# Effective clustering for Single Cell Sequencing cancer data

Simone Ciccolella[1,†], Murray Patterson[1,*,†], Paola Bonizzoni[1] and Gianluca Della Vedova[1]

[1] Department of Informatics, Systems, and Communication,
University of Milano-Bicocca, Milan, Italy
[†] Joint first author
[*] Corresponding author: murray.patterson@unimib.it

## Abstract

Single Cell Sequencing (SCS) technologies provide a level of resolution that makes it indispensable for inferring from a sequenced tumor, evolutionary trees or phylogenies representing an accumulation of cancerous mutations. A drawback of SCS is elevated false negative and missing value rates, resulting in a large space of possible solutions, which in turn makes infeasible using some approaches and tools. While this has not inhibited the development of methods for inferring phylogenies from SCS data, the continuing increase in size and resolution of these data begin to put a strain on such methods.

One possible solution is to reduce the size of an SCS instance — usually represented as a matrix of presence, absence and missing values of the mutations found in the different sequenced cells. Previous approaches have used $k$-means to this end, clustering groups of mutations and/or cells, and using these means as the reduced instance. Such an approach typically uses the Euclidean distance for computing means. However, since the values in these matrices are of a *categorical* nature, we explore techniques for clustering categorical data — commonly used in data mining and machine learning — to SCS data, with this goal in mind.

In this work, we explore such categorical clustering techniques, and show on a study of simulated cancer phylogenies that the *$k$-modes* technique is the most effective strategy, when coupled with our novel *dissimilarity* measure — called the *conflict* dissimilarity measure — for computing the resulting modes, or centroids. We demonstrate this by showing that $k$-modes clusters mutations with high precision: never pairing too many mutations that are unrelated in the ground truth, but also obtains accurate results in terms of the phylogeny inferred downstream from a method, when applied to the resulting reduced instance produced by $k$-modes. Finally, we apply the entire pipeline (clustering + inference method) to a real dataset which was previously too large for the inference method alone, showing that our clustering procedure is effective in reducing the running time, hence raising considerably the threshold on the instance size that can be solved.

**Availability:** Our approach, Celluloid: *clustering single cell sequencing data around centroids*, which uses $k$-modes along with our conflict dissimilarity measure, is available at `https://github.com/AlgoLab/celluloid/` under an MIT license.

# 1 Introduction

A tumor, at the time of detection, usually by performing a biopsy on the extracted tissue, is the result of a tumultuous evolutionary process, originating from a single tumor cell — the founder cell [21] — that has acquired a *driver* mutation, which inhibits control on the proliferation of subsequent cancer cells. From that moment, the combination of unrestrained proliferation and a very hostile environment — as immune system fights for survival, and so the tumor cells, under extreme selection pressure, have to disguise themselves to avoid being attacked, compete with each other, all while having to thrive while getting low levels of oxygen — produces the accumulation of highly elevated number of mutations, including structural variations. This model of tumor evolution is called the clonal model [21], since a clone is a population of cells carrying the same set of mutations. Understanding this clonal evolutionary history can help clinicians understand the cell heterogeneity in the tumors of various types of cancer and, more importantly, gives insights on how to devise therapeutic strategies [20, 25].

The rise of *next-generation sequencing* (NGS) technologies has led to the computational problem of tumor phylogeny inference from NGS bulk sequencing data [6, 17, 23, 9]. This idea is very cost-effective, since NGS data is cheap to obtain and very reliable. The procedure consists of extracting different samples of the tumor and of aligning the NGS reads against the reference genome: this allows to determine the approximate fraction of reads of each sample that are affected by any given mutation. This fraction is taken as a proxy of the fraction of cells in each sample that are affected by that mutation. The main difficulties of this technique are the fact that a sample contains a mix of both healthy cells and cancer cells, while the cancer cells are an unknown mixture of different clones.

Since both the composition of the samples and the evolutionary history are unknown in this model, most of the approaches have needed some simplifying assumption, such as the *infinite sites assumption* (ISA) that postulates that each mutation is acquired exactly once, and never lost, in the entire tree. While this assumption has only limited biological validity [15, 2] it reduces greatly the space of all possible solutions, making feasible several approaches [6, 9], which are at least partially inspired by a classical linear time algorithm [8] for reconstructing the phylogeny from noise-free character data: this latter computational problem is called the *perfect phylogeny* problem.

The newer *single cell sequencing* (SCS) technology provides a much finer level of resolution: in fact we can determine whether or not a given cell *has* a mutation, therefore avoiding the notion of sample and the approximations implied by the use of samples. Still, SCS is expensive, and is plagued by high dropout, *i.e.*, missing values, and false negative rates — the latter problem has greater consequences from a computational point of view.

