Bayesian non-parametric clustering of single-cell mutation profiles

Nico Borgsmüller[1,2*], Jose Bonet[3,4*], Francesco Marass[1,2], Abel Gonzalez-Perez[3,4], Nuria Lopez-Bigas[3,5], and Niko Beerenwinkel[1,2✉]

[1] Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland
[2] SIB, Swiss Institute of Bioinformatics, 4058 Basel, Switzerland
[3] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain
[4] Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
[5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
* These authors contributed equally to this work
✉ Corresponding author: niko.beerenwinkel@bsse.ethz.ch

Abstract. The high resolution of single-cell DNA sequencing (scDNA-seq) offers great potential to resolve intra-tumor heterogeneity by distinguishing clonal populations based on their mutation profiles. However, the increasing size of scDNA-seq data sets and technical limitations, such as high error rates and a large proportion of missing values, complicate this task, rendering the applicability of existing methods more limited. Here we introduce BnpC, a novel non-parametric method to cluster individual cells into clones and infer their genotypes based on their noisy mutation profiles. BnpC employs a Dirichlet process mixture model coupled with a Markov chain Monte Carlo sampling scheme, including a modified non-conjugate split-merge move and a novel posterior estimator to predict clones and genotypes. Our method was comprehensively benchmarked against state-of-the-art methods on simulated data using various data sizes and was applied to three cancer scDNA-seq data sets. On simulated data, BnpC compared favorably against current methods in terms of accuracy, runtime, and scalability. On tumor scDNA-seq data, BnpC was able to identify clonal populations missed by the original cluster analysis but supported by supplementary experimental data. As scDNA-seq data size constantly grows, scalable, efficient and accurate methods such as BnpC will become increasingly relevant, not only to solve intra-tumor heterogeneity, but also as a pre-processing step to reduce data size. BnpC is freely available under MIT license at https://github.com/cbg-ethz/BnpC.

Keywords: single-cell DNA sequencing · clustering · genotyping · non-parametric Bayes · Chinese Restaurant Process · Dirichlet Process Mixture Model · Markov Chain Monte Carlo

1 Introduction

Cancer is an evolutionary process characterized by the accumulation of mutations that drive tumor initiation, progression, and treatment resistance [1]. The underlying evolutionary process can follow different modes but ultimately leads to multiple coexisting cell populations differing in their genotype [2, 3]. This genomic heterogeneity, also known as intra-tumor heterogeneity (ITH), poses major challenges for cancer treatment as treatment-surviving subpopulations are likely to cause cancer recurrence [**burrell2013**, 4]. Therefore, it is beneficial to identify the clonal composition of the tumor and to adapt the treatment accordingly.

Recent advances in the field of single-cell DNA sequencing (scDNA-seq) have led to new insights into cancer evolution and ITH. Examples include the detection of rare subclones in breast cancer patients [5], the identification of new cell types like disease-associated microglia [6], and major advancements in the reconstruction of cancer evolution [7]. Compared to bulk sequencing, scDNA-seq offers the possibility to directly access clonal genotypes at the cellular level and to more easily detect branching in clonal evolution. However, scDNA-seq data tends to be very noisy. DNA amplification through processes like multiple displacement amplification as well as mutation calling algorithms introduce a large fraction of errors and missing values [8]. Errors can be either true mutations not identified, namely false negatives (FN), or mutations not present in a cell but falsely reported, namely false positives (FP). Characteristic of scDNA-seq data are high FN rates, arising from technical failure to measure both alleles at a mutated locus, and a large fraction of missing values, resulting from non-uniform coverage and drop-outs.

Since the inception of single-cell sequencing, scDNA-seq data has been used to infer the clonal composition and the underlying evolutionary history of tumor samples based on mutation profiles, i.e. the absence or presence of called mutations in each cell. Generic clustering algorithms, such as centroid- or density-based methods, do not account for these characteristics and are therefore unsuitable for scDNA-seq data.Hence, a variety of tailored methods was introduced recently, varying in the main objective, model choice, and inference. The probabilistic framework BitPhylogeny [9] and the nested effects model OncoNEM [10] jointly cluster cells into clones and infer their phylogenetic relations, for which the former employs a Markov Chain Monte Carlo (MCMC) scheme and the latter a heuristic search algorithm. SCITE [11] employs a Bayesian model to infer mutation or cell lineage trees via MCMC. The methods above operate under the infinite sites assumption, which states that mutations in the lineage can only be acquired once and never be lost. On scDNA-seq data, this is a simplifying assumption which may not be supported by the data [12]. Recently, several phylogenetic tree inference tools not make this assumption were published, such as SiFit [13], SiCloneFit[14], PhISCS [15], SPhyR [16], and SASC [17]. SPhyR and SASC operate under the Dollo parsimony model [18] which postulates that mutations can only be acquired once but lost several times. SiFit, SiCloneFit, and PhISCS follow the finite sites assumption, allowing mutations to be acquired and lost more than once. These approaches focusing focus on resolving the phylogenetic relationship among cells and in doing so can also provide clusters and genotypes. However, due to the inherent tree search task, these approaches are computationally expensive limiting their scalability to larger data sets. Currently, the only method focused on clustering and genotyping scDNA-seq data is SCG (Single Cell Genotyper) [19], which uses a parametric model and applies variational inference to learn the genotypes and clonal composition. To determine the unknown number of clusters, the inference is run several times with varying cluster numbers and the best result is reported. This approach is computationally efficient but the parametric nature of the model and the variational inference limit the robustness of its predictions.

Here, we introduce BnpC, a fully Bayesian method to determine the clonal composition and genotypes of scDNA-seq data. BnpC clusters individual cells into clones based on their noisy mutation profiles. For the estimation of the unknown number of clusters, we employ a representation of the Dirichlet Process mixture model, the Chinese restaurant process (CRP). We benchmark our approach against state-of-the-art methods on simulated data and demonstrate that BnpC performs similarly or outperforms current approaches. In the analysis of scDNA-seq cancer data, BnpC provides more detailed insights into the clonal evolution of individual patients by identifying additional clonal populations. These additional clones were confirmed in the original analysis by independent experimental copy number data but were not identified from the scDNA-seq data alone. In addition, we demonstrate that without applying any data pre-processing, BnpC reproduces results from a previous analysis, which included a laborious manual filtering step.

