Submitted to the Annals of Applied Statistics

SUPPLEMENTARY MATERIALS FOR: PHYLOGENY-BASED TUMOR SUBCLONE IDENTIFICATION USING A BAYESIAN FEATURE ALLOCATION MODEL

By Li Zeng, Joshua L. Warren and Hongyu Zhao

Yale University

Figure 1-4: Z matrix estimation from the three methods
Subclones in the figures are sorted so that they resemble true **Z** matrices as much as possible.

Figure 5-8: Point estimation of parameters in simulations

Figure 9-12: Point estimations of parameters from the breast cancer dataset

Figure 13-14: CNVs detected by SIFA from the breast cancer dataset (PD9771, PD9849)

Figure 15-18: VAF for each SNV cluster identified by SIFA from the breast cancer dataset

Figure 19-20: Estimated evolution tree structure, subclone fractions and model selection criterion values for patients PD9771 and PD9849.
The left panels present the inferred phylogenetic tree, with breast cancer related genes listed on its right. Red gene names indicate loci with copy gain, green names indicate loci with copy loss, and the others are copy neutral loci. The middle panels show fractions of each subclone in all WGS samples. Subclones are represented by colors, and the lengths of each colored segment are proportional to their fractions. The right panels present the Bayes free energy values calculated in the model selection step.

Figure 21-24: VAF for each SNV cluster identified by Cloe from the breast cancer dataset

Figure 25-26: VAF for each SNV cluster identified by Pyclone from

the breast cancer dataset
Due to the missing copy number information for PD9771 and PD9777, Pyclone was only applied to patients PD9694 and PD9849. For PD9694, Pyclone reported 17 clusters. We only present the clusters with at least 5 mutations in the figure.

Figure 27: VAF fitting error from `Cloe`, `SIFA` and `Pyclone` in the analysis of the four breast cancer patients.

Figure 28: Computation speed of SIFA. We calculate computation speed in terms of seconds/100 sampling iterations, under different loci numbers (100, 200 and 300) and number of subclones (K = $3, 4, \ldots, 8$**). The figure shows that both data size and the number of subclone** K **are relevant to computation efficiency.**

Figure 29: Geweke's Statistics for testing convergence of Bayesian samples. Top, middle,and bottom panels present the statistics for SIFA, Cloe, and Pyclone, respectively.
For each method, we calculated the Geweke diagnosis statistics for each subclone fraction parameter (θ_{kt}). They are presented using boxplots in Figure 29. In case of multiple trees in posterior samples, it was not feasible to calculate the statistics using all samples, because θ_{kt} for different trees hold different meanings. Therefore, in such cases, the Geweke statistics were calculated only using the most frequent tree, to ensure tree-wise convergence. In all simulation settings the statistics generally have absolute values below two, indicating the good quality of convergence.

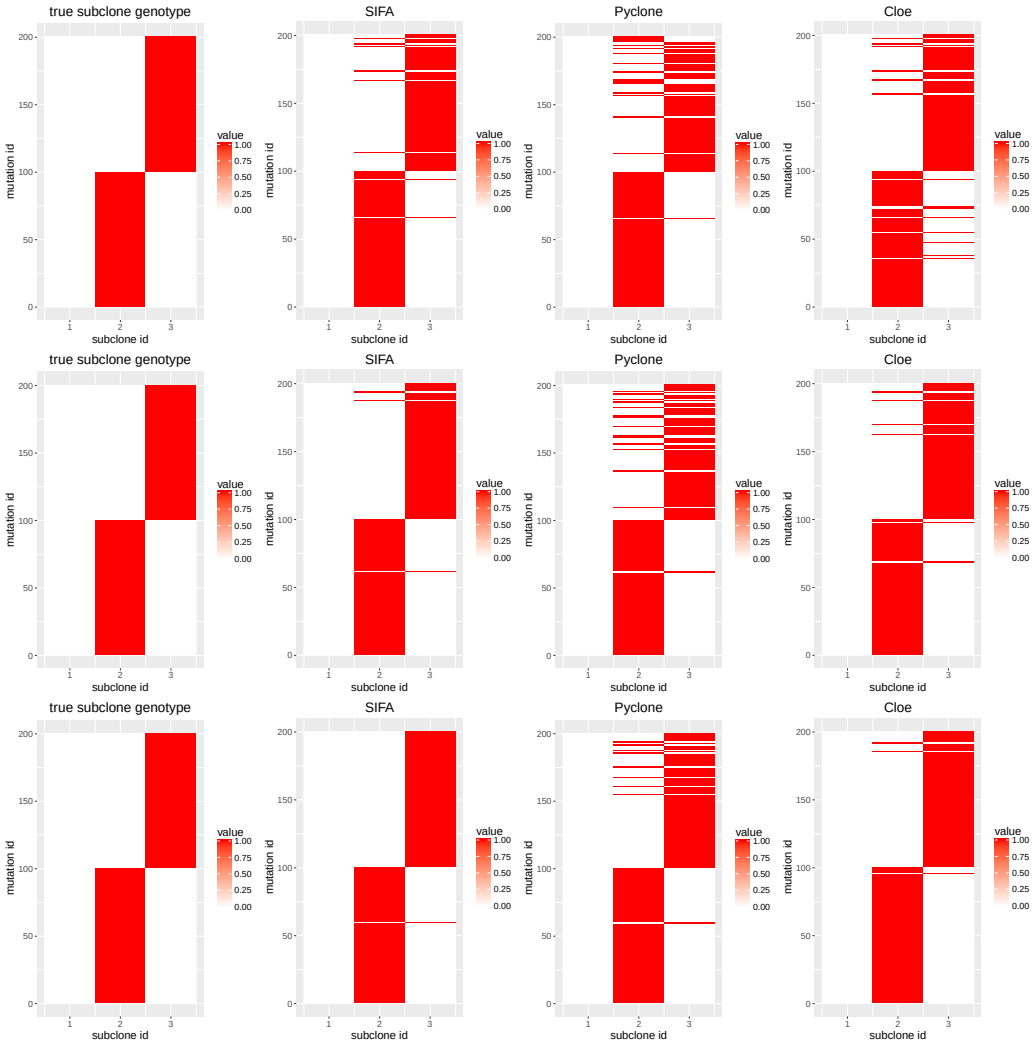

Fig 1: Results for scenarios where $K = 3$. The three rows have simulated sequencing depth $40, 60$, and 80, respectively.

