants into a hierarchical tree structure by progressively grouping variants with similar VAFs into a cluster. Starting with leaf nodes each consisting of an individual variant, we iteratively merge a pair of nodes with the shortest distance among all pairs to create a new cluster till all variants are merged into one root cluster. Given two nodes (i.e., clusters), $C_1\ and\ C_2$ consisting of m_1 and m_2 variants, respectively, we define their distance d as a weighted sum of negative log likelihood that VAFs of all variants in $C_1\ and\ C_2$ are drawn from the same beta distribution,

$$d(C_1,\ C_2) = w\ \sum_{i\in\{C_1,\ C_2\}} -\log(\mathrm{P}(v_i\ ;\ Beta(\alpha,\beta))) \quad (2)$$

where α and β are calculated by solving equation (1) and the weight $w = 1/(m_1 + m_2) \cdot var(v) \cdot range(v)$. Because the distance is down-weighted by the variance and range of VAFs, given two pairs of clusters with similar values of log likelihood, MAGOS will choose the pair with a smaller variance and a narrower range to merge at an earlier step.

In the second phase, we identify boundaries of distinct beta distributions by traversing and partitioning the tree into clades (i.e., aggregation of clusters below a branching point). Unlike traditional approaches that cut the tree at a fixed branch level, we perform an adaptive splitting (**Fig. 1B**). Along the root-to-leaves path, we examine the clade at each branching point and test the null hypothesis that VAFs in this clade are drawn from the same beta distribution. This is done by comparing the observed variance of VAFs with the expected variance of VAFs. Specifically, given a clade containing m variants, we assume they belong to the same subclone and compute α and β by solving equation (1). We then draw m random samples $x_{1:m}{\sim}Beta(\alpha,\beta)$ and calculate $var(x)$. By repeating this process 1,000 times, we derive 1,000 $var(x)$ values representing the null distribution. We then use one-sample one-sided t-test to evaluate if $var(v)\ \leq \overline{var(x)}$. We reject the null hypothesis if the p-value < 0.01, which indicates VAFs of this clade are from heterogeneous beta distributions and needs to be partitioned further. Otherwise, we consider this clade as homogeneous and stop traversing below this branching point. We repeat this process till we find homogeneous clades along all branches or we reach the leaf nodes. Each of the resulted homogeneous clade represents a unique cluster.

When multiple samples of a tumor are analyzed, we expect that VAFs representing the same subclone change concordantly across all samples. However, because the sequencing depth and cellular prevalence of a subclone vary across samples, we need to estimate the beta distribution of this subclone in each sample separately. We then extend equation (1) to

$$\begin{cases} \alpha^s + \beta^s = \overline{e^s} \\ \frac{\alpha^s}{\alpha^s+\beta^s} = \overline{v^s} \end{cases} \tag{3}$$

where $\overline{e^s}$ is the mean sequencing depth and $\overline{v^s}$ is the mean VAF of these variants in sample s, and α^s and β^s are the two shape parameters of a beta distribution specific to this sample. To determine the between-cluster distance in each tumor sample, we extend equation (2) to

$$d^s(C_1,\ C_2) = w^s\ \sum_{i\in\{C_1,\ C_2\}} -\log(\mathrm{P}(v_i^s\ ;\ Beta(\alpha^s, \beta^s))) \tag{4}$$

where weight w^s is computed for each sample s as $1/m \cdot var(v^s) \cdot range(v^s)$. We then define the between-cluster distance across all S samples as

$$d(C_1,\ C_2) = \max(d^1, \ldots, d^s) \tag{5}$$

Thus, variants with concordant VAFs across all samples will be merged prior to variants with discordant VAFs. During tree partitioning, we evaluate if $var(v^s)\ \leq\ \overline{var(x^s)}$ for each sample s. We accept a clade as a single subclone if no sample produces a p-value < 0.01.

Because the above analyses are performed on SNVs not affected by CNVs, the mean VAF $\bar{v}_c$ of variants in a cluster c is linearly correlated with the cellular prevalence $\rho_c = 2\bar{v}_c$ in a sample. The cluster with a mean VAF of 0.5 corresponds to heterozygous SNVs in the founding clone fc. If a tumor sample is contaminated with normal cells, the mean VAF of the founding clone $\bar{v}_{fc}$ will deviate from 0.5 and the difference is proportional to the fraction of normal cells in the admixture. We then calculate the tumor purity $r = \rho_{fc} = 2\bar{v}_{fc}$ where $\bar{v}_{fc}$ is the mean VAF of variants in the cluster closest to 0.5. Other clusters with lower VAFs consist of SNVs emerged at various time points after the founding clone expansion.

For variants located in CNV regions, we assume they do not form new clusters but instead belong to clusters identified from the SNV analysis. Given a variant i, the expected VAF v_i' reflects the cellular prevalence ρ, average ploidy φ of the genomic region it resides and the number k of copies carrying the mutant allele,

$$v_i' = \frac{k\rho}{\varphi} \tag{6}$$

Note that φ is the average ploidy of the focus region in the entire sample and takes a continuous value. Finding the cluster assignment g from among existing SNV clusters $\{1, \dots, C\}$ is to solve

$$\arg\min_{k,g} |v_i - v_i'| = \arg\min_{k,g} \left| v_i - \frac{k\rho_g}{\varphi} \right| \tag{7}$$

where v_i is the observed VAF. We limit the search space of k to integers between 1 and $10 \times \varphi$. Extensions to accommodate CNVs in multiple samples are described in Supplementary Method.

Performance on simulated single tumor samples

To estimate the lower bound of sequencing depth and difference of mean VAFs ($\Delta\bar{v}$) between subclones that can be detected by MAGOS and two other methods, namely PyClone and SciClone, we simulated tumor samples with a simple two-population structure. A previous study has sequenced a primary leukemia sample at various depths from 60x to over 10,000x (16). This sample had an estimated purity of 90.3% and consisted of 1,343 somatic variants in the founding clone. The distributions of VAFs of these variants confirmed that the variance of VAFs was negatively correlated with the sequencing depth (Pearson correlation coefficient= –0.54, correlation test p-value=0.02). Using these variants as a pool, we randomly drew two populations each containing 100 variants. We then adjusted the $\bar{v}$ of each population between 0.05 and 0.45 with an interval of 0.05, combined these two populations and simulated read counts at an average sequencing depth of 30x, 50x, 100x, 200x, 300x and 500x via a Poisson-based down-sampling procedure (Supplementary Method). For each combination of $\bar{v}$ values and sequencing depth, we created 10 artificial admixtures. To quantify decomposition accuracies, we computed a weighted Jaccard index (*J*) that considered both the number of clusters identified and the assignment of variants (Supplementary Methods). A *J* score takes a value between 0 and 100, with 100 indicating a perfect match between true compositions and inferred compositions.