Nowadays, we are witnessing a decrease in SCS costs, coupled with improvements in the dropout and false negative rates, stimulating the research on tools for tumor evolution inference from SCS data [13, 22, 3, 4, 26, 5]. We believe that this line of research is going to become even more important in the next few years, since currently available SCS data is associated with a very large solution space — the high missing value and false negative rates allow a huge number of possible phylogenies — with near optimal values of the objective function: this fact makes difficult to determine which methods actually produce better solutions.

These advances in costs and quality of the data produced will result in larger, but more constrained, instances with a comparatively smaller number of likely solutions. Since most of the currently available methods do not scale well to large instances (usually their running time is quadratic with respect to the number of mutations), a preprocessing step can be useful to reduce the size of an instance: for example, SPhyR [5] uses $k$-means [18, 1] to such purpose. However, $k$-means is designed for continuous data — where means are usually based on a Euclidean distance

— while SCS data, specifying the presence (1) or absence (0) of a mutation in cell, or the fact that it is missing (2), can be thought of as categorical.

Clustering categorical data is an active field of research in data mining [14], where massive databases of categorical data are handled, *e.g.*, finding groups of members of an insurance policy who regularly travel overseas.

In this paper we analyze some clustering algorithms, and compare them experimentally on SCS data. Our comparison is twofold: we first compare the quality of the clusters computed, and then the effect of the clustering procedure when coupled with a tumor phylogeny inference tool (SASC [3]).

# 2 Methods

We now give an overview of the methods we consider for clustering mutations (columns) of single-cell datasets. While a full exposition of all methods we consider is reserved — for lack of space — for the full version of the paper, we consider $k$-modes (Section 2.2) because it performed the best of all other other methods for our purpose, while $k$-means (Section 2.1) serves as a baseline for the most commonly used method — and it has even been used in this context [5].

## 2.1 $k$-means

Given $m$ objects $X = \{\overline{x}_1, \ldots, \overline{x}_m\}$ on $n$ real values, *i.e.*, each $|\{\overline{x}_j\}| = n$, the $k$-means algorithm [18, 1] finds the vector of $k$ values, called *means*, $Q = \{\overline{q}_1, \ldots, \overline{q}_k\}$ minimizing the cost function

$$\sum_{l=1}^{k} \sum_{j=1}^{m} p_{jl} d(\overline{x}_j, \overline{q}_l), \qquad \text{where} \qquad d(\overline{x}_j, \overline{q}_l) = \sum_{i=1}^{n} (\overline{x}_{ji} - \overline{q}_{li})^2, \tag{1}$$

and $p_{jl}$ is an element of an $m \times k$ *partition matrix* $P$ [10]. Here, $d$ is a *distance measure* that is usually the Euclidean distance. The $k$-means algorithm starts with some initial set $Q^0$ of $k$ means, and an initial collection $\mathbb{C}^0$ of $k$ disjoint subsets of $X$, and then iterates the operations

1. compute $d(\overline{x}_j, \overline{q}_l)$, where $\overline{q}_l$ is the mean of cluster $C_l^0 \in \mathbb{C}$, for each $\overline{x}_j, C_l^0$;
2. allocate $\overline{x}_j$ to the cluster $C_{l'}^1$ minimizing $d(\overline{x}_j, q_{l'})$; and
3. recompute the new set of means $Q^1$ according to the new clusters $\mathbb{C}^1$, minimizing Equation 1;

until convergence, *i.e.*, $Q^{t+1} = Q^t$. This algorithm runs in time $O(tkmn)$, where $t$ is the number of iterations of the above set of operations. It has been shown to converge in [18].

We implemented the $k$-means algorithm used here — see `https://github.com/AlgoLab/celluloid/`

## 2.2 $k$-modes

Given $m$ objects $X$ on $n$ *categorical* attributes $A = \{\overline{a}_1, \ldots, \overline{a}_n\}$, the $k$-modes algorithm [11, 12] finds $k$ *modes* $Q$ — that is, for each $\overline{q}_l \in Q$, $\overline{q}_{li}$ is one of the $\overline{a}_i$ possible categories at position $i$. Note that every $\overline{a}_i = \{0, 1, 2\}$ in our case. Instead of the Euclidean distance between object $\overline{x}_j$ and (mode) $\overline{q}_l$ we compute a *dissimilarity*

$$d_1(\overline{x}_j, \overline{q}_l) = \sum_{i=1}^{n} \delta(\overline{x}_{ji}, \overline{q}_{li}). \tag{2}$$

according to some dissimilarity measure $\delta$. A common measure is the simple *matching dissimilarity*

$$\delta_M(\overline{x}_{ji}, \overline{q}_{li}) = \begin{cases} 0 & \text{if } \overline{x}_{ji} = \overline{q}_{li} \\ 1 & \text{o.w. } (\overline{x}_{ji} \neq \overline{q}_{li}) \end{cases} . \tag{3}$$