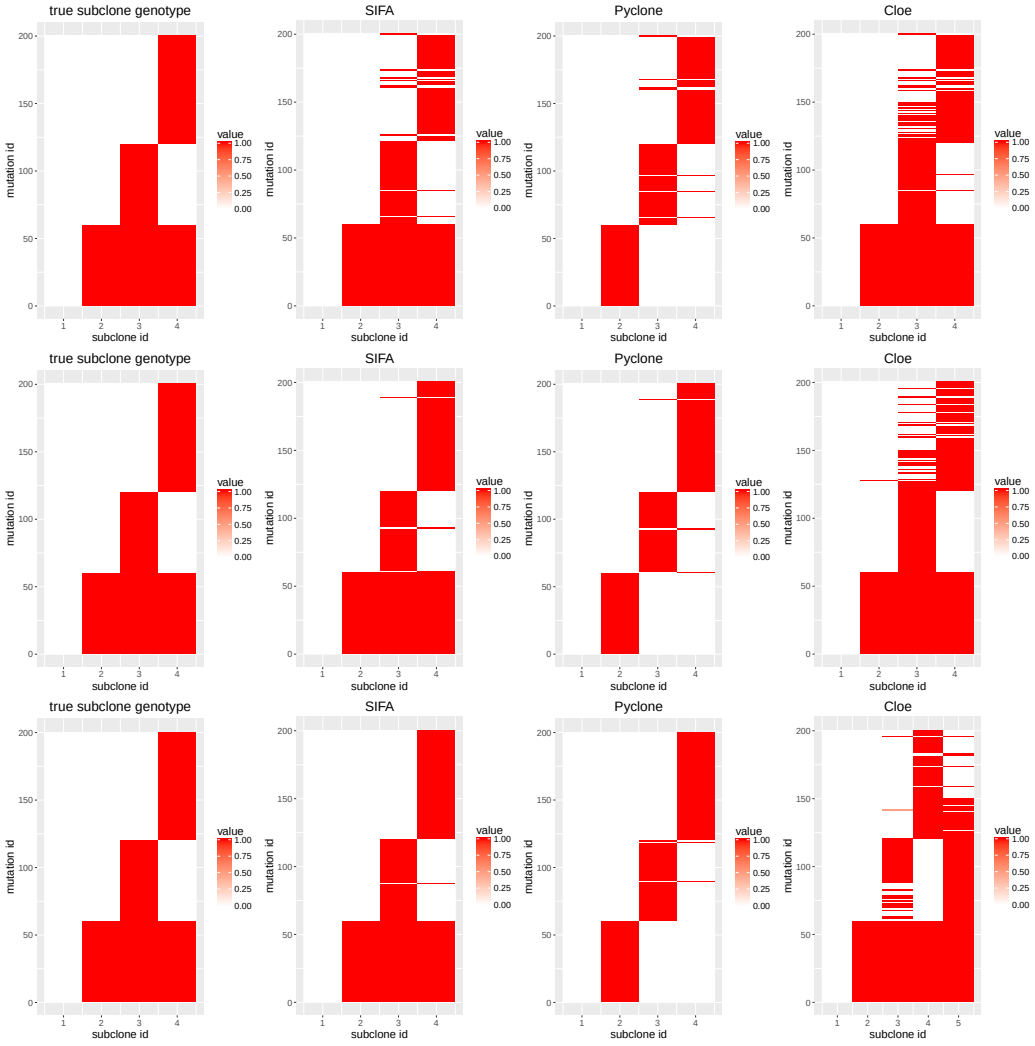

Fig 2: Results for scenarios where $K = 4$. The three rows have simulated sequencing depth $40, 60$, and 80, respectively.

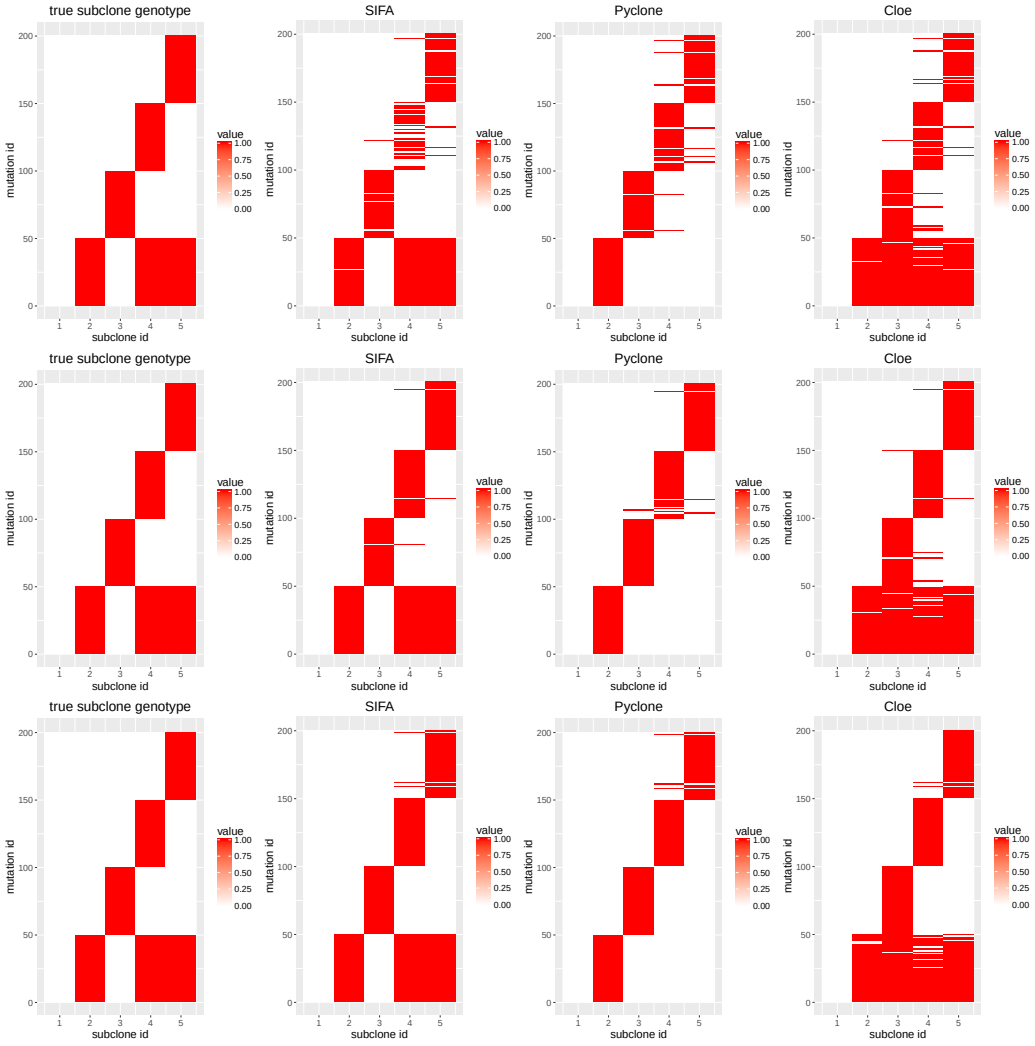

Fig 3: Results for scenarios where $K = 5$. The three rows have simulated sequencing depth $40, 60$, and 80, respectively.