We first examined admixtures in which the $\bar{v}$ of one population was 0.45 representing a founding clone and the $\bar{v}$ of the other population was lower than 0.45 representing a derived clone. We presented three examples to illustrate different decomposition results of these methods. In an admixture with a $\Delta\bar{v} = 0.2$ sequenced at a 30x depth, the distributions of VAFs of the two populations showed substantial overlaps covering a continuous spectrum of VAFs from 0.09 to 0.85 (**Fig. 2A**). While MAGOS found the correct number of clusters, PyClone and SciClone reported excessive clusters, mistaking the large technical variance as variance caused by mixture of multiple populations. At the 300x depth, the same admixture produced well-separated distributions of VAFs (**Fig. 2B**). Both MAGOS and SciClone then found two clusters correctly. PyClone however still reported more than two clusters with many clusters containing one or two variants. When we reduced the $\Delta\bar{v}$ to 0.05, it was extremely challenging to divide the two populations even at the 300x coverage (**Fig. 2C**). In this case, SciClone reported one cluster and PyClone reported four clusters, respectively. Although MAGOS successfully recognized the existence of two clusters, it assigned only 75% of the variants to the correct cluster.

Using J >80 as the accuracy threshold, we recorded the minimum $\Delta\bar{v}$ value between the two populations at a given sequencing depth for each method. The advantage of MAGOS was the most prominent at the depths of 30x – 50x (**Fig. 2D**). In these simulations, MAGOS could produce accurate decompositions with $\Delta\bar{v}$ as low as 0.25. PyClone required a $\Delta\bar{v}$ of at least 0.35. SciClone could not achieve J >80 at any level of $\Delta\bar{v}$. MAGOS retained the leading position till the sequencing depth increased to 200x, beyond which both MAGOS and SciClone could decompose the admixtures equally well. Interestingly, the minimum $\Delta\bar{v}$ for PyClone remained at 0.35 across all sequencing depths.

Next, we examined the decomposition accuracies of all admixtures. The J score of all three methods was positively correlated with the $\Delta\bar{v}$ value (linear regression coefficients for MAGOS, PyClone and SciClone are 1.30, 1.03 and 0.87, respectively, all p-values $<10^{-12}$). The J score was positively correlated with the sequencing depth for MAGOS and SciClone (coefficients are 0.08, 0.15, respectively, p-value $<10^{-16}$), but not for PyClone (coefficient=0.006, p-value=0.51). At the 30x depth, MAGOS could achieve an average J score ≥ 80 when $\Delta\bar{v} \geq 0.25$ (**Fig. 2E**). In a total of 100 such admixtures, the average J score of MAGOS was 86.6, which was significantly better than that of PyClone (73.7, t test p-value=0.008) and SciClone (54.4, p-value=$3.5\text{x}10^{-8}$). As the depth increased to 300x, MAGOS could achieve an average J score ≥ 80 when $\Delta\bar{v} \geq 0.15$ (**Fig. 2F**). In a total of 210 such admixtures, the average J score of MAGOS was 97.2, which was significantly better than that of PyClone (64.9, t test p-value=$2\text{x}10^{-8}$) but slightly worse than SciClone (98.7, p-value=0.02).

Performance on simulated multiple tumor samples

To evaluate the performance on delineating complicated subclonal structure embedded in multiple samples from an individual tumor, we used an established method (17) to simulate the admixtures. Each simulation contains 200 variants distributed among 3 subclones, and 40 replicated were generated at sequencing depth of 30x, 50x, 100x and 300x. The largest improvement was at 30x depth where MAGOS had a mean *J* score of 0.82 whereas SciClone and PyClone had a mean J score of 0.66 and 0.78, respectively (paired t test p-values$<10^{-4}$, **Fig 3**). MAGOS remained at the leading performance at the sequencing depths increased to 50x and 100x ($p<0.05$). As the sequencing depth reached 300x, SciClone a performed equally well as MAGOS, achieving similar mean *J* scores of 0.95. Interestingly, although all three methods showed better performances at higher sequencing depth, PyClone was the least affected.

Performance on real-world sequencing data

We used the ultra-deep sequencing data published by Griffith et.al (16) to assess the accuracy and reproducibility of MAGOS and the other two methods. This dataset contains 1,337 high-quality somatic SNVs detected in a primary sample and a relapsed sample from a patient with acute myeloid leukemia. Each sample was sequenced at up to 10,000x depth for validation that represents the most comprehensively sequenced tumor. Although the true subclonal structures of these samples are unknown, we followed the authors' suggestion and used clusters detected at highest depths as the benchmark to evaluate clusters found at lower depths (data at 30x, 60x and 300x were available). The "best truth" consisted of five clusters with high confidences and two clusters with low confidences.

At the depth of 300x, MAGOS, PyClone and SciClone performed equally well, each reporting five to seven tight clusters (**Fig. 4A-C**). When the depth drops to lower than 60x, VAFs of these subclones showed large overlaps. However, MAGOS was still able to decompose the structure correctly, reporting six clusters (**Fig. 4D**). SciClone had great difficulties in separating overlapping clusters and reported only 3 subclones (**Fig. 4E**). Results from PyClone was similar to MAGOS but included small interspersed clusters (**Fig. 4F**). At the depth of 30x, MAGOS was the only method reporting the correct number of clusters and assigning variants to the correct cluster with high accuracies (**Fig. 4G**). SciClone reported results similar to 60x (**Fig. 4H**). PyClone added more than 10 small interspersed clusters in its result **(Fig. 4I**).