2 Methods

2.1 Model

BnpC takes as input a binary matrix with missing values $\boldsymbol{X} = (x_{ij}) \in \{0, 1, -\}^{N \times M}$ of N cells and M mutations, where 0 indicates the absence of a mutation, 1 its presence, and $-$ a missing value (Fig. 1 A). We assume that the N cells were sampled from an unknown number K of clones, each with a distinct mutation profile $\boldsymbol{\theta}_k \in [0, 1]^M$, coming from a prior distribution G_0. To model the cluster assignments, we use a Chinese Restaurant Process (CRP) [20]. The CRP is a probability distribution over partitions of the natural numbers, which in this context serves as a prior for clusterings. The prior probability of assigning a cell to a clone is given by

$$\begin{aligned} p(c_i = k|\alpha_0; c_1, \ldots, c_{i-1}) &= \frac{n_{k,-i}}{N - 1 + \alpha_0} \\ p(c_i = k_{+1}|\alpha_0; c_1, \ldots, c_{i-1}) &= \frac{\alpha_0}{N - 1 + \alpha_0} \end{aligned} \tag{1}$$

where c_i is the clonal assignment of cell i, $n_{k,-i}$ the number of cells assigned to clone k excluding cell i, k_{+1} is an unoccupied clone, and α_0 is the CRP concentration parameter. Finally, the probabilities of observing a FP or FN in the data matrix $\mathbf{X}$ are given by the parameters α and β, respectively.

With the parameters described above, we can formulate the likelihood of BnpC as

$$P(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{c}, \alpha, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{M} \theta_{c_i,j} \Big[(1-\beta)^{x_{i,j}} \cdot \beta^{1-x_{i,j}}\Big] + (1 - \theta_{c_i,j}) \Big[(1-\alpha)^{1-x_{i,j}} \cdot \alpha^{x_{i,j}}\Big] \tag{2}$$

where the first term accounts for the presence of a mutation in a clone and the observation of a true positive or FN, and the second term accounts for the absence of a mutation in a clone and the observation of a true negative or a FP. Missing values are skipped.

The posterior distribution over the latent variables is

$$P(\boldsymbol{\theta}, \boldsymbol{c}, \alpha, \beta, \alpha_0|\boldsymbol{X}, \boldsymbol{H}) \propto P(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{c}, \alpha, \beta)\ P(\boldsymbol{\theta}, \boldsymbol{c}, \alpha, \beta, \alpha_0|\boldsymbol{H}) \tag{3}$$

where $\boldsymbol{H}$ is the set of fixed hyperparameters $\boldsymbol{H} = \{G_0, \mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta\}$.

Following the graphical model (Fig. 1 B), the prior probabilities can be factored as

$$P(\boldsymbol{\theta}, \boldsymbol{c}, \alpha, \beta, \alpha_0|\boldsymbol{H}) = P(\boldsymbol{\theta}|G_0) P(\boldsymbol{c}|\alpha_0) P(\alpha_0) P(\alpha|\mu_\alpha, \sigma_\alpha) P(\beta|\mu_\beta, \sigma_\beta). \tag{4}$$

2.2 Inference

As the posterior distribution is not analytically tractable, we employed a Markov chain Monte Carlo (MCMC) sampling scheme to obtain samples from the posterior distribution. Cluster parameters and error rates are updated via Metropolis-Hastings algorithm; the concentration parameter α_0 is learned as described in [21]; cluster assignments are updated with Gibbs sampling and a modified non-conjugate split-merge move [22, 23].

The split-merge move introduced by Jain and Neal [24] samples two observations uniformly. If both observations are drawn from the same cluster, it proposes to split that cluster, otherwise it proposes to merge the two clusters. In our model, observations correspond to cells and clusters to clones. We modified this procedure by first deciding which move to perform. If a merge move is selected, two cells are drawn from two different clones, themselves chosen in a manner inversely proportional to their size. This increases the probability of merging spurious clusters. For a split move, two random cells are sampled from a clone chosen proportionally to its size. The definition of fixed split-merge rates and the incorporation of clone size into the clone selection allow us to adapt the split-merge move to the posterior landscape and to circumvent bottlenecks of low probability regions in a reasonable time.

We account for these changes by extending the proposals ratio in Metropolis-Hastings update as follows. For a split move, we introduce the ratio

$$\frac{\tilde{K}_i \tilde{K}_j}{(\sum_l \tilde{K}_l)^2} \cdot \left(\frac{|K|}{N} \binom{|K|}{2}^{-1} \right)^{-1}. \tag{5}$$

The second term describes the probability of choosing cluster K (of size $|K|$) according to its size, and choosing two cells (i and j) from it. After the split, let K_i and K_j denote the two different clusters to which i and j belong. Let $\tilde{K_i} = \left(\frac{|K_i|}{N}\right)^{-1}$ represent the inverse cluster size of the cluster with cell i. The first term in Eq. 5 denotes the probability of choosing the clusters with cells i and j to reverse the split move.

Similarly, for a merge move we extend the Metropolis-Hastings ratio with the following factor:

$$\frac{|K|}{N}\frac{1}{|K|(|K|-1)} \cdot \left(\frac{\tilde{K_i}\tilde{K_j}}{(\sum_l \tilde{K_l})^2}\frac{1}{|K_i||K_j|}\right)^{-1}. \tag{6}$$

Here, the second term accounts for choosing two distinct clusters in a manner inversely proportional to their size, and then two cells i and j uniformly from each cluster. The first term undoes the merge move by selecting the merged cluster K according to its size, and selecting cells i and j from it.

2.3 Estimators

Downstream analyses and interpretation generally require a single clustering and genotypes for all cells. There are several options to summarise the posterior distribution and obtain such information. A simplistic approach is to use the maximum likelihood (ML) or maximum a posteriori (MAP) point estimators, which return the model parameters that achieved the highest likelihood or posterior probability. These point estimators are easy to use but ignore the shape of the posterior distribution. In contrast, the MPEAR [25] estimator operates on the posterior similarity matrix, which contains the posterior co-occurrences for every pair of observations. To identify the final number of clones, the hierarchical clustering of this matrix is cut according to the posterior expected adjusted Rand index (ARI).