Since a 2 represents *missing* data in our context, we slightly relax the matching dissimilarity, to give what we call the *conflict dissimilarity*

$$\delta_C(\overline{x}_{ji}, \overline{q}_{li}) = \begin{cases} 0 & \text{if } \overline{x}_{ji} = 2 \text{ or } \overline{q}_{li} = 2 \\ \delta_M(\overline{x}_{ji}, \overline{q}_{li}) & \text{o.w.} \end{cases} . \tag{4}$$

Now, the *mode* of cluster $C_l \subseteq X$ is a vector $\overline{q}_l$ which minimizes

$$D(C_l, \overline{q}_l) = \sum_{\overline{x}_j \in C_l} d_1(\overline{x}_j, \overline{q}_l). \tag{5}$$

Note that $\overline{q}_l$ is not necessarily an element of $C_l$. The $k$-modes algorithm then starts with some initial set $Q^0$ of $k$ modes, and an initial collection $\mathbb{C}^0$ of $k$ disjoint subsets of $X$, and then iterates operations 1–3 as in the $k$-means algorithm, but using *dissimilarity* instead of distance, and computing *modes* instead of means, while minimizing instead

$$\sum_{l=1}^{k} \sum_{j=1}^{m} \sum_{i=1}^{n} p_{jl} \delta(x_{ji}, q_{li}). \tag{6}$$

by finding modes according to the following theorem, where $c_{ri}$ is the $r$-th category of attribute $\overline{a}_i$ and $f(\overline{a}_i = c_{ri}|X)$ is its relative frequency in $X$:

**Theorem 1 ([12])** *Eq. 5 is minimized iff* $f(\overline{a}_i = q_{li}|X) \geq f(\overline{a}_i = c_{ri}|X)$ *for* $q_{li} \neq c_{ri}$ $\forall\ i \in [n]$.

In other words, Eq. 5 is minimized by selecting the mode category for each attribute. Note that this theorem implies that the mode of $X$ is not unique.

Since the solution depends on the initial set $Q^0$ of $k$ modes, we consider two procedures for initializing $Q^0$. The first one is quite simple: a random selection of $k$ objects from the set $X$ of objects as the initial $k$ modes — which we refer to as *random initialization*. The second one, devised in [12], is a more complicated procedure, based on the frequencies $f(\overline{a}_i = c_{ri}|X)$ of all categories, which we refer to as *Huang initialization*. This second procedure is as follows:

1. order the categories of each attribute $\overline{a}_i$ in descending order of frequency, *i.e.*, $f(c_{r_1 i}) \geq f(c_{r_2 i}) \geq f(c_{r_3 i})$, etc.;
2. assign uniformly at random the most frequent categories to the initial $k$ modes; and
3. for each $\overline{q}_l$, select the $\overline{x}_j \in X$ most similar to $\overline{q}_l$ and make this the mode $\overline{q}_l^0 \in Q^0$, such that $\overline{q}_l^0 \neq \overline{q}_{l'}^0$ for $l \neq l'$.

This final step is to avoid empty clusters in $\mathbb{C}^0$. This initial selection procedure is to have a diverse set of initial modes $Q^0$, which can lead to better clustering results — see more details in [12].

The *Python Package Index* (pypi) has a package for computing a clustering based on $k$-modes `https://pypi.org/project/kmodes/`, including an implementation of Huang (and random) initialization. Since the dissimilarity measure used to infer the $k$-modes is customizable, we implemented our novel conflict dissimilarity, and used it with this software.

# 3 Results

To evaluate the accuracy of the clustering methods we designed a two-fold experimentation on synthetic datasets: we first measure the quality of the clusters found by the methods, and later we evaluate how such clusters impact the quality of the phylogenies returned by a cancer progression inference method.

The simulated data are generated as follows. First we simulate a random tree topology on $s$ nodes, each representing a tumor clone, by first creating a root (the germline) and then iteratively attaching the $s-1$ remaining nodes uniformly at random to any other node in the tree. The nodes are then randomly labeled with $m$ mutations — meaning that each mutation is acquired at the node that it labels. Then, a total of $n$ cells are associated to the nodes, uniformly at random. A binary matrix $M$ is then extracted from these cells giving rise to a genotype profile for each cell (a row in $M$), which is the presence (1) or absence (0) of each mutation (column in $M$) in the cell, given the the presence or absence of the mutations on the path in three from the root to this cell. The binary matrix is then perturbed according to the false negative, false positive and missing value rates, to simulate a real-case scenario. Each of the $s$ nodes is therefore considered as a natural (true) cluster of the simulated dataset.

For each experiment, we consider 100, 200 and 300 cells, respectively Experiment 1, 2, and 3. For each such value we generated 50 simulated datasets, where we fixed the number $s$ of clones to 20 and the number $m$ of mutations to 1000. While this number of mutations is at the high end in terms of currently available real cases, it will be a typical size in the near future — some such cases already exist today (see, *e.g.*, Section 3.3). Such a target already causes difficulties for the current state-of-the-art cancer inference methods which employ a more general model for SCS data [3, 4, 26, 5].