Analyzing TCGA liver cancer data: We applied the MAGOS method to whole-exome sequencing data of 331 liver hepatocellular carcinoma samples from the TCGA project. The majority (79.2%) of these tumors contained 3 or 4 subclones (**Table 1**). Using Cox proportional hazard regressions, we tested if the number of subclones in a tumor was significantly associated with patient overall survival via age at diagnosis, sex and tumor stages as covariates. We found a significant association among tumors of stage III (p-value=0.01, HR=1.67, **Fig. 5**). For comparisons, the number of mutations in a tumor is not a significant prognostic factor among these tumors (p-value=0.44). Therefore, the subclone number is a novel prognostic factor for stage-3 liver cancers that is independent of age at diagnosis, sex and total number of mutations.

DISCUSSIONS:

Cancer, as an evolutionary process, is born with a heterogeneous and dynamic nature(2, 3, 8). Precision identification and intervention of cancers shall consider the past, present and future of each tumor. With NextGen sequencing technologies, we can now catch snapshots of this process and potentially reconstruct the evolutionary history and trajectory of a tumor(6, 11, 12, 18-20). While single-cell sequencing is a promising technology to examine the genetic compositions of individual cells, uneven genome coverage, low accuracy of variant calls and prohibitive cost limit its usage in subclonal investigations (21-24). The majority of current studies and likely many others in the near future rely on bulk sequencing of mixed tumor cells and computational decomposition to identify variants that occur and evolve together. Several challenges emerge in these analyses.

First, sequencing depth is a key factor affecting the accuracy of identified subclones (15). Shown in both the simulated and real-world data (**Fig. 2-4**), as the sequencing depth drops, the centroid of each cluster remains unchanged while each cluster becomes more scattered, eventually leading to overlaps. Explicit modeling of this correlation enables MAGOS to accommodate large variances at a lower depth. However, variants in overlapping areas are impossible to separate. Instead of using VAF cutoffs to create artificial borders, a more informative measure is the probability of each variant belonging to a specific clone, which MAGOS reports.

Second, the difference of mean VAFs between subclones limits the power of distinguishing them. Increasing sequencing depth does not change the centroid of each cluster, thus helps little on detecting subclones with similar VAFs. Contrarily, additional samples from the same tumor helps segregate these clusters that are otherwise undiscernible. As MAGOS enables subclonal identifications from genomes sequenced at standard depths, the saved cost can be better invested on analyzing more samples. The benefit of sequencing additional samples is more evident at low sequencing depth, as shown in our test of PyClone. PyClone performs significantly worse than MAGOS when only a single sample is analyzed at low depth (**Fig. 2**). However, when analyzing 2 samples at depth 60x, PyClone results are similar to that from MAGOS (**Fig. 4**).

Third, as multiple samples from a tumor help identify segregating subclones and whole-genome sequencing reveals noncoding variants, these additional data also increase the computational complexity, which in turn requests efficient algorithms. We optimized the efficiency of MAGOS that tested its CPU time using simulated tumors. We varied the number of samples for each tumor, the number of mutations in each sample and the mean sequencing depth. Across all configurations, MAGOS showed 3-20x acceleration as compared to SciClone and PyClone (**Fig. 6A-C**), making it a fast and reliable method for subclone decompositions. The identified clusters can be used for further analyses, such as tumor phylogenetic inferences.

We implemented MAGOS as an open-source R package that is available through github (https://github.com/liliulab/magos).

Figure Legends

Figure 1. MAGOS algorithm. (**A**) Beta mixture distribution. The observed distribution (black curve) of VAFs is a combination of multiple hidden groups of VAFs, each forming a beta distribution (shaded curves) defined by different parameters. (**B**) Hierarchical clustering and adaptive partitioning. In this example, ten variants at the leaf nodes are progressively grouped into clusters based on VAF similarities to form a tree structure. To partition the tree, we follow the root-to-leave paths. At each branching point, the variance of the VAFs of the clade is compared with the expected variance. A cluster is accepted if the variance is lower than the expected value. Otherwise, it is rejected and partitioning continues (marked by black crosses). In this example, the red, blue and green cluster are accepted.

Figure 2. Performance of MAGOS, PyClone and SciClone on simulated single tumor samples, each consisting of two subclones. (**A, B, C**) The scatter plots show two simulated clusters of variants in a tumor sample. We varied the mean VAF of each cluster and the average sequencing depth. In panels A and B, the mean VAFs of the two clusters are 0.45 and 0.20, respectively, sequenced at an average depth of 30x (A) and 300x (B). In panel C, the mean VAFs of the two clusters are 0.45 and 0.4, respectively, sequenced at an average depth of 300x. (**D**) Minimum ΔVAF of two subclones that can be decomposed with an accuracy *J* score >80 by each method. Broken tops on bars indicate *J* scores >80 cannot be achieved. (**E, F**) Number of reported clusters (upper triangle) and *J* scores (lower triangle) at sequencing depths of 30x (D) and 300x (E). Displayed values are averages of 10 simulations. Perfect decompositions shall report 2 clusters and a *J* score value of 100.

Figure 3. Performance on simulated multiple tumor samples, each consisting of three subclones. Accuracies of different methods tested on tumors sequenced at depth from 30x to 1,000x. Asterisks indicate a significant better performance of MAGOS as compared to the other two methods.

Figure 4. Performance on empirical data of two samples (primary tumor and relapsed tumor) from the same patient. Scatter plots show VAFs of variants sequenced at average depths of

300x (**A-C**) and 30x (**D-F**). Dots of the same color are variants assigned to the same cluster by each method. Shaded ellipses represent “true” clusters inferred from data at ~10,000x sequencing depth. Radiuses of an ellipse correspond to 2 standard deviations of VAFs of variants belonging to a true cluster.

Figure 5. Kaplan-Meier plot of liver cancers at stage III. Tumors were stratified into groups based on the number of subclones.

Figure 6. Computational efficiency comparisons tested on simulated tumors. (A) Each tumor has two samples and each sample contains 50 to 1,000 mutations sequenced at 100x depth. (B) Each tumor has two samples and each sample contains 100 mutations sequenced at depth from 30x to 1,000x. (C) Each tumor has 1 to 4 samples and each sample contains 100 mutations sequenced at 100x depth.