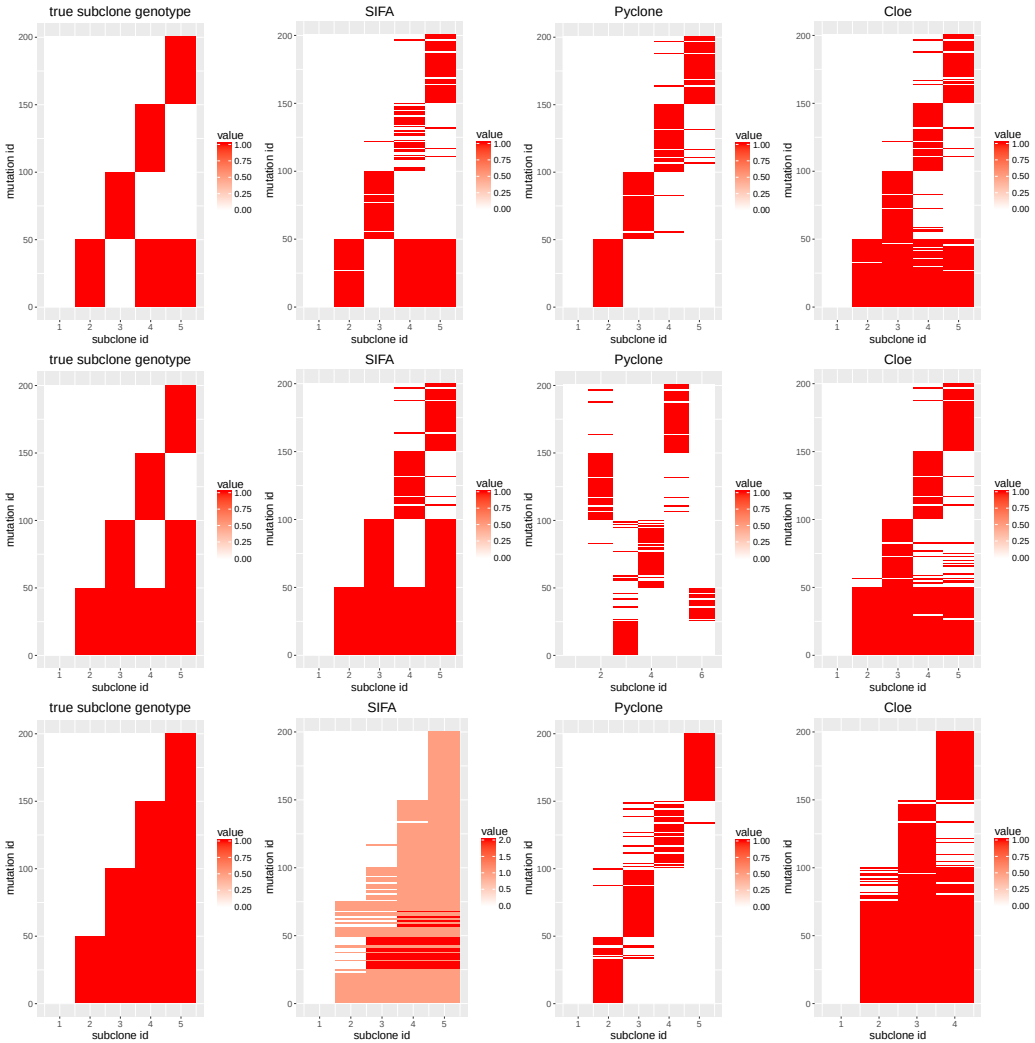

Fig 4: Results for scenarios $K = 5$ with different tree structures.

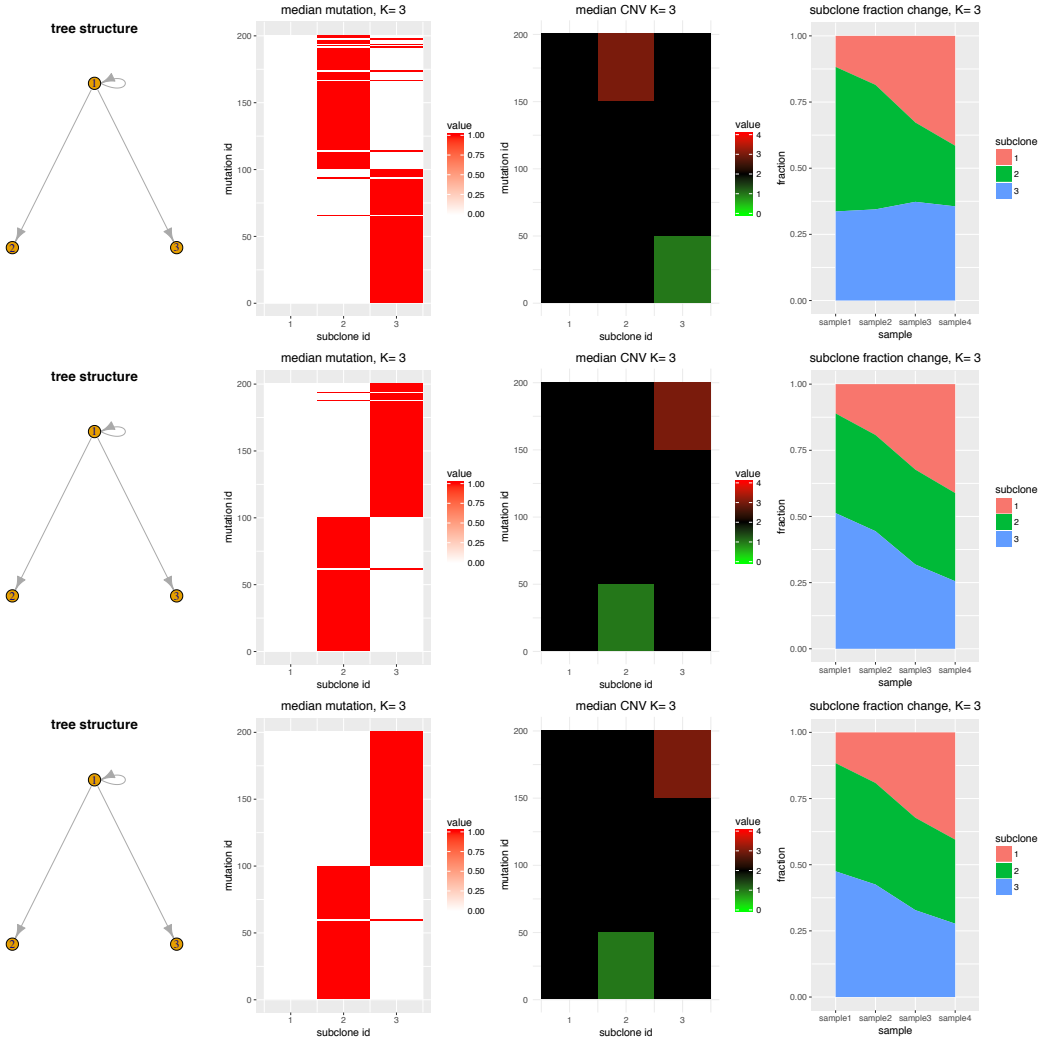

Fig 5: Results for scenarios where $K = 3$. The three rows have simulated sequencing depth $40, 60$, and 80, respectively.