We followed a similar approach by using the posterior similarity matrix for agglomerative clustering. However, instead of using an additional metric like the ARI to predict the optimal number of clones, we leveraged the non-parametric nature of our model and selected the average number of clones over all posterior samples. The genotypes were subsequently inferred independently for each clone from a selected subset of posterior samples. For each clone, we selected posterior samples based on two criteria: *(1)* all cells assigned the clone are clustered together; *(2)* no other cell is clustered with these cells. If no posterior sample fulfills both criteria, we selected samples fulfilling only criteria 1. The final genotype is the rounded mean over the corresponding cluster parameters from the selected posterior samples.

For clustering, we evaluated our estimator against the MPEAR, ML, and MAP estimators on simulated data (Fig. S5). The inferred clustering measured by the V-Measure varied only slightly between the estimators, not favoring any significantly. Because MPEAR only predicts a clustering, it was excluded from the genotype comparisons. The inferred genotypes, evaluated by the Hamming distance between true and inferred genotypes of each cell, indicate that our novel estimator predicts genotypes more accurately than point estimators (Fig. S6). In all simulated cases, our novel estimator achieved better genotyping than the point estimators, independently of the error rates and fraction of missing values. All estimators showed the same tendencies regarding varying error rates and missing value rates, with higher rates resulting in less accurate clustering/genotyping, and vice versa.

3 Results

3.1 Benchmarking on simulated data

To benchmark BnpC against state-of-the-art methods, we generated two synthetic data sets differing in size and clone number. For the first, we simulated data consisting of 50 mutations and 50 cells, forming 5 distinct clones (data set: 50×50). For the second, we generated data of 50 mutations and 200 cells, forming 10 clones (data set: 50×200). For all three data sets, we varied the evolutionary history, error rates, and the number of missing values. A description of the simulation process is provided in the supplementary data – Simulations section. We benchmarked BnpC against OncoNEM [10], SCG [19] and SiCloneFit [26]. SiFit and SCITE were not considered as they only infer the phylogenetic relation, not the clonal composition or genotypes. BitPhylogeny follows an approach similar to OncoNEM but was shown to produce less accurate results;

hence, we excluded it from our benchmarking. Clustering accuracy was evaluated using the V-Measure [27]. Genotyping errors were measured by the Hamming distance between the predicted cellular genotypes and the true ones. SCG was run with 10^9 steps, BnpC with 5000, SiCloneFit with 500, and OncoNEM with 2000. Step numbers for SiCloneFit and OncoNEM were selected so as to not exceed one hour on the 50×200 data set. For SCG and BnpC we assumed convergence after the selected number of steps: performance did not improve with greater number of steps. We provided the true error rates to OncoNEM as it does not learn them, the other algorithms were run with their default priors on error rates.

On the 50×50 data set, OncoNEM showed the lowest clustering accuracy, independently of error or missing value rates (Fig. S2). Because of OncoNEM's low clustering accuracy, despite being run with the true error rates, and its long runtime on the smallest data set, we excluded it from further benchmarking analysis on larger data sets.

Figure 2 shows the V-measure and genotyping error on the 50×200 data set. In general, all three methods performed similarly. SCG predicted clusters with highest accuracy, BnpC and SiCloneFit predicted them slightly less accurately. On high FN and missing value rates, SiCloneFit outperformed BnpC, but SiCloneFit performed worst on the data with varying FP. For genotyping, BnpC outperformed the other methods, except for high FN and missing value rates where SiCloneFit performed best. Both SCG and SiCloneFit showed higher variance in their results than BnpC. At a large fraction of FP, SCG's genotyping performance dropped eminently.

The same trends between algorithms were observed on the different simulated evolutionary histories. Unsurprisingly, the simulation of different evolutionary histories showed that frequent and early branching events, resulting in clones with highly diverse mutation profiles, lead to a higher clustering accuracy of all methods than linear evolution and late branching events (Fig. S3).

In the previous analysis, SCG ran shortest, BnpC second shortest, and SiCloneFit longest. To access the scalability and runtime of the algorithms, we simulated two larger data sets, one with 1000 cells, 200 mutations, and 20 clones (data set: 200×1000) and one with 3000 cells, 200 mutations, and 30 clones (data set: 200×3000). FN, FP, and missing value rates were fixed at 0.3, 0.01, and 0.2, respectively. The evolutionary history was also fixed (minimal trunk size of 0.1 and mutation rate of 0.25). On these data sets, we ran BnpC, SCG, and SiCloneFit with different numbers of steps. As the computational time per step varies highly between algorithms, the number of steps was chosen differently for each algorithm to cover runtimes from approximately 0.5 hours to a maximum time limit of 10 hours. The selected number of steps are shown in Fig. 3. Algorithms were run on a high performance computing cluster, each algorithm ran on a single core with 10 GiB memory and 2.4 GHz CPU.

On the 200×1000 data set, SCG had the shortest runtime, independent of step size. However, clustering accuracy varied between runs and the genotyping error was high in comparison to the other algorithms. SiCloneFit performed less accurate clustering, even when ran longer than SCG and BnpC, and showed better genotyping than SCG but worse than BnpC (Fig. 3). BnpC performed the most accurate clustering and genotyping, with less variance in the clustering accuracy than SCG, independent of the runtime, which was longer in comparison to SCG and shorter compared to SiCloneFit. On the 200×3000 data set, SCG and BnpC showed similarly high clustering accuracy, but SCG predicted the least accurate genotypes. Again, SCG ran in the shortest amount of time, but BnpC run with a small number of steps achieved similar clustering but better genotyping in a similar runtime. SiCloneFit results are only displayed for two different number of steps as none of the SiCloneFit runs with 100 steps produced a result within the 10 hours time limit. Again, SiCloneFit performed poorly in clustering but outperformed SCG in genotyping. BnpC achieved similar clustering accuracies than SCG and outperformed the other algorithms in the genotyping, independent of runtime.

3.2 Application to real data

To demonstrate the application of BnpC on real tumor data, we analyzed the sequencing data of five patients with childhood leukemia [28], one high-grade serous ovarian cancer (HGSOC) patient [29] and two colorectal cancer (CRC) patients [30].