REFERENCES

1. Miura S, Gomez K, Murillo O, Huuki LA, Vu T, Buturla T, Kumar S. Predicting clone genotypes from tumor bulk sequencing of multiple samples. Bioinformatics. 2018;34(23):4017-26. doi: 10.1093/bioinformatics/bty469. PubMed PMID: 29931046; PMCID: PMC6247940.
2. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306-13. doi: 10.1038/nature10762. PubMed PMID: 22258609; PMCID: PMC3367003.
3. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8. doi: 10.1126/science.959840. PubMed PMID: 959840.
4. Andor N, Harness JV, Muller S, Mewes HW, Petritsch C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics. 2014;30(1):50-60. doi: 10.1093/bioinformatics/btt622. PubMed PMID: 24177718; PMCID: PMC3866558.
5. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer. 2013;108(3):479-85. doi: 10.1038/bjc.2012.581. PubMed PMID: 23299535; PMCID: PMC3593543.
6. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jonsson G, Mariani O, Boyault S, Miron P, Fatima A, Langerod A, Aparicio SA, Tutt A, Sieuwerts AM, Borg A, Thomas G, Salomon AV, Richardson AL, Borresen-Dale AL, Futreal PA, Stratton MR, Campbell PJ, Breast Cancer Working Group of the International Cancer Genome C. The life history of 21 breast cancers. Cell. 2012;149(5):994-1007. doi: 10.1016/j.cell.2012.04.023. PubMed PMID: 22608083; PMCID: PMC3428864.
7. Ma QC, Ennis CA, Aparicio S. 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. 2012;22(1):3-9. doi: 10.1016/j.gde.2012.01.008. PubMed PMID: 22386266.
8. Aktipis CA, Kwan VS, Johnson KA, Neuberg SL, Maley CC. Overlooking evolution: a systematic analysis of cancer relapse and therapeutic resistance research. PLoS One. 2011;6(11):e26100. doi: 10.1371/journal.pone.0026100. PubMed PMID: 22125594; PMCID: PMC3219640.
9. Gerlinger M, Swanton C. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. Br J Cancer. 2010;103(8):1139-43. doi: 10.1038/sj.bjc.6605912. PubMed PMID: 20877357; PMCID: PMC2967073.
10. Schwartz R, Schaffer AA. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet. 2017;18(4):213-29. doi: 10.1038/nrg.2016.170. PubMed PMID: 28190876; PMCID: PMC5886015.
11. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, Wan Y, Zhang W, Shukla SA, Vartanov A, Fernandes SM, Saksena G, Cibulskis K, Tesar B, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, Getz G, Wu CJ. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152(4):714-26. doi: 10.1016/j.cell.2013.01.019. PubMed PMID: 23415222; PMCID: PMC3575604.
12. Egan JB, Shi CX, Tembe W, Christoforides A, Kurdoglu A, Sinari S, Middha S, Asmann Y, Schmidt J, Braggio E, Keats JJ, Fonseca R, Bergsagel PL, Craig DW, Carpten JD, Stewart AK. Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. Blood. 2012;120(5):1060-6. doi: 10.1182/blood-2012-01-405977. PubMed PMID: 22529291; PMCID: PMC3412329.
13. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, Vij R, Tomasson MH, Graubert TA, Walter MJ, Ellis MJ, Schierding W, DiPersio JF, Ley TJ, Mardis ER, Wilson RK, Ding L. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns

of tumor evolution. PLoS Comput Biol. 2014;10(8):e1003665. doi: 10.1371/journal.pcbi.1003665. PubMed PMID: 25102416; PMCID: PMC4125065.

14. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Cote A, Shah SP. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11(4):396-8. doi: 10.1038/nmeth.2883. PubMed PMID: 24633410; PMCID: PMC4864026.
15. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-32. doi: 10.1038/nrg3642. PubMed PMID: 24434847.
16. Griffith M, Miller CA, Griffith OL, Krysiak K, Skidmore ZL, Ramu A, Walker JR, Dang HX, Trani L, Larson DE, Demeter RT, Wendl MC, McMichael JF, Austin RE, Magrini V, McGrath SD, Ly A, Kulkarni S, Cordes MG, Fronick CC, Fulton RS, Maher CA, Ding L, Klco JM, Mardis ER, Ley TJ, Wilson RK. Optimizing cancer genome sequencing and analysis. Cell Syst. 2015;1(3):210-23. doi: 10.1016/j.cels.2015.08.015. PubMed PMID: 26645048; PMCID: PMC4669575.
17. El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics. 2015;31(12):i62-70. doi: 10.1093/bioinformatics/btv261. PubMed PMID: 26072510; PMCID: PMC4542783.
18. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, Shannon WD, Payton JE, Kulkarni S, Westervelt P, Walter MJ, Graubert TA, Mardis ER, Wilson RK, DiPersio JF. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481(7382):506-10. doi: 10.1038/nature10738. PubMed PMID: 22237025; PMCID: PMC3267864.
19. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, Wartman LD, Lamprecht TL, Liu F, Xia J, Kandoth C, Fulton RS, McLellan MD, Dooling DJ, Wallis JW, Chen K, Harris CC, Schmidt HK, Kalicki-Veizer JM, Lu C, Zhang Q, Lin L, O'Laughlin MD, McMichael JF, Delehaunty KD, Fulton LA, Magrini VJ, McGrath SD, Demeter RT, Vickery TL, Hundal J, Cook LL, Swift GW, Reed JP, Alldredge PA, Wylie TN, Walker JR, Watson MA, Heath SE, Shannon WD, Varghese N, Nagarajan R, Payton JE, Baty JD, Kulkarni S, Klco JM, Tomasson MH, Westervelt P, Walter MJ, Graubert TA, DiPersio JF, Ding L, Mardis ER, Wilson RK. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012;150(2):264-78. doi: 10.1016/j.cell.2012.06.023. PubMed PMID: 22817890; PMCID: PMC3407563.
20. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A. 2008;105(35):13081-6. doi: 10.1073/pnas.0801523105. PubMed PMID: 18723673; PMCID: PMC2529122.
21. Hughes AE, Magrini V, Demeter R, Miller CA, Fulton R, Fulton LL, Eades WC, Elliott K, Heath S, Westervelt P, Ding L, Conrad DF, White BS, Shao J, Link DC, DiPersio JF, Mardis ER, Wilson RK, Ley TJ, Walter MJ, Graubert TA. Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. PLoS Genet. 2014;10(7):e1004462. doi: 10.1371/journal.pgen.1004462. PubMed PMID: 25010716; PMCID: PMC4091781.
22. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90-4. doi: 10.1038/nature09807. PubMed PMID: 21399628; PMCID: PMC4504184.
23. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge J, Grubor V, Levy D, Lundin P, Maner S, Zetterberg A, Hicks J, Wigler M. Inferring tumor

progression from genomic heterogeneity. Genome Res. 2010;20(1):68-80. doi: 10.1101/gr.099622.109. PubMed PMID: 19903760; PMCID: PMC2798832.

24. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17(3):175-88. doi: 10.1038/nrg.2015.16. PubMed PMID: 26806412.

Figure 1

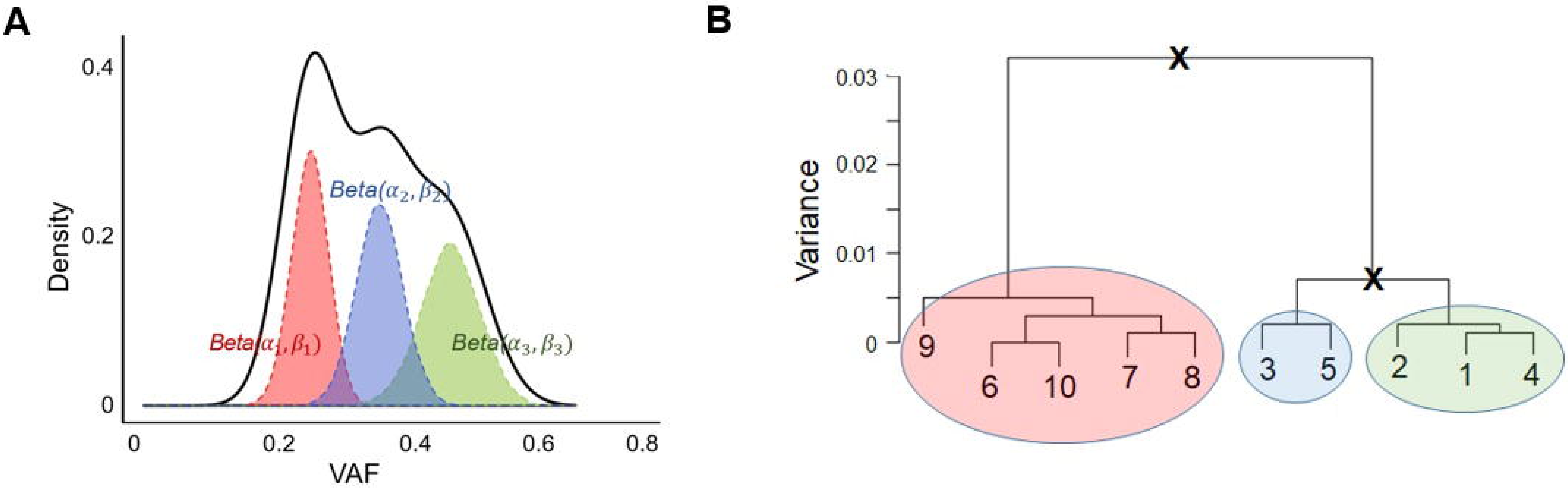

A
0.4
0.2
0
Density
Beta(α_1,β_1)
Beta(α_2,β_2)
Beta(α_3,β_3)
0
0.2
0.4
0.6
0.8
VAF
B
X
X
0.03
0.02
0.01
0
Variance
9
6
10
7
8
3
5
2
1
4