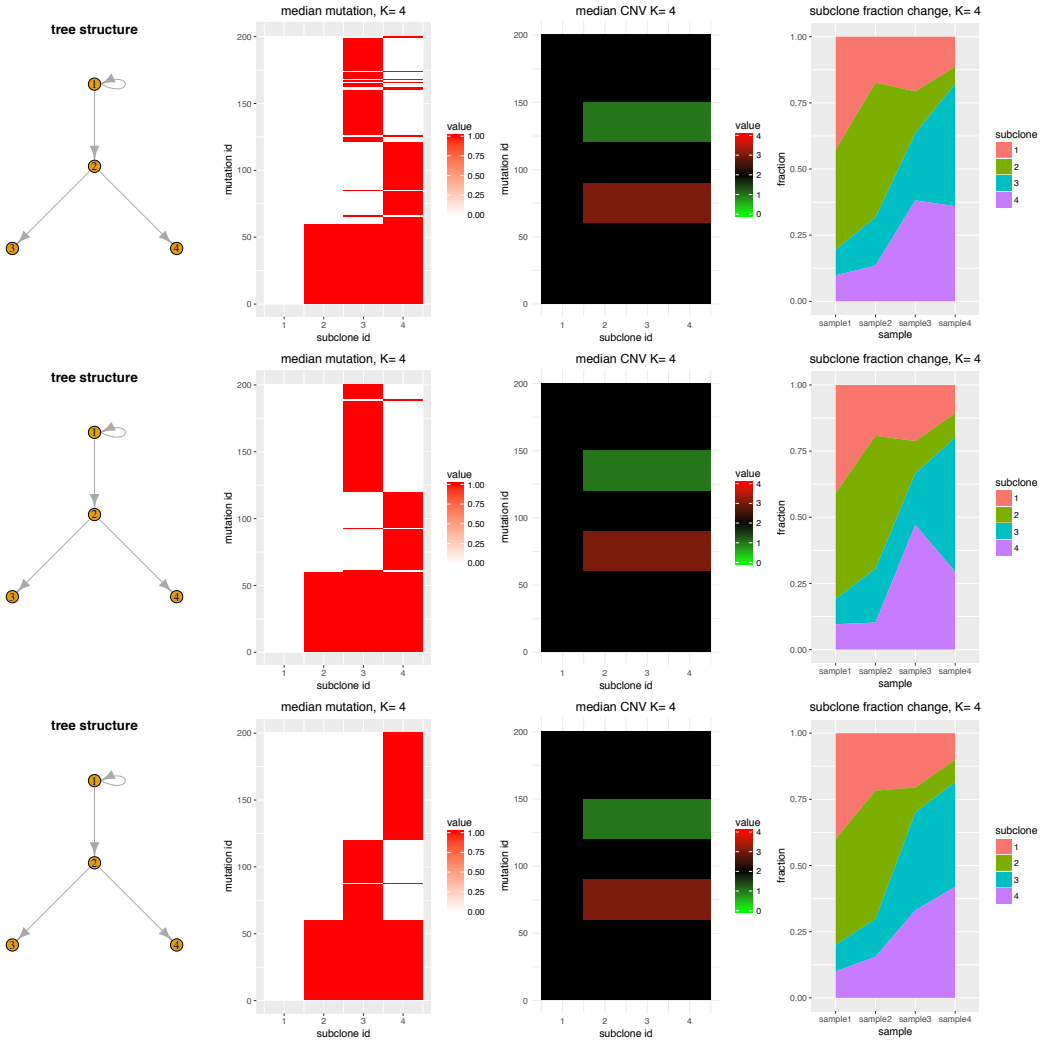

Fig 6: Results for scenarios where $K = 4$. The three rows have simulated sequencing depth $40, 60$, and 80, respectively.

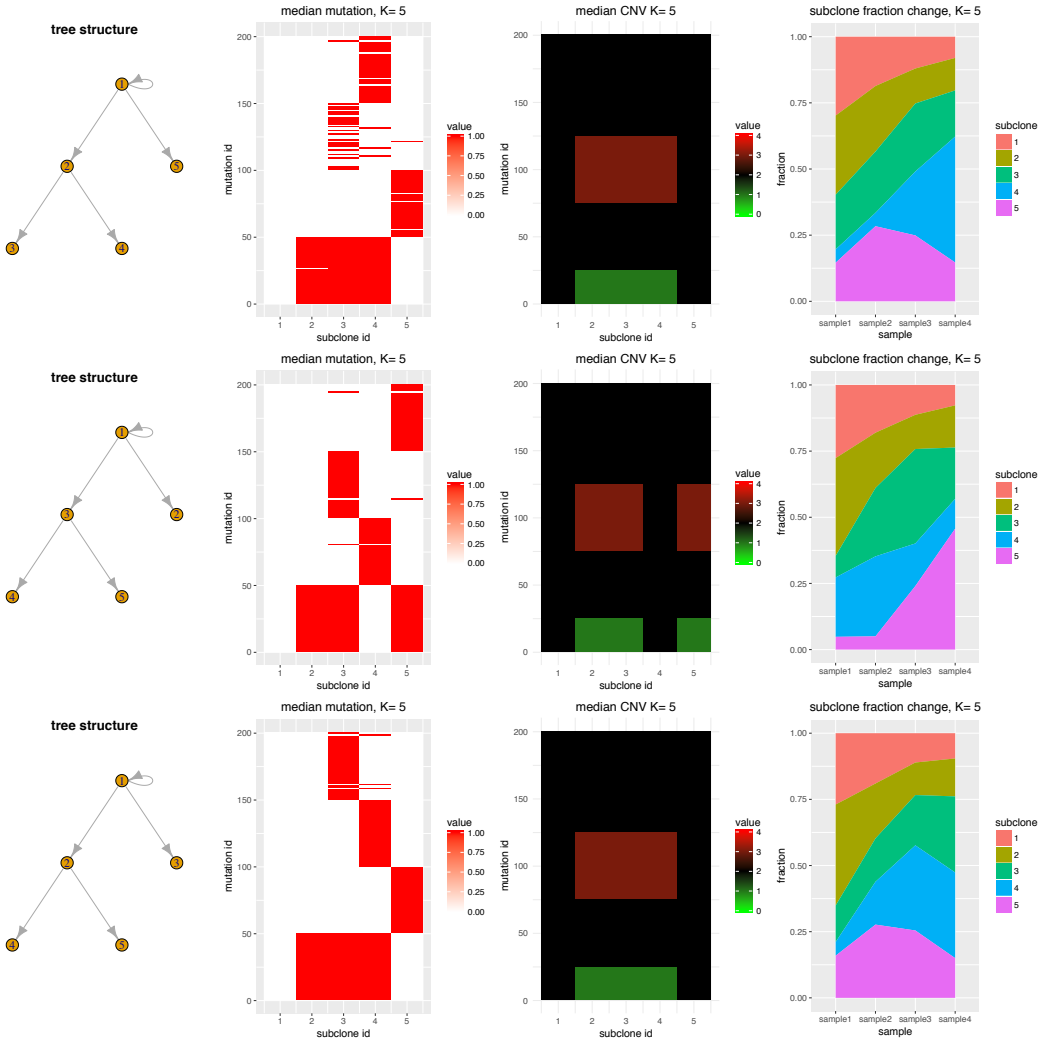

Fig 7: Results for scenarios where $K = 5$. The three rows have simulated sequencing depth $40, 60$, and 80, respectively.