Acute Lymphoblastic Leukemia We reanalyzed scDNA-seq data of five Acute Lymphoblastic Leukemia (ALL) patients [28]. The data contains between 16 and 105 mutations and between 96 and 143 cells per

patient. Gawad *et al.* used a combination of a multivariate Bernoulli model and the Jaccard distance to predict the clonal composition and to infer genotypes. Inferred genotypes and clones by Gawad *et al.* as well as the ones inferred by BnpC are displayed in Fig. S4. Genotypes and clones predicted by BnpC are largely in accordance with those previously determined. BnpC predicted some additional clones of small size. BnpC predictions were of partly higher resolution. Specifically for patient 4, BnpC was able to detect an additional clone (orange) differing from the closest clone by five mutations (Fig. 4 A). The identification of this particular clone results in a different and more accurate evolutionary pattern, as a common ancestor for the two tumor branches is obtained (Fig. 4 B). Gawad *et al.* confirmed the existence of this additional clone in their subsequent analysis by incorporating copy number data. These findings show that our approach is sensitive to small clones and able to recover biological meaningful results.

High-Grade Serous Ovarian Cancer The HGSOC data of patient 9 from the McPherson data set [29] was obtained by whole-genome sequencing of five samples taken from three tumor sites: left ovary, right ovary, and omentum. The data consists of 420 cells, 43 SNVs, and five breakpoints. We compared our predictions to the results obtained by Roth *et al.* using SCG. Their initial clustering analysis identified a normal population and eight tumor clones, of which they filtered out three clones due to a high fraction of missing values in the corresponding cells (mean $\geq$ 20% SNV events missing per cell).

BnpC was able to produce the same findings as SCG [19] without applying any additional filtering step (Fig. S7). By excluding the three clusters, 28 cells which represent 7% of the patient data were discarded.

The clonal prevalence shows differences between the two samples coming from the left ovary (LOv) (Fig. S7 B). Populations within one of the two samples (LOv2) contain the amplification in ERBB2, while the other (LOv1) does not. These populations harboring the amplification correspond to clones 0 (purple) and 1 (orange). Knowing that the primary site of the tumor was in the left ovary and that all other clones carry this amplification, our findings are in accordance with Roth et al.

Colorectal Cancer Patients CRC0827 and CRC0907 from Wu *et al.* [30] were collected by single-cell Whole Exome Sequencing on CRC tissue samples (C1 and C2) and matched normal tissue (N). Additional samples from normal polyp (NP, CRC0907) and adenomatous polyp tissue (AP, CRC0827) were sequenced for the analysis. While BnpC recapitulated the results for patient CRC0827, we identified an additional clone for patient CRC0907 (green clone in Fig. 5). This new clone suggests another step in the clonal evolution of the tumor. For patient CRC0907, Wu *et al.* identified two tumor clones harboring somatic mutations. They subsequently analyzed a subset of functionally related mutations to CRC development and separated them into unique clonal (detected by bulk sequencing) and unique subclonal (not detected). The results obtained from BnpC allows us to further classify the unique subclonal mutations into early subclonal (contained in the green clone) or late subclonal events (only present in the blue clone). Therefore, our method suggests an early mutation of LAMA4 compared to the other subclonal mutations annotated by Wu et al. (PDE3A, AB13BP, LHCGR, and CFHR5), which are only present in the later evolved population. Besides, we observed one of their annotated unique clonal mutations (STXBP1) to be present only in the blue clone, suggesting a later acquisition of the mutation. In summary, these results indicate that BnpC can give new insights into the evolution of the tumor and the order in which mutations are acquired by better identifying the clonal composition within tumor samples.

4 Discussion

The identification of the heterogeneous tumor composition and the clonal genotypes is potentially advantageous for cancer treatment. scDNA-seq provides the opportunity to resolve ITH in greater detail and to detect rare clones, despite experimental protocols still producing a high fraction of FN and missing events. We have introduced the novel non-parametric probabilistic method BnpC, specially designed for the clustering and genotyping of scDNA-seq data, accounting for uneven error rates and high fractions of missing values. The method implements a modification of a non-conjugate split-merge move and employs a novel estimator inferring from the posterior distribution for more accurate genotype predictions.

We compared our method with the state-of-the-art methods SCG and SiCloneFit on simulated and biological data. SCG achieved the highest clustering accuracy, but the least accurate genotype inference, and

it predicted unreliable genotypes at high FP rates (1%). While often small, depending on the sequencing platform and protocol, FP rates of several percent points can be observed in practice [8]. In the runtime analysis, SCG ran fastest and scaled well with data size thanks to its variational inference. On small data sets, SiCloneFit predicted clusters and genotypes with high accuracy. However, on larger data sets and within the limit of 10 hours, SiCloneFit failed to reach convergence and predicted worse clusters and genotypes than BnpC. This indicates that the underlying tree model is computationally expensive and not scalable to the larger data sets that will be produced in the near future. BnpC outperformed the other methods on small data sets in genotyping accuracy and performed similarly in clustering. Only for very high error or missing value rates (≥ 0.3) did SiCloneFit predict more accurate genotypes and SCG more accurate clusters. On the large data sets, BnpC achieved the most accurate genotype prediction and similar clustering accuracy than SCG. Its runtime scaled nearly linearly with data size (Fig. 3) and optimal results were already reached after one hour on the largest data sets. We argue that clustering with inaccurate genotypes misrepresents the data and can mislead downstream analyzes and biological interpretation. BnpC is so far the only method providing accurate clonal genotypes on large data sets within reasonable time.

On real data, our method did not only reproduce previous findings for three different data sets but identified additional clones not detected in the original analysis but confirmed by additional data in Patient 4 from Gawad *et al.* and Patient CRC0907 from Wu *et al.* For patient 9 from McPherson *et al.*, we were able to recapitulate previous results without the manual pre-processing step conducted in the original analysis.

A limitation of the BnpC model is the absence of a phylogenetic structure on cells. The information given by the mutation order could be used to correct errors in the data or to infer missing values. However, tree structures are computationally expensive to infer and the trade-off between accuracy and runtime needs to be investigated further. A possible extension of BnpC could be the incorporation of doublets, two single cells pooled and measured together during sequencing. Currently, doublets would be reported as separate clones. Identifying and handling them explicitly as doublets could improve the clustering and genotyping, especially of the clones corresponding with the two doublet cells.