We performed clustering on the datasets of our experiments to obtain instances with a reduced number of columns (mutations), which can in turn be given as input to such a method above. The clustering methods we used were $k$-means, as well as $k$-modes with all combinations of both Huang and random initialization procedures, and both the matching dissimilarity and our conflict dissimilarity measures. Since the cancer inference methods tend to scale quadratically with the number of mutations, we reduced with all clustering methods the size of mutations to a total of 100, which is a reasonable number of mutations in the currently available literature — and what the current methods can handle.

## 3.1 Evaluating a clustering

To evaluate the quality of the clusters, we used standard precision and recall measures, adapted to the particular goal, as follows.

**Precision:** measures how well mutations are clustered together. For each pair of mutations appearing in the same clone in the simulated tree, we check if they are in the same cluster, resulting in a true positive ($TP$). For each pair of mutations clustered together that are not in the same clone, we encounter a false positive ($FP$). The value of the precision is then calculated with the standard formula: $\frac{TP}{TP+FP}$.

**Recall:** measures how well mutations are separated. For each pair of mutations in the same clone, we now also check if they are not in the same cluster, resulting in a false negative ($FN$). The recall is then calculated as: $\frac{TP}{TP+FN}$.

It is important to highlight that we are mostly interested in obtaining a high *precision* since, while cancer phylogeny inference algorithms can later cluster together mutations by assigning them

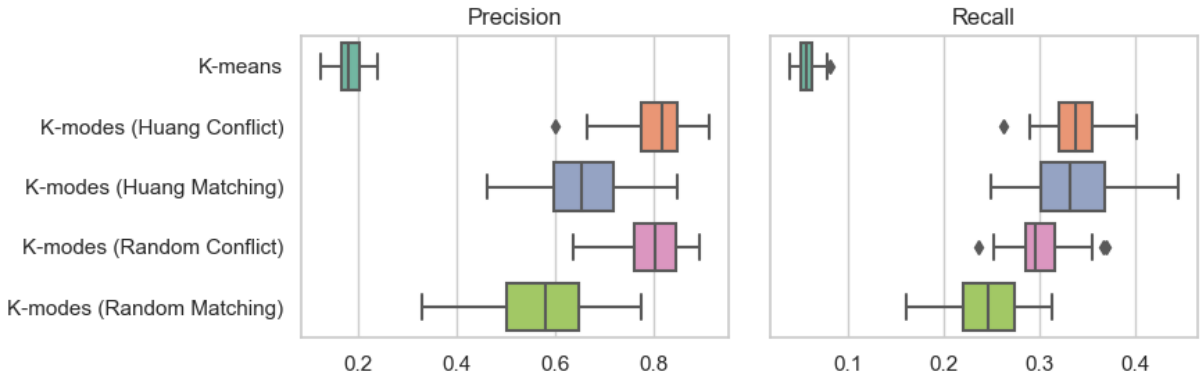


Figure 1: Precision and recall results for experiment 1, generated with a total of 1000 mutations, 100 cells and a clustering size of $k = 100$. The plots include results for $k$-means and $k$-modes clustering with Huang and random initialization procedures and conflict and matching dissimilarities.

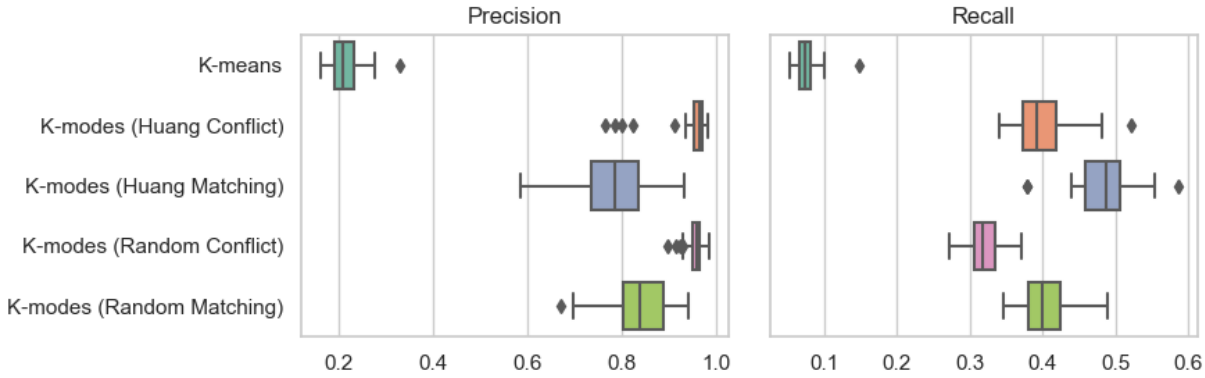


Figure 2: Precision and recall results for experiment 2, generated with a total of 1000 mutations, 200 cells and a clustering size of $k = 100$. The plots include results for $k$-means and $k$-modes clustering with Huang and random initialization procedures and conflict and matching dissimilarities.

to same subtree or the same path, they cannot separate mutations that have been erroneously clustered together.