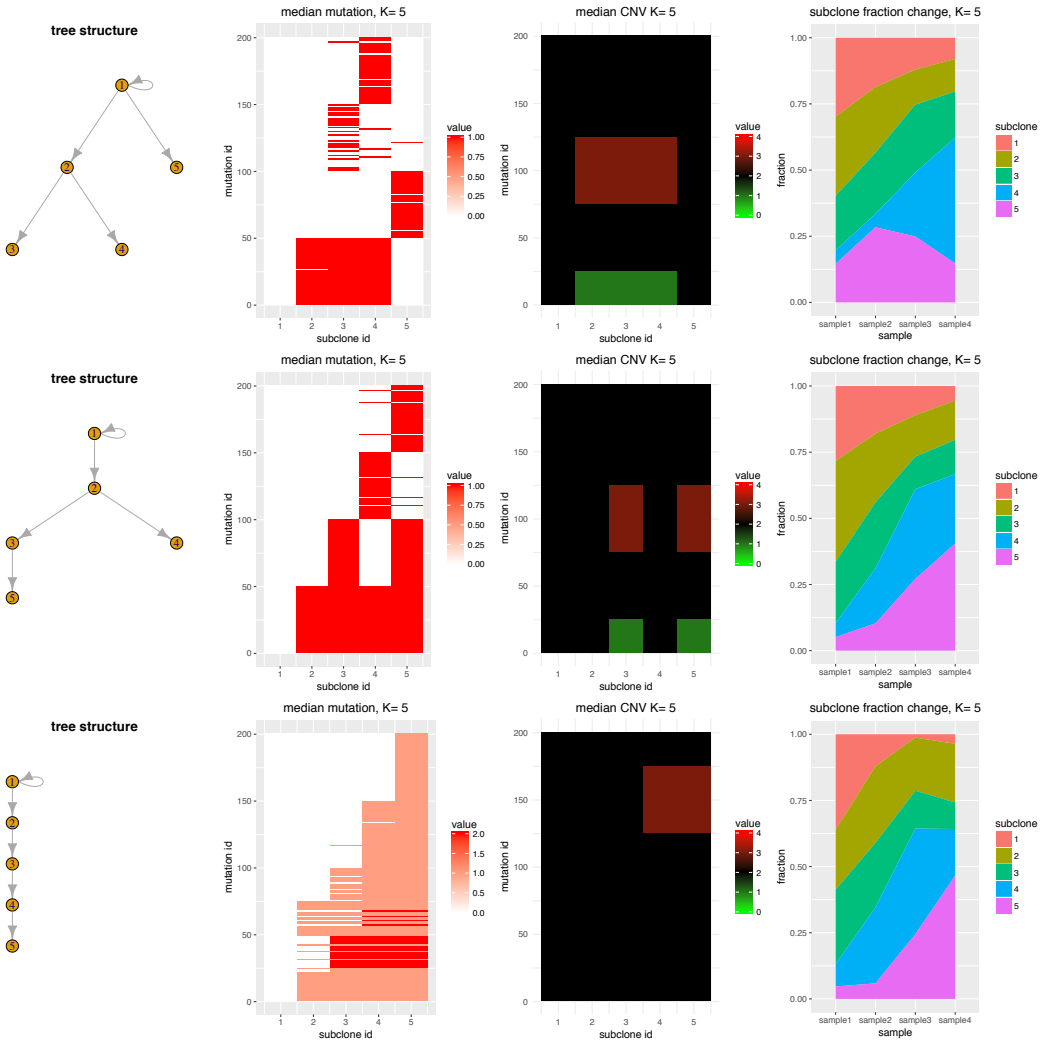

Fig 8: Results for scenarios $K = 5$ with different tree structures.

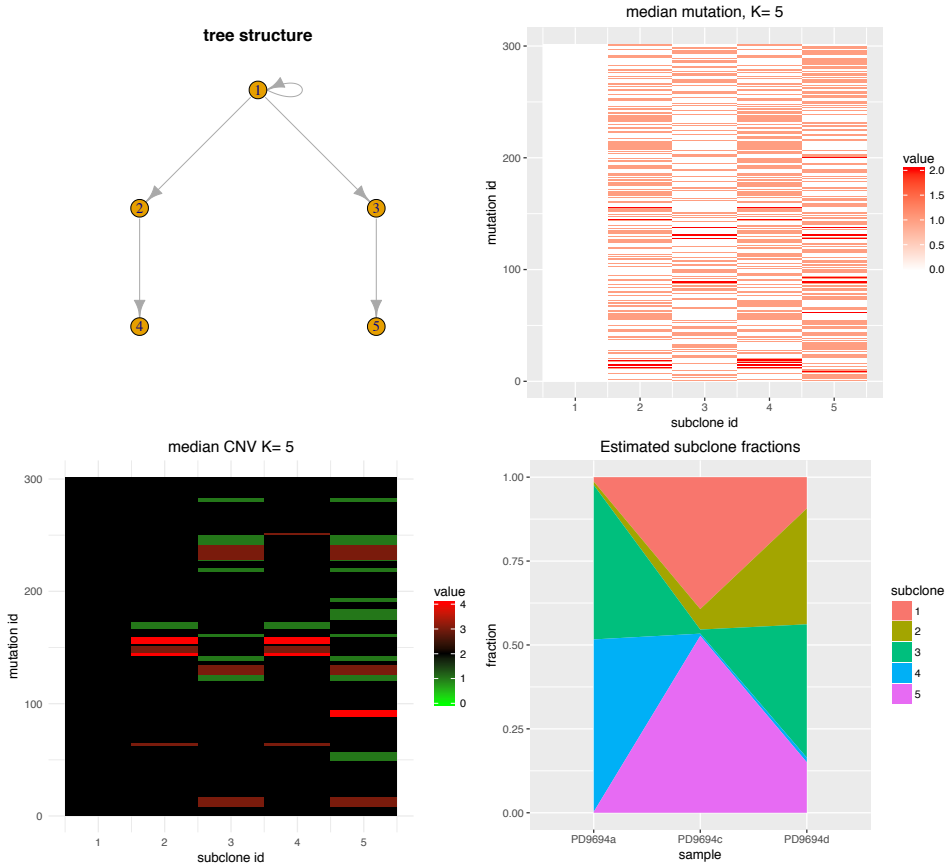

Fig 9: SIFA parameter estimations for patient PD9694.

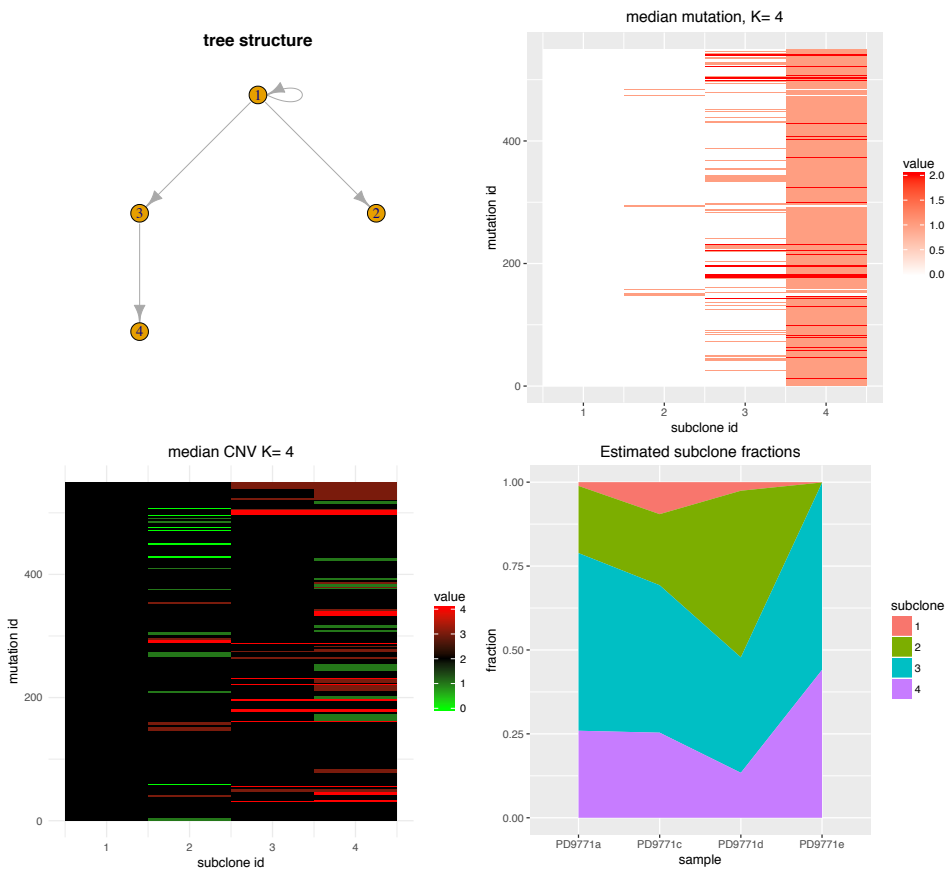

Fig 10: SIFA parameter estimations for patient PD9771.