In summary, the non-parametric nature and sampling scheme of our model results in robust clonal composition and genotype predictions in reasonable computational time. Besides their relevance for personalized treatment, the inferred clusters and genotypes can be used to reduce data size significantly, thereby facilitating downstream analyses. As scDNA-seq data size grows, this problem will become increasingly relevant. Additionally, the generic implementation not assuming a tree-structure makes our method applicable to other fields. For example, our method could be used for the analysis of methylation profiles or the analysis of microbiome data, where the input matrix indicates the presence or absence of species in samples.

5 Software availability

BnpC was implemented in Python 3.6 and is freely available under MIT license at: https://github.com/cbg-ethz/BnpC.

6 Supplementary Material

Additional figures and a description of the simulation scheme.

7 Author Contribution

N.Bo., J.B., and F.M. designed the study. N.Bo., J.B., and F.M. developed the methodology. N.Bo. and J.B. implemented the methodology. N.Bo. and J.B. performed analyses. N.Be., A.GP., and N.LB. supervised the study. All authors drafted the manuscript and approved the final version.

8 Acknowledgement

This work was funded by ITN-CONTRA EU grant H2020 MSCA-ITN-2017-766030. IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO; Government of Spain) and is supported by CERCA (Generalitat de Catalunya). Part of this work was supported by the European Research Council, ERC Synergy Grant 609883 (to N.B.)

9 Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Weinberg, Robert Allan. *The biology of cancer*. Garland Science, 2014.

[2] Davis, Alexander, Gao, Ruli, and Navin, Nicholas. "Tumor evolution: Linear, branching, neutral or punctuated?" In: *Biochim Biophys Acta Rev Cancer* 1867.2 (2017), pp. 151–161.

[3] Turajlic, Samra et al. "Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal". In: *Cell* 173.3 (2018), 595–610.e11. ISSN: 1097-4172.

[4] Gillies, Robert J., Verduzco, Daniel, and Gatenby, Robert A. "Evolutionary dynamics of carcinogenesis and why targeted therapy does not work". In: *Nature Reviews Cancer* 12.7 (2012), pp. 487–493.

[5] Wang, Yong Xin et al. "Clonal Evolution in Breast Cancer Revealed by Single Nucleus Genome Sequencing". In: *Nature* 512 (2014), pp. 155–160.

[6] Deczkowska, Aleksandra et al. "Disease-Associated Microglia: A Universal Immune Sensor of Neurodegeneration". In: *Cell* 173.5 (May 2018), pp. 1073–1081. ISSN: 0092-8674.

[7] Schwartz, Russell and Schäffer, Alejandro A. "The evolution of tumour phylogenetics: principles and practice". In: *Nature Reviews Genetics* 18 (Feb. 2017). Review Article, pp. 213–219.

[8] Estévez-Gómez, Nuria et al. "Comparison of single-cell whole-genome amplification strategies". In: (2018).

[9] Yuan, Ke et al. "BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies". In: *Genome Biology* 16.1 (2015), p. 36.

[10] Ross, Edith M. and Markowetz, Florian. "OncoNEM: inferring tumor evolution from single-cell sequencing data". In: *Genome Biology* 17.1 (2016).

[11] Jahn, Katharina, Kuipers, Jack, and Beerenwinkel, Niko. "Tree inference for single-cell data". In: *Genome Biology* 17.1 (May 2016).

[12] Kuipers, Jack et al. "Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors". In: *Genome research* 27.11 (Nov. 2017), pp. 1885–1894.

[13] Zafar, Hamim et al. "SiFit: Inferring tumor trees from single-cell sequencing data under finite-sites models". In: *Genome Biology* 18 (Dec. 2017).

[14] Zafar, Hamim et al. "SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data". In: *Genome Research* (2019).

[15] Malikic, Salem et al. "PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data". In: *Genome Research* (2019).

[16] El-Kebir, Mohammed. "SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error". In: *Bioinformatics* 34.17 (2018), pp. i671–i679.

[17] Ciccolella, Simone et al. "Inferring Cancer Progression from Single Cell Sequencing while allowing loss of mutations". In: *bioRxiv* (2018).

[18] DOLLO, L. "Lesloisdel'évolution". In: *Bul. Soc. Belge Géol. Pal. Hydr* 7 (1893), pp. 164–166.

[19] Roth, Andrew et al. "Clonal genotype and population structure inference from single-cell tumor sequencing". In: *Nature Methods* 13.573 (2016), pp. 573–576.

[20] Pitman, Jim. "Exchangeable and partially exchangeable random partitions". In: *Probability Theory and Related Fields* 102.2 (June 1995).

[21] Escobar, Michael D. and West, Mike. "Bayesian Density Estimation and Inference Using Mixtures". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 577–588.

[22] Neal, Radford M. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models". In: *Journal of Computational and Graphical Statistics* 9.2 (2000), pp. 249–265.

[23] Jain, Sonia and Neal, Radford M. "Splitting and merging components of a nonconjugate Dirichlet process mixture model". In: *Bayesian Anal.* 2.3 (2007), pp. 445–472.

[24] Jain, Sonia and Neal, Radford M. "A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model". In: *Journal of Computational and Graphical Statistics* 13.1 (Mar. 2004), pp. 158–182.

[25] Fritsch, Arno and Ickstadt, Katja. "Improved criteria for clustering based on the posterior similarity matrix". In: *Bayesian Anal.* 4.2 (June 2009), pp. 367–391.

[26] Zafar, Hamim et al. "SiCloneFit: Bayesian inference of population structure, genotype,and phylogeny of tumor clones from single-cell genome sequencing data". In: *bioRxiv* (2018).

[27] Rosenberg, Andrew and Hirschberg, Julia. "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure". In: *Proc. 2007 Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007, pp. 410–420.

[28] Gawad, Charles, Koh, Winston, and Quake, Stephen R. "Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics". In: *PNAS* 111.50 (2014), pp. 17947–17952.

[29] McPherson, Andrew et al. "Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer". In: *Nature Genetics* 48 (May 2016), pp. 758–767.