In Figures 1,2 and 3 — representing the respective experiments 1,2 and 3 — a common trend is evident: indeed $k$-means clustering performs much more poorly than the others, presenting a notable gap from all the other methods in terms of both precision and recall. With $k$-modes, when coupled with our conflict dissimilarity measure, the initialization procedures (Huang and random) differ slightly in terms of precision, while in terms of recall, Huang initialization performs slightly better in most cases. On the other hand the matching dissimilarity is lacking in every experiment in terms of precision, but it tends to give slightly better recall in experiments 2 and 3.

It is interesting to notice that the precision of the conflict dissimilarity rapidly increases when the amount of cells increase, thus being well-suited for future increases on the size of SCS datasets.

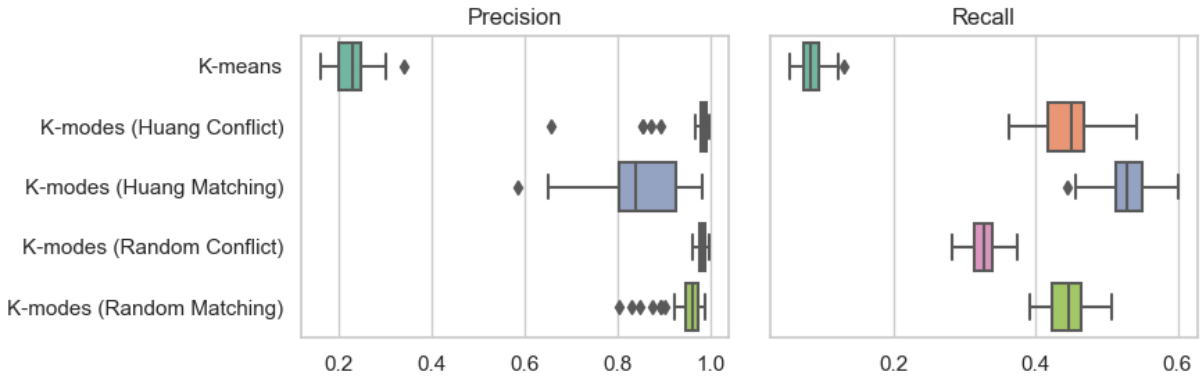


Figure 3: Precision and recall results for experiment 3, generated with a total of 1000 mutations, 300 cells and a clustering size of $k = 100$. The plots include results for $k$-means and $k$-modes clustering with Huang and random initialization procedures and conflict and matching dissimilarities.

### 3.2 Assessing the impact of a clustering

To better understand the impact of the clustering on the actual cancer progression inference, we selected three clustering algorithms and executed SASC [3] on the obtained clusters. Since the most important aspect of the clustering for this application is to have a high precision, we selected $k$-modes with both the random and the Huang initialization procedures, coupled with our conflict dissimilarity measure. We also included $k$-means in the evaluation as a baseline for the most commonly used method. The other techniques were discarded because of their lack of precision during the clustering evaluation.

For assessing the accuracy of the methods we utilized the measures defined in [4, 3]:

**Ancestor-Descendant accuracy:** for each pair of mutations in an ancestor-descendant relationship in the ground truth, we check whether the relationship is conserved in the inferred tree ($TP$) or whether it is not ($FN$). For each pair of mutations in an ancestor-descendant relationship in the inferred tree, we also check if such relationship does not exist in the ground truth tree ($FP$).

**Different lineages accuracy:** similarly to the previous measure, we check whether mutations in different branches are correctly inferred or if any pair of mutation is erroneously inferred in different branches.

To conduct the second part of the experimentation we selected the cancer inference tool SASC, since the objective of this work is to evaluate the clustering, and how it can aid such an inference tool in inferring a phylogeny from massive dataset — so massive that it is normally out of reach. The choice of the inference tool is not relevant and any other method would lead to the same conclusions, independently from the definition or complexity of the method itself.

Figures 4,5 and 6 show a clear drop in performance when $k$-means is used as a preprocessing step, while $k$-modes clustering produces very accurate results. This fact is supported by the gap in the precision of the methods — a low precision indeed leads to a low accuracy in the tree reconstruction. The trend is still present in the different lineages accuracy, but to a lower extent — this is because, as previously discussed, a cancer inference method can separate *clusters* of mutations, but when a cluster is computed it is not possible to separate mutations *within* this cluster.