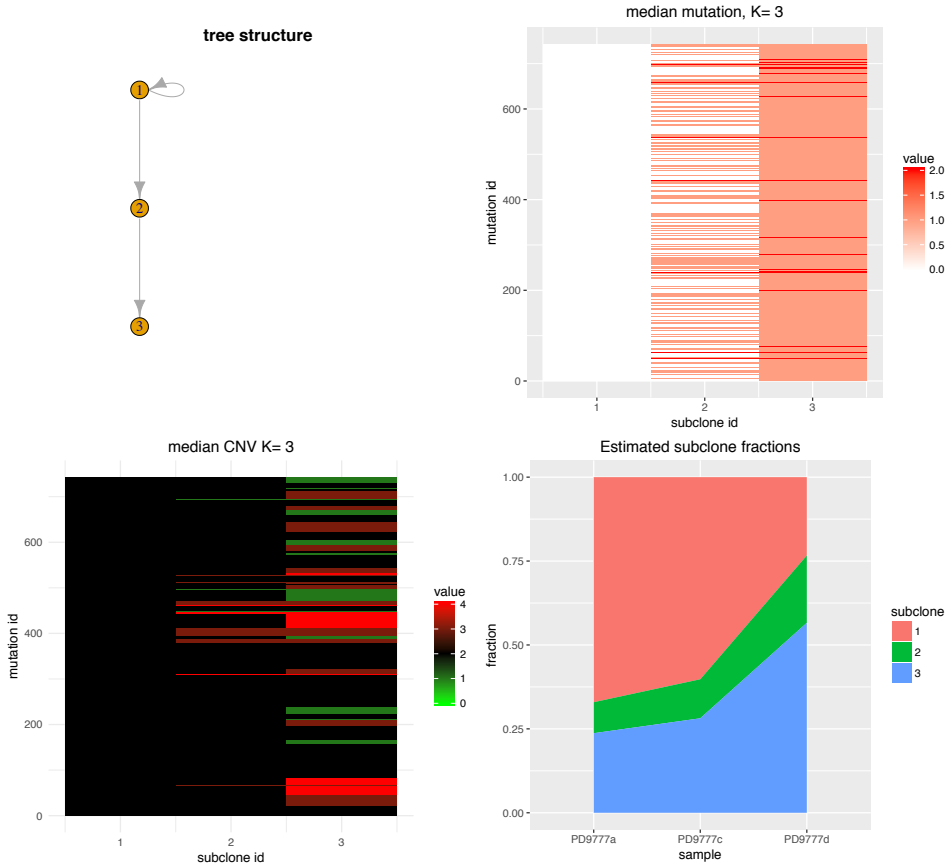

Fig 11: SIFA parameter estimations for patient PD9777.

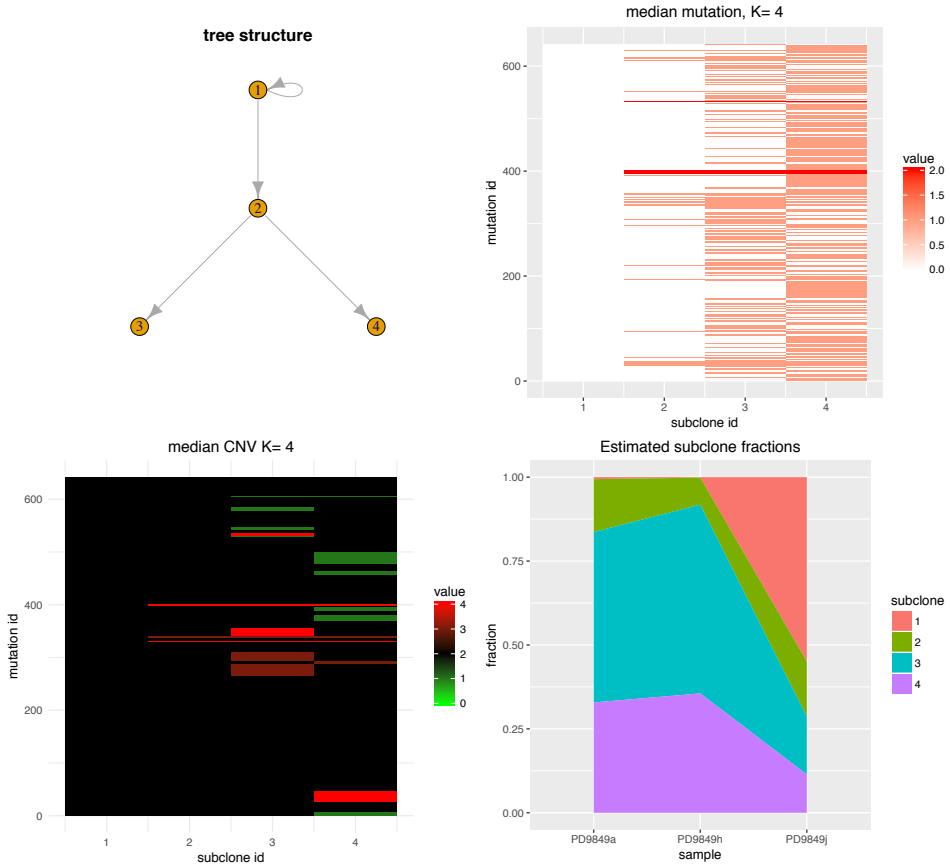

Fig 12: SIFA parameter estimations for patient PD9849.