Figure 2

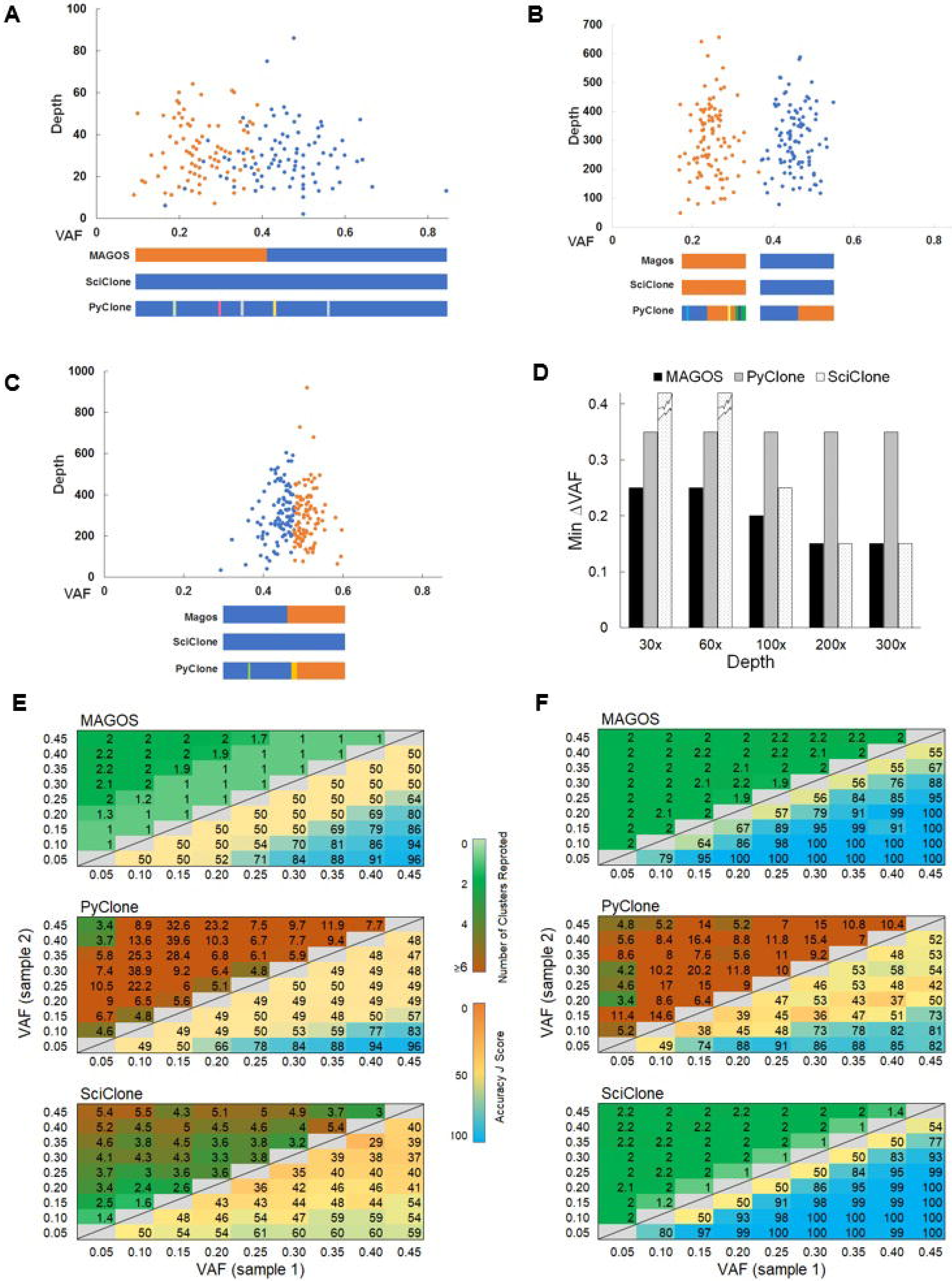
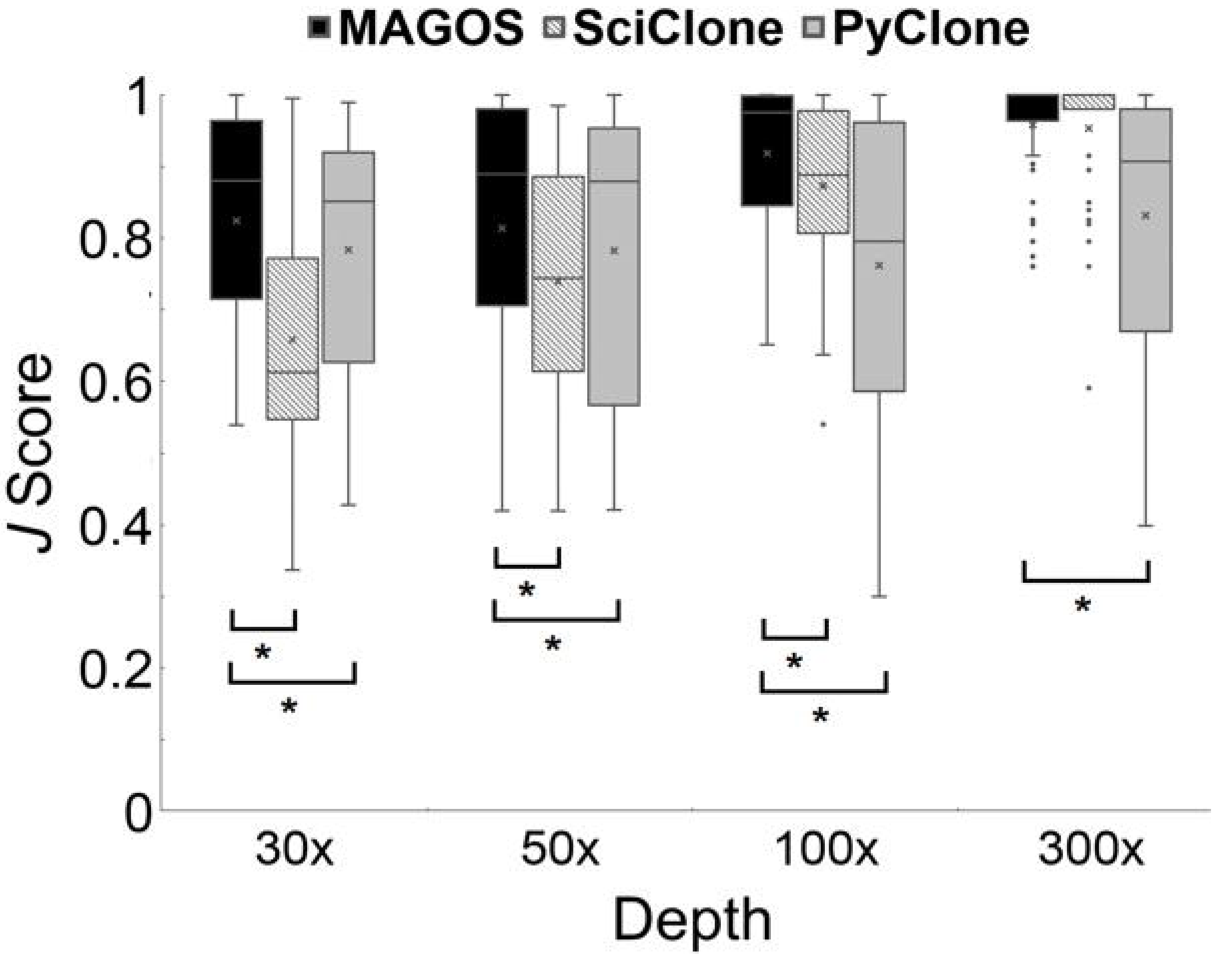

A
Depth
VAF
MAGOS
SciClone
PyClone
B
Depth
VAF
Magos
SciClone
PyClone
C
Depth
VAF
Magos
SciClone
PyClone
D
MAGOS
PyClone
SciClone
Min ΔVAF
30x
60x
100x
200x
300x
Depth
E
MAGOS
PyClone
SciClone
VAF (sample 2)
VAF (sample 1)
Number of Clusters Reproted
≥6
Accuracy J Score
F
MAGOS
PyClone
SciClone
VAF (sample 2)
VAF (sample 1)

Figure 3



■ MAGOS ▨ SciClone ■ PyClone
J Score
1
0.8
0.6
0.4
0.2
0
*
*
*
*
*
*
*
30x
50x
100x
300x
Depth