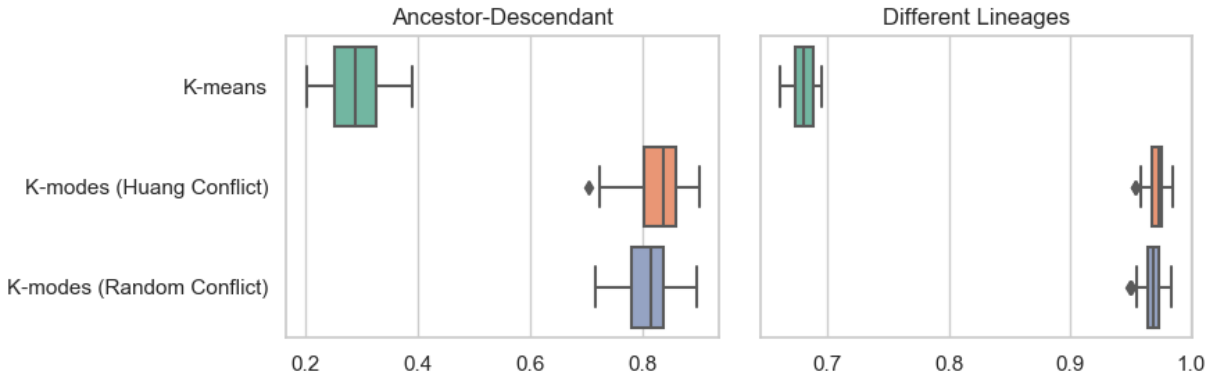


Figure 4: Ancestor-Descendant and Different lineages accuracy measures for experiment 1. Cancer phylogenies are inferred by SASC using as input the clusters obtained by $k$-means, $k$-modes with random and Huang initialization procedures and our conflict dissimilarity measure.

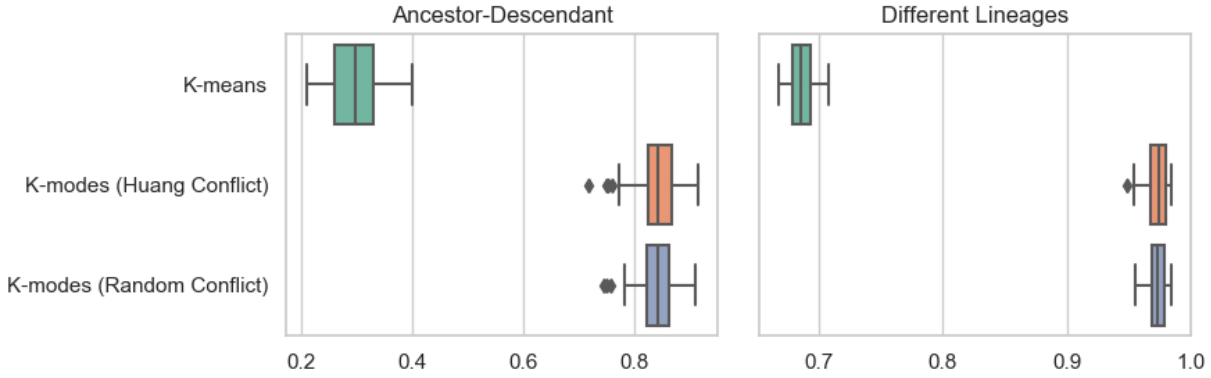


Figure 5: Ancestor-Descendant and Different lineages accuracy measures for experiment 2. Cancer phylogenies are inferred by SASC using as input the clusters obtained by $k$-means, $k$-modes with random and Huang initialization procedures and our conflict dissimilarity measure.

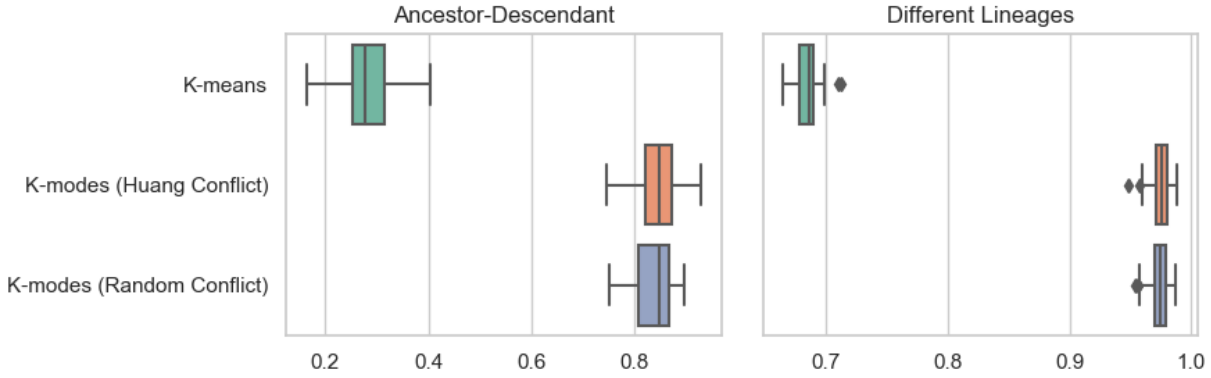


Figure 6: Ancestor-Descendant and Different lineages accuracy measures for experiment 3. Cancer phylogenies are inferred by SASC using as input the clusters obtained by $k$-means, $k$-modes with random and Huang initialization procedures and our conflict dissimilarity measure.