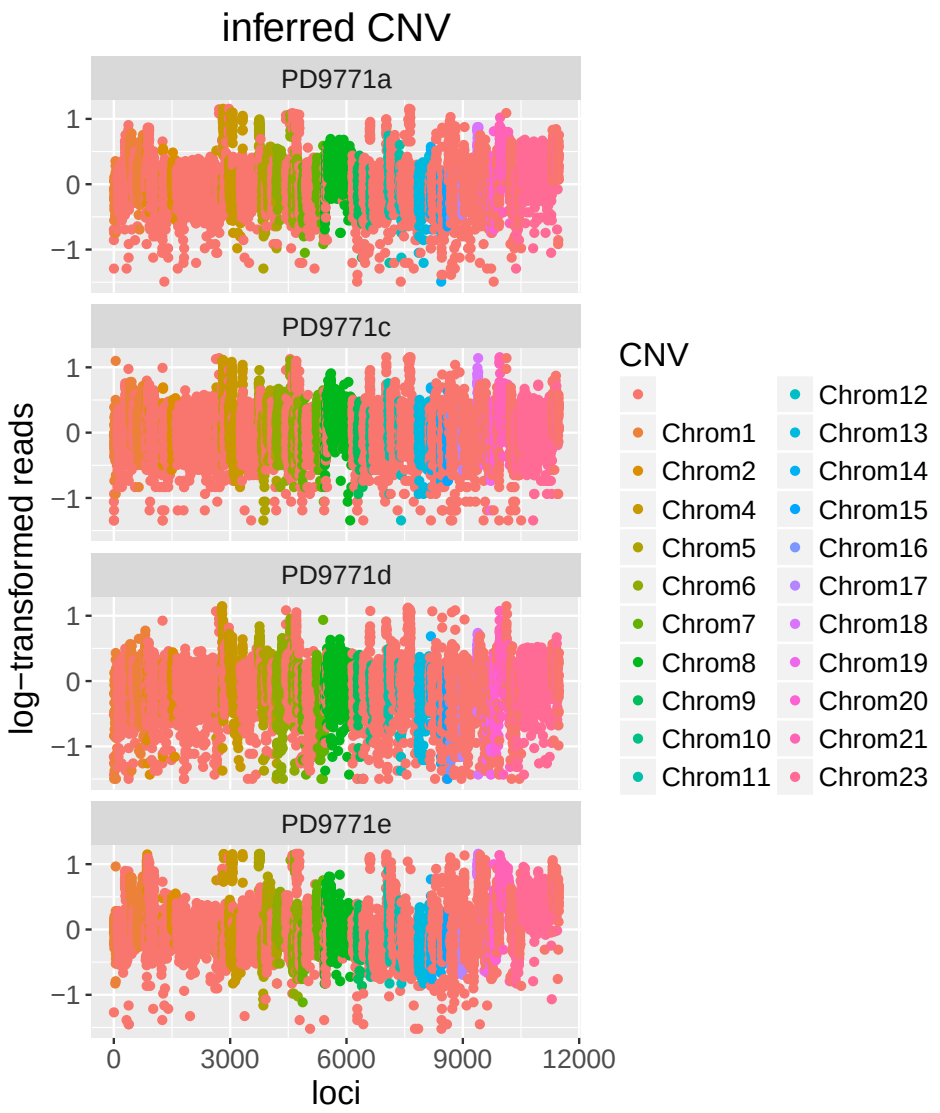

Fig 13: CNVs detected by SIFA for PD9771.

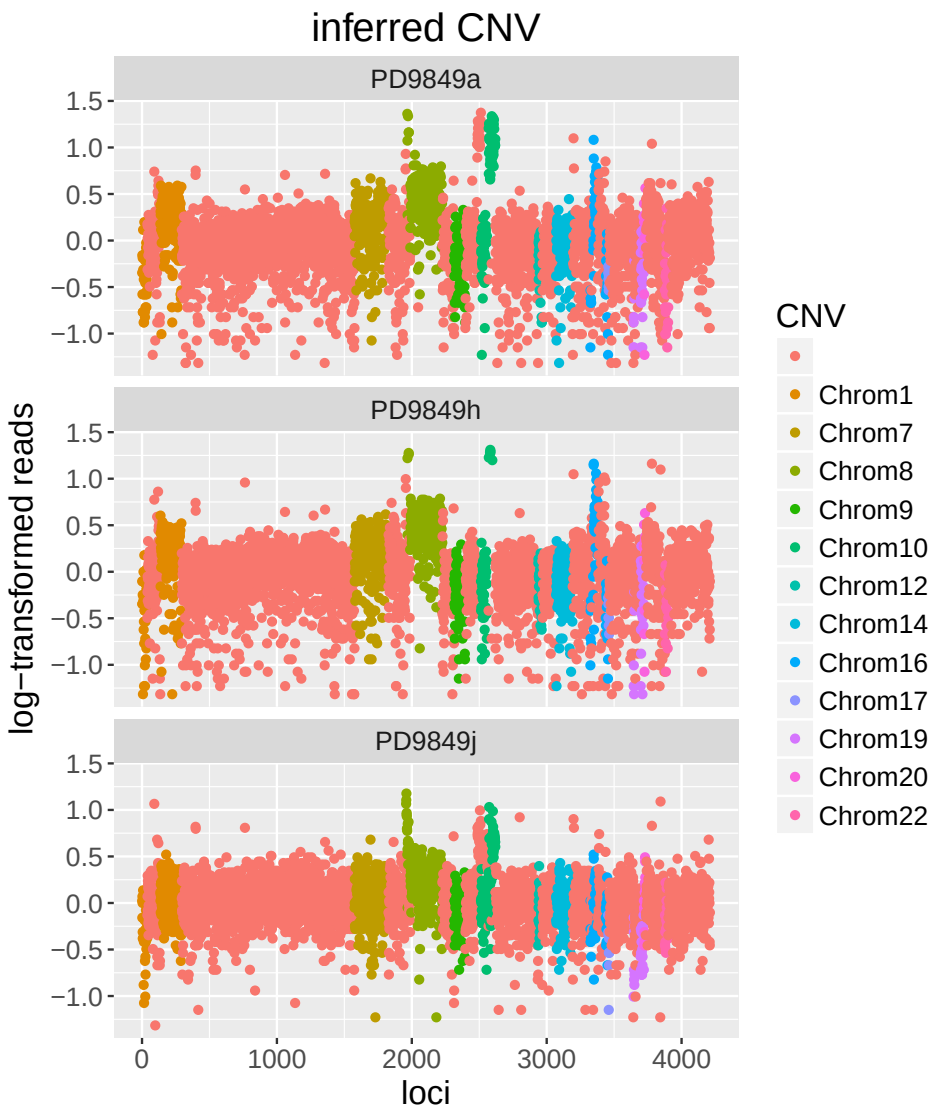

Fig 14: CNVs detected by SIFA for PD9849.