[30] Wu, H. et al. "Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing". In: *Oncogene* 36.20 (2017), pp. 2857–2867. ISSN: 1476-5594.

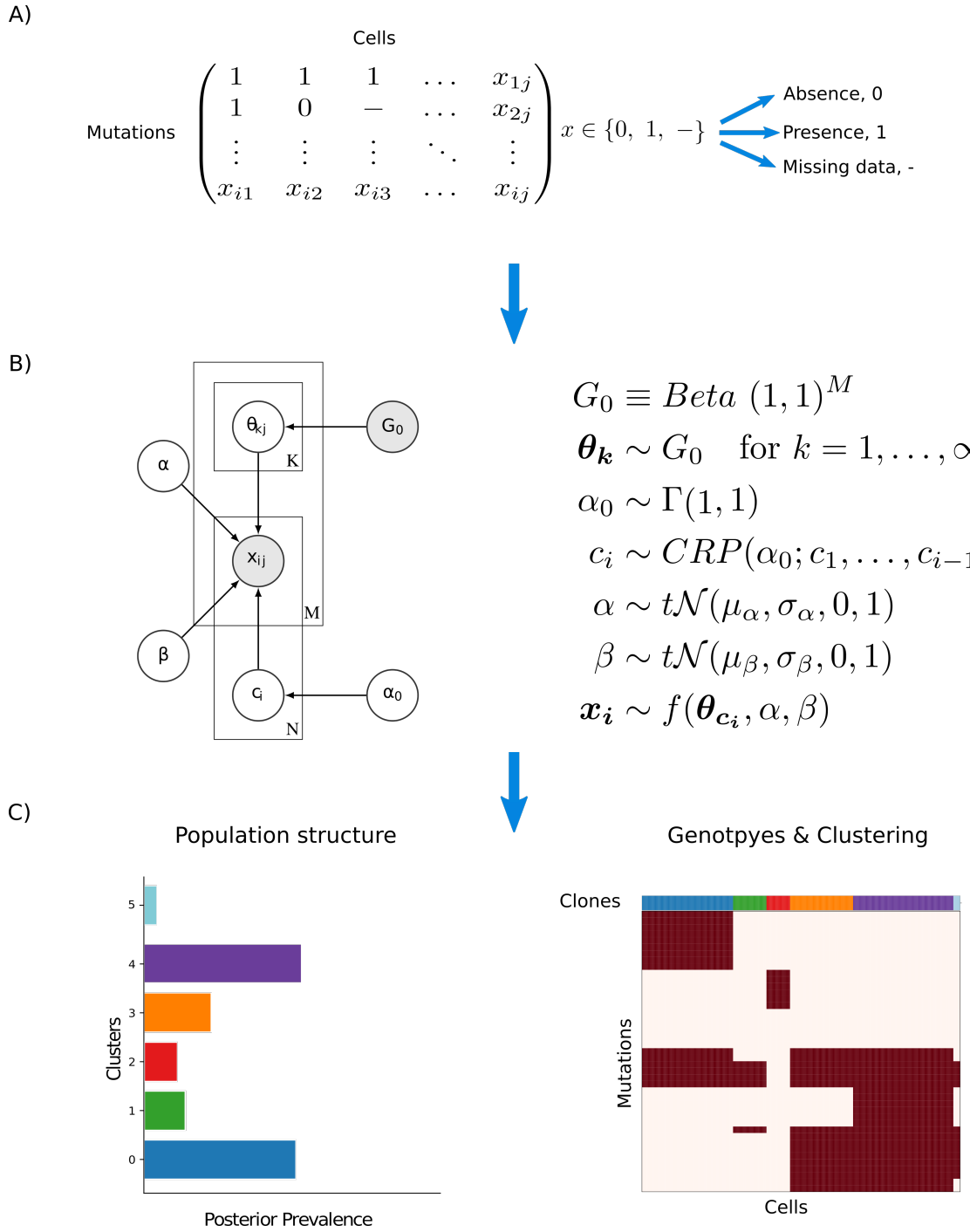

Fig. 1: BnpC model overview. **A)** The model's input is a binary mutation matrix, where each row represents a mutation and each column represents a single cell. Possible values are 0, indicating the absence of a mutation, 1, indicating the presence of a mutation, and missing values. **B)** BnpC's probabilistic graphical model. The binary input data X, consisting of N cells and M clones, contains a fraction of FP and FN entries, indicated by α and β, respectively. G_0 is a base distribution over the genotypes θ of an infinite number of clones. c is the assignment of cells to the clones, sampled from a CRP with concentration parameter α_0, and $f(\cdot)$ is the model's likelihood. Shaded nodes represent observed or fixed values, while the values of unshaded nodes are learned using MCMC. **C)** BnpC predicts clonal composition, corresponding genotypes, and the population structure.