Figure 4

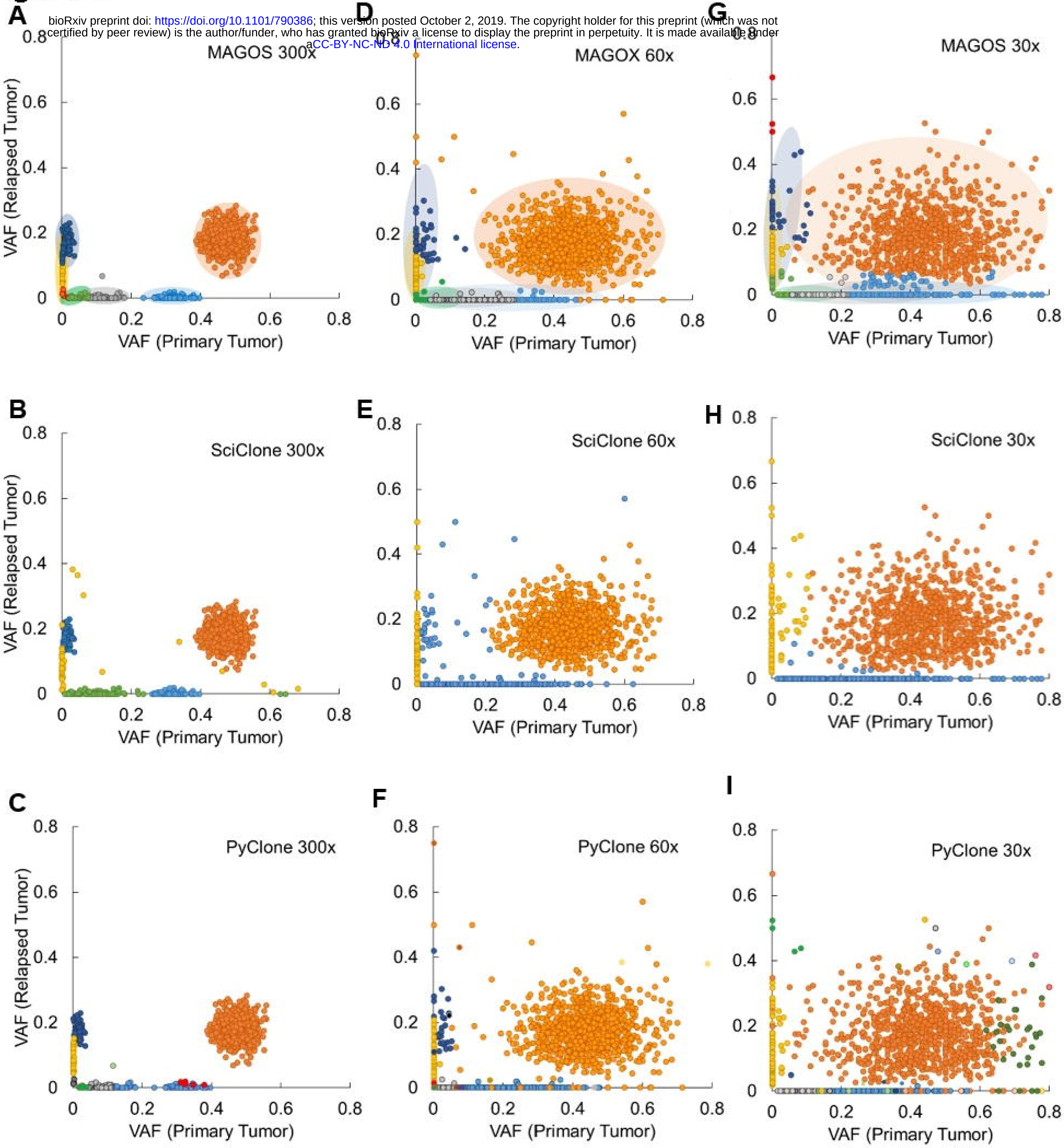
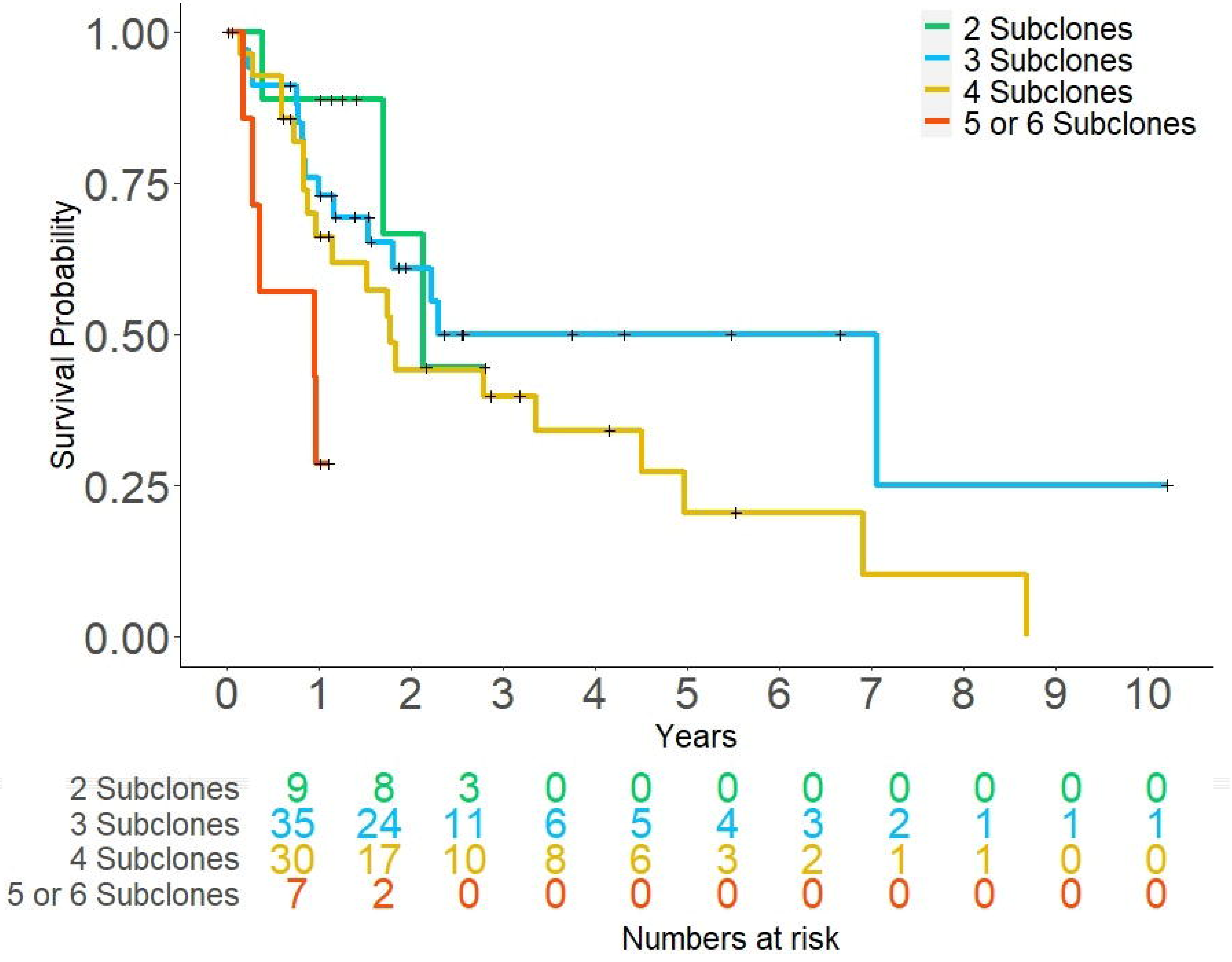