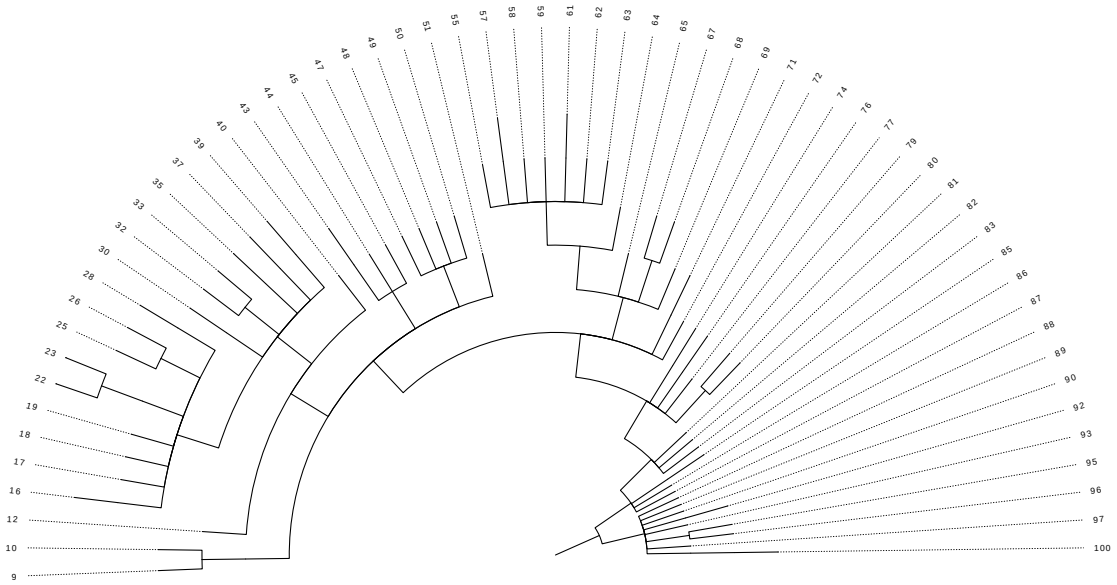


Figure 7: Tree computed on IDH-mutated Glioblastoma from [24]. The labels on the nodes represent the 100 clusters obtained by the $k$-modes clustering with the Huang initialization method and conflict dissimilarity. The tree was inferred using the SASC [3] tool.

### 3.3 Application on real data

Finally, we run our entire pipeline (clustering + inference method) on a Glioblastoma (GBM) IDH-mutated tumor [24] consisting of 1842 mutations over 926 cells. Such computation was previously not possible with the use of SASC alone, due to the running time required — SASC was not able to provide a solution in a time-frame of three weeks. With the combined use of the $k$-modes clustering with the Huang initialization procedure and our conflict dissimilarity (one with best performance in Section 3.1) we were able to compute the tree within 12 hours.

We clustered the 1842 original mutations into 100 clusters of mutations, using the resulting set of 100 modes (or centroids) as the resulting matrix for input to SASC. Figure 7 shows the tree inferred, where each node is one of 100 clusters.

## 4 Discussion

We have analyzed various clustering methods on single cell sequencing data showing that $k$-modes with random and Huang initialization procedures clearly outperforms $k$-means on those data. More precisely, we have run a two-fold experiment on synthetic data: the first experiment measured the quality of the clusters computed by the methods, while the second experiment considered the effect of the clustering on the quality of trees obtained from a phylogeny inference tool. The experiment on real data shows the usefulness of the clustering procedure, as it has reduced the running time of SASC at least 40x.

We conjecture that the major pitfall with the $k$-means approach, is that $k$-means has been designed (and is routinely used) for numeric values, while single cell data is *categorical* data. Indeed, matrices associated with single cell data encode the presence (1), and absence (0) of a mutation (a

column of the matrix), as well as missing values (2), in a cell (a row). A consequence of the meaning of the matrix entries, is that the Euclidean distance between a pair of mutations (columns) has little meaning, hence ruling out clustering methods relying on such a notion of distance.