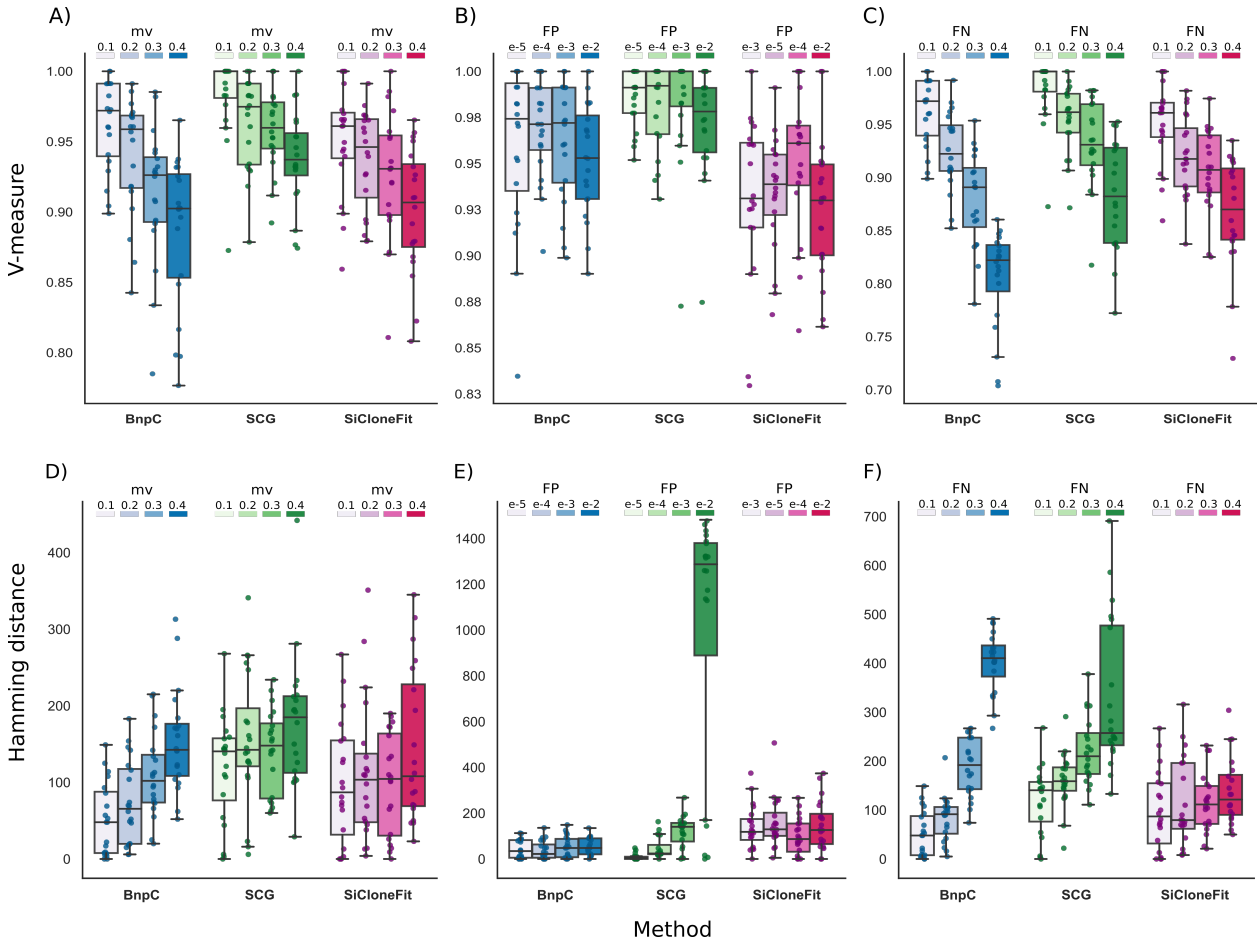

Fig. 2: (**A, B, C**) Clustering accuracy and (**D, E, F**) genotyping error of BnpC, SCG, and SiCloneFit on a synthetic data set of 50 mutations, 200 cells, and 10 clusters. **A)** 10% FN rate, 0.01% FP rate, and a variable rate of missing values (mv). **B)** 10% missing values, 10% FN rate, and a variable FP rate. **C)** 10% missing values, 0.01% FP rate, and a variable FN rate. **D)** 10% FN rate, 0.1% FP rate, and a variable rate of missing values (mv). **E)** 10% FN rate, 10% missing values, and a variable FP rate. **F)** 10% missing values, 0.1% FP rate and a variable FN rate. Each combination of error rates and missing values was simulated 20 times.