A
MAGOS 300x
D
MAGOX 60x
G
MAGOS 30x
B
SciClone 300x
E
SciClone 60x
H
SciClone 30x
C
PyClone 300x
F
PyClone 60x
I
PyClone 30x
VAF (Relapsed Tumor)
VAF (Primary Tumor)
0
0.2
0.4
0.6
0.8

Figure 5



2 Subclones	9	8	3	0	0	0	0	0	0	0	0
3 Subclones	35	24	11	6	5	4	3	2	1	1	1
4 Subclones	30	17	10	8	6	3	2	1	1	0	0
5 or 6 Subclones	7	2	0	0	0	0	0	0	0	0	0

Figure 6

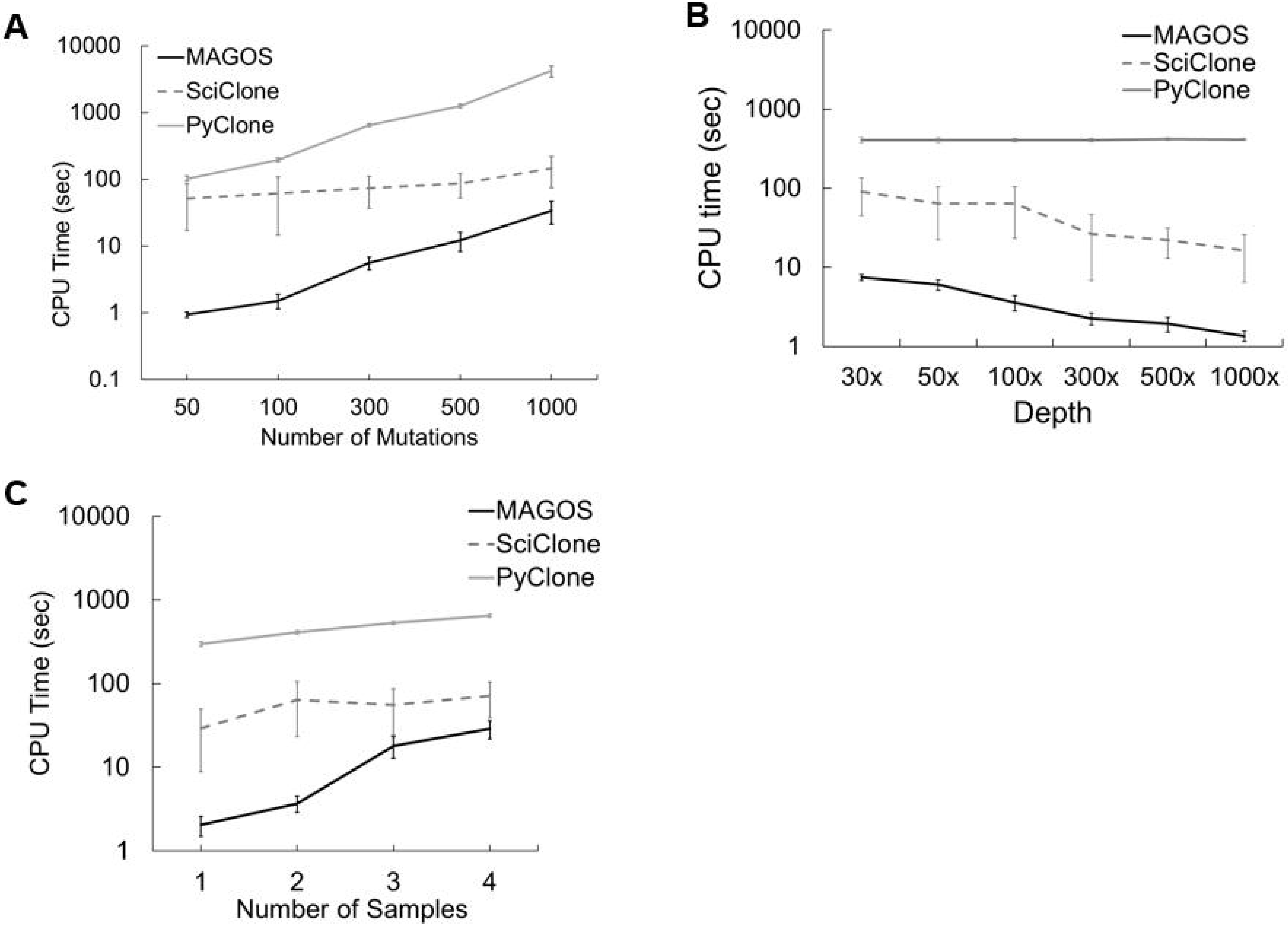

A
10000
1000
100
10
1
0.1
CPU Time (sec)
50
100
300
500
1000
Number of Mutations
MAGOS
SciClone
PyClone
B
10000
1000
100
10
1
CPU time (sec)
30x
50x
100x
300x
500x
1000x
Depth
MAGOS
SciClone
PyClone
C
10000
1000
100
10
1
CPU Time (sec)
1
2
3
4
Number of Samples
MAGOS
SciClone
PyClone

Table 1

		NUMBER OF SUBCLONES IN A TUMOR						
		1	**2**	**3**	**4**	**5**	**6**	Subtotal
TUMOR STAGE	**I**	0	25	66	62	13	0	166 (50.2%)
	II	1	7	38	28	6	0	80 (24.2%)
	III	0	9	35	30	6	1	81 (24.5%)
	IV	0	1	1	2	0	0	4 (1.2%)
	subtotal	1 (0.3%)	42 (12.7%)	140 (42.3%)	122 (36.9%)	25 (7.6%)	1 (0.3%)	