A direction of further research is to extend the set of clustering methods studied and to develop a new clustering method that is tailored for SCS data. In fact, clustering categorical data is an active field of research in data mining [14], where massive databases of categorical data are handled. Another possibility to explore is conceptual clustering algorithms [19, 7, 16], which have been initially developed in machine learning to reason with categorical data: however their efficiency relies on good search strategies. For SCS data, the search space is so large that decided to take an approach which strikes a good balance between the meaningful results delivered by the above categorical methods, and computational efficiency, namely $k$-modes [11, 12].

The $k$-modes clustering performed very well on SCS data, both in terms of clustering, and the resulting inferred tree by a downstream phylogeny method. However, it was when it was coupled with our conflict dissimilarity measure that it performed the best, rather than the simple matching dissimilarity measure that is usually used for clustering categorical data. We conjecture that it is the meaning of a mutation profile for a cell that makes this difference: the fact that a 2 represents missing data, makes it a special category, needing a special treatment. Namely a mutation profile with many missing values (2's) should not be deemed "far" from any other profile, because such a profile is very uncertain. It is because we do not overly penalize an elevated number of missing values in a profile — something very possible, given that a 25% missing value rate is typical — that we get more meaningful centroids, and respective clusters, as a result in this context. There may be even more suitable dissimilarity measures, however — an interesting future direction.

## Acknowledgments

We thank Iman Hajirasouliha and Dana Silverbush for several illuminating discussions. We acknowledge the support of 2017-ATE-0534 and 2016-ATE-0332 grants.

## References


[1] M.R. Anderberg. *Cluster Analysis for Applications*. Academic Press, 1973.

[2] David Brown, Dominiek Smeets, Borbála Székely, et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nature Communications*, 8:14944, 2017.

[3] Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. Inferring cancer progression from single cell sequencing while allowing loss of mutations. *bioRxiv*, 268243, 2018.

[4] Simone Ciccolella, Mauricio Soto, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. gpps: An ILP-based approach for inferring cancer progression with mutation losses from single cell data. In *IEEE Computational Advances in Bio and medical Sciences, 8th International Conference (ICCABS)*, 2018. to appear.

[5] Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.

[6] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

[7] D.H Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987.

[8] Dan Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.

[9] Iman Hajirasouliha and Benjamin J. Raphael. *Reconstructing Mutational History in Multiply Sampled Tumors Using Perfect Phylogeny Mixtures*, pages 354–367. Lecture Notes in Computer Science. Springer Nature, 2014.

[10] D.J. Hand. *Discrimination and Classification*. John Wiley & Sons, 1981.

[11] Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 1–8, 1997.

[12] Z. Huang. Extensions to the k-modes algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

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

[14] W. Klosgen and J.M. Zytkow. Knowledge discovery in databases terminology. In *Advances in Knowledge Discovery and Data Mining*, pages 573–592. AAAI Press/The MIT Press, 1996.

[15] Jack Kuipers, Katharina Jahn, Benjamin J. Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, 27:1885–1894, 2017.

[16] M Lebowitz. Experiments with incremental concept formation. *Machine Learning*, 2(2):103–138, 1987.

[17] S. Malikic, A.W. McPherson, N. Donmez, and S.C. Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 2015.

[18] J.B. McQueen. Some methods for classification and analysis of multivariate observations. In *the 5th Berkely Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[19] R.S Michalski and R.E. Stepp. Automated construction of classifications: Conceptual clustering versusnumerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(4):396–410, 1983.

[20] A. Sorana Morrissy, Livia Garzia, David J. H. Shih, et al. Divergent clonal selection dominates medulloblastoma at recurrence. *Nature*, 529(7586):351–357, 2015.

[21] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[22] Edith M. Ross and Florian Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17(1):69, 2016.

[23] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):e165, 2013.

[24] Andrew S. Venteicher, Itay Tirosh, Christine Hebert, Keren Yizhak, Cyril Neftel, Mariella G. Filbin, Volker Hovestadt, Leah E. Escalante, McKenzie L. Shaw, Christopher Rodman, Shawn M. Gillespie, Danielle Dionne, Christina C. Luo, Hiranmayi Ravichandran, Ravindra Mylvaganam, Christopher Mount, Maristela L. Onozato, Brian V. Nahed, Hiroaki Wakimoto, William T. Curry, A. John Iafrate, Miguel N. Rivera, Matthew P. Frosch, Todd R. Golub, Priscilla K. Brastianos, Gad Getz, Anoop P. Patel, Michelle Monje, Daniel P. Cahill, Orit Rozenblatt-Rosen, David N. Louis, Bradley E. Bernstein, Aviv Regev, and Mario L. Suvà. Decoupling genetics, lineages, and microenvironment in idh-mutant gliomas by single-cell rna-seq. *Science*, 355(6332), 2017.

[25] Jiguang Wang, Emanuela Cazzato, Erik Ladewig, et al. Clonal evolution of glioblastoma under therapy. *Nature Genetics*, 48(7):768–776, 2016.

[26] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology*, 18(1):178, 2017.