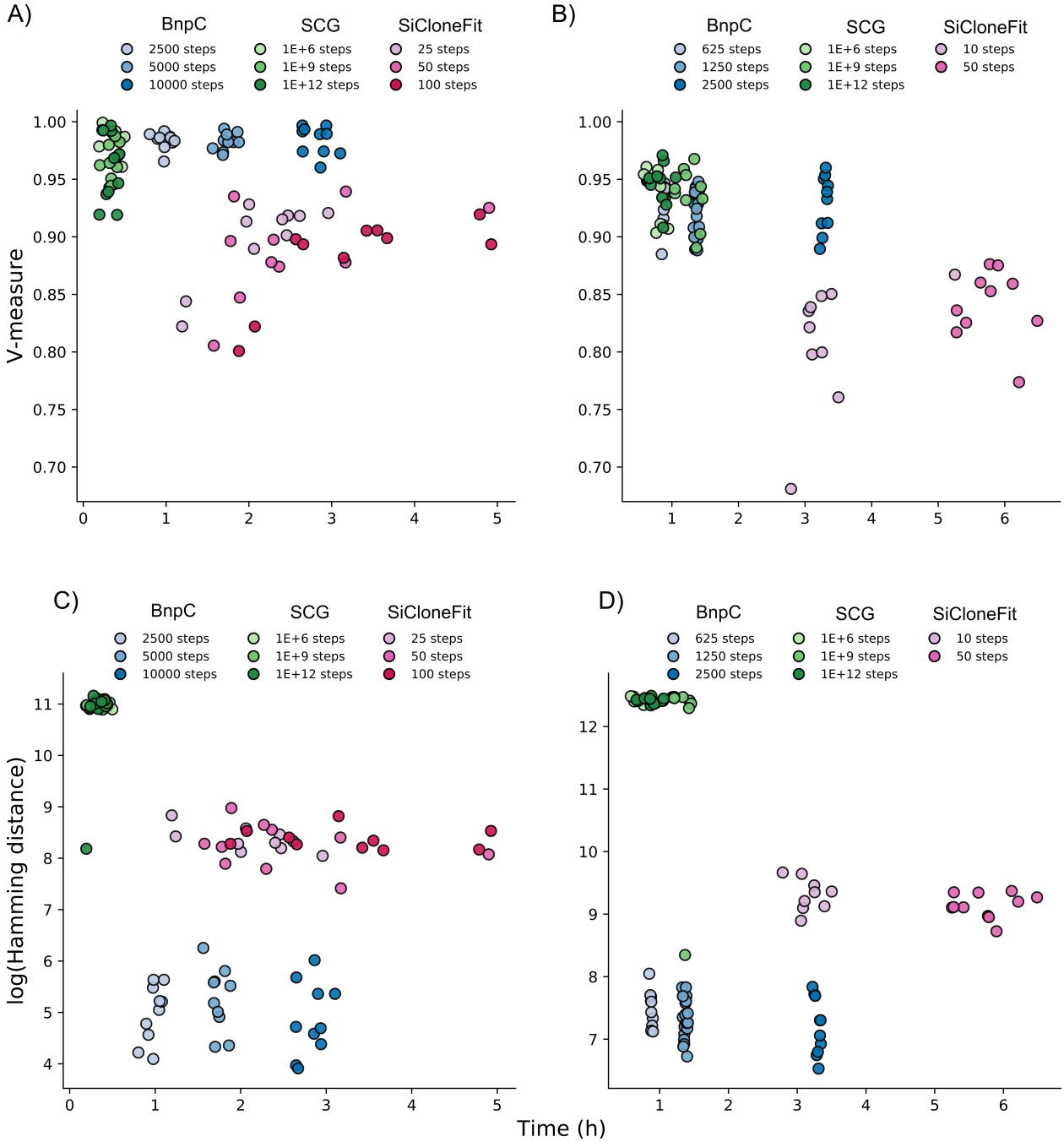

Fig. 3: Clustering accuracy measured by the V-Measure and genotyping error measured by the Hamming distance of BnpC, SCG, SiCloneFit. **A, B)** Clustering accuracy. **C, D)** Genotyping error. **A, C)** Data set of 1000 cells, 200 mutations, and 20 clusters. **B, D)** Data set of 3000 cells, 200 mutations, and 30 clusters. Both data sets had a FN rate of 30%, FP rate of 1%, and a missing value fraction of 20%. Algorithms were run 10 times per number of steps each.

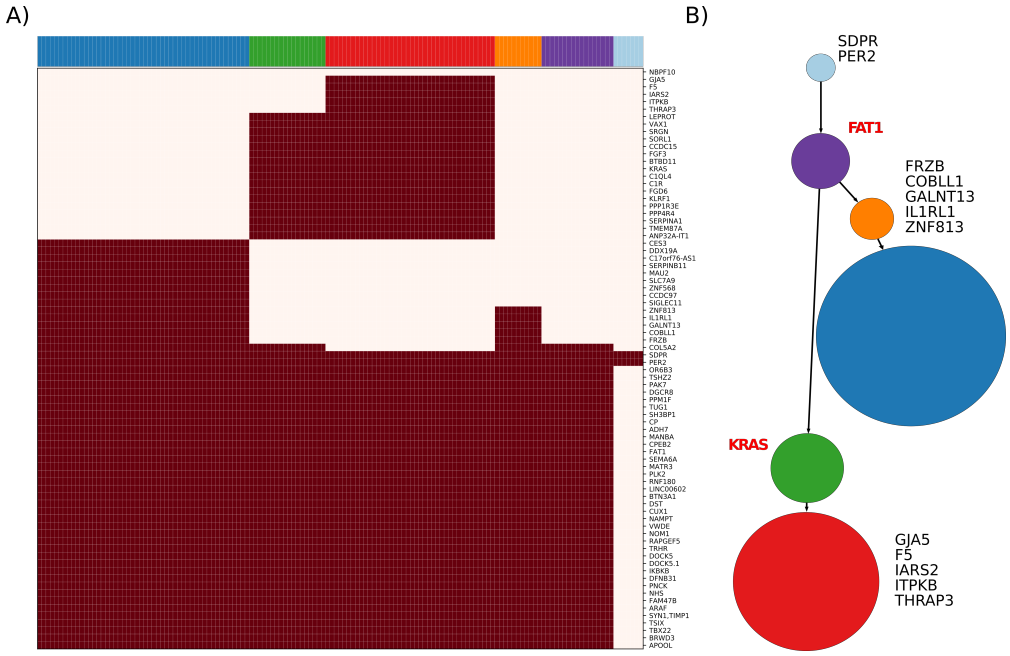

Fig. 4: **A)** Clones and genotypes inferred by BnpC for patient 4 of the Gawad data set. Heatmap depicts absence (white) or presence (red) of mutations for every mutation (row) in every cell (column). **B)** Resulting minimum spanning tree from the clonal genotypes as obtained in Gawad et al. Gene labels in the tree determine either mutations leading to a new clone (black) or known ALL driver genes (red). Node size corresponds with the clonal size.

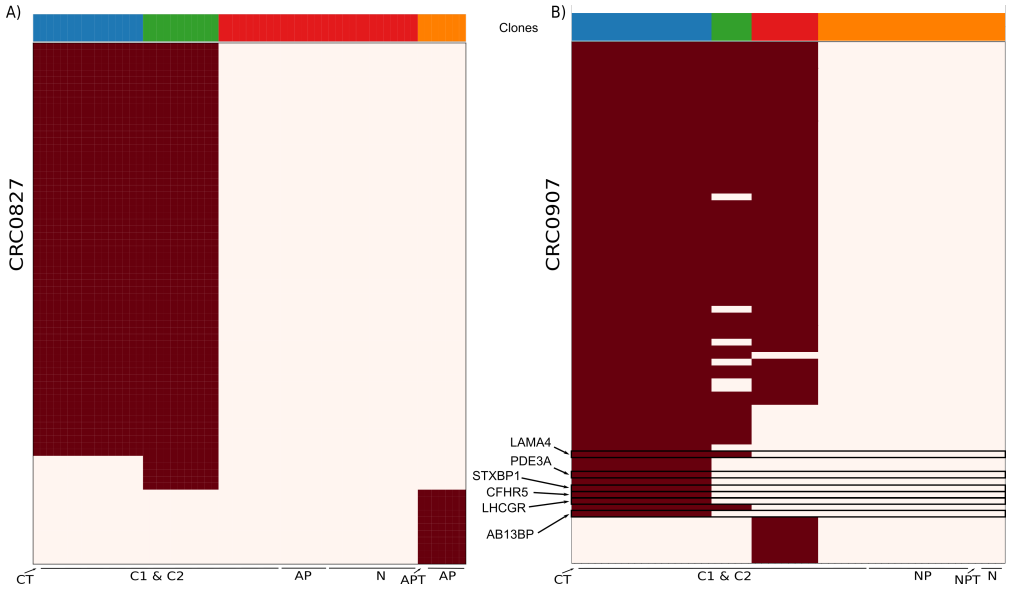

Fig. 5: Inferred clones and genotypes by BnpC for patients CRC0827 **A)** and CRC0907 **B)** of the Wu data set. Heatmap depicts absence (white) or presence (red) of mutations for every mutation (row) in every cell (column).