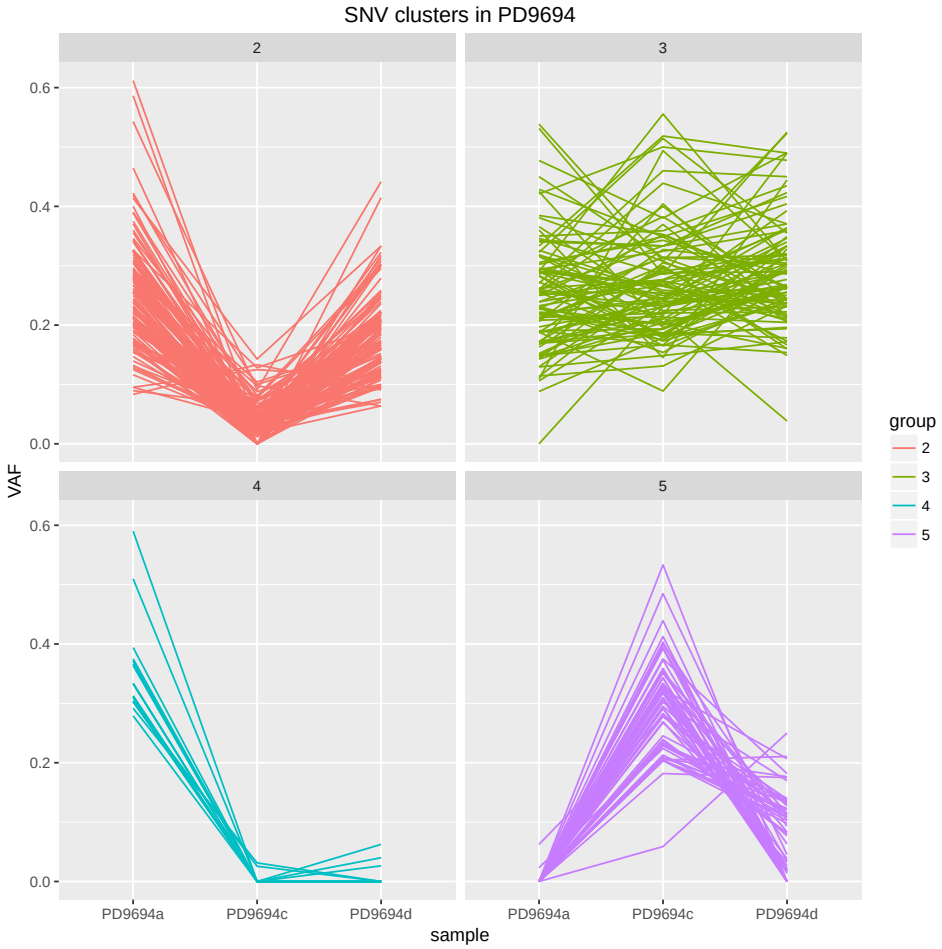

Fig 15: VAF for different SNV clusters of patient PD9694.

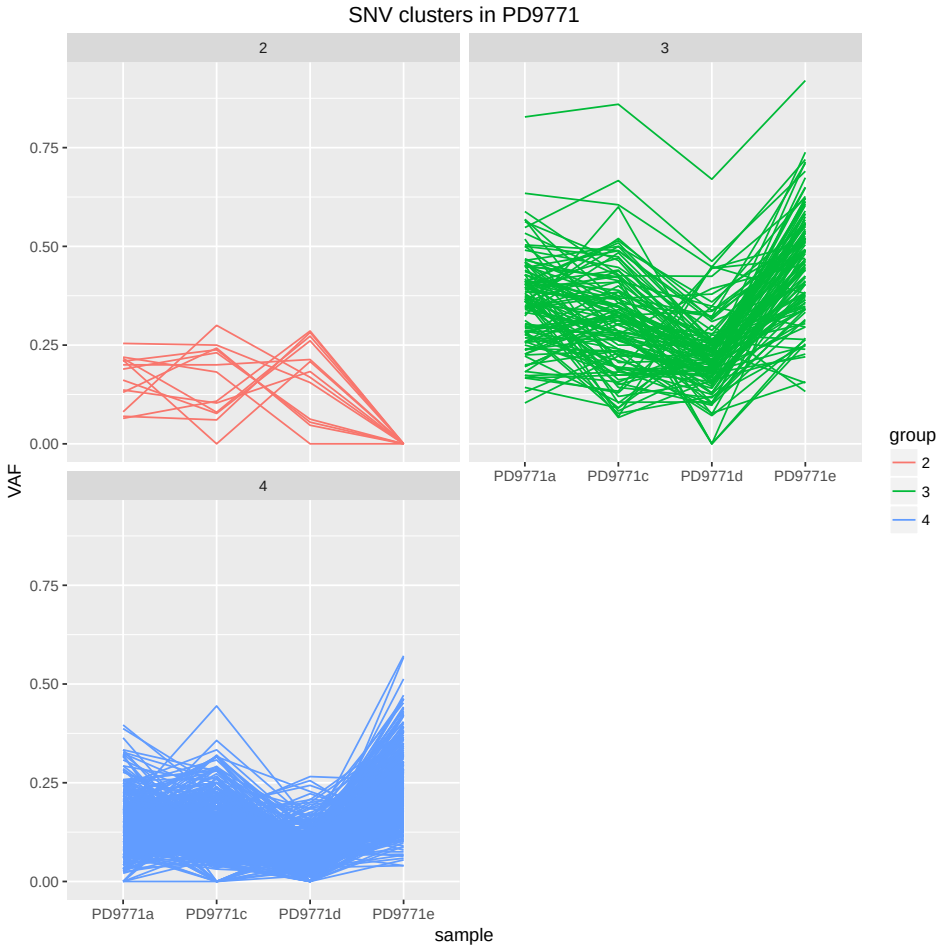

Fig 16: VAF for different SNV clusters of patient PD9771.

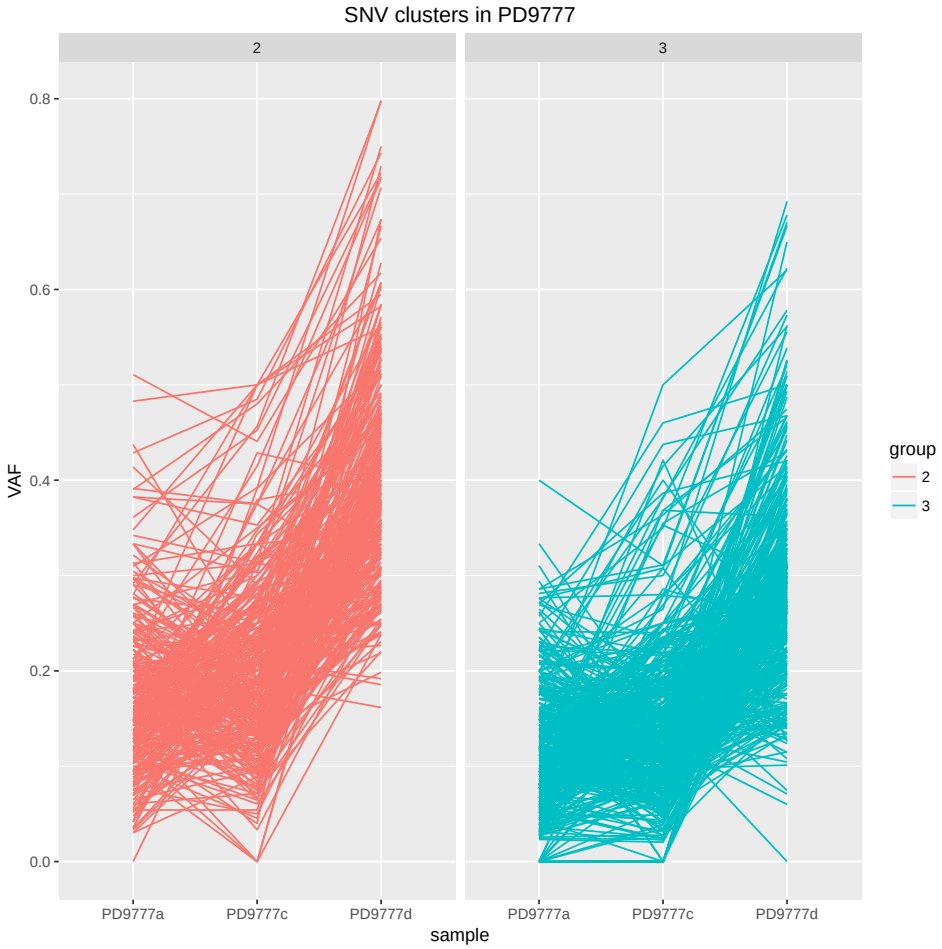

Fig 17: VAF for different SNV clusters of patient PD9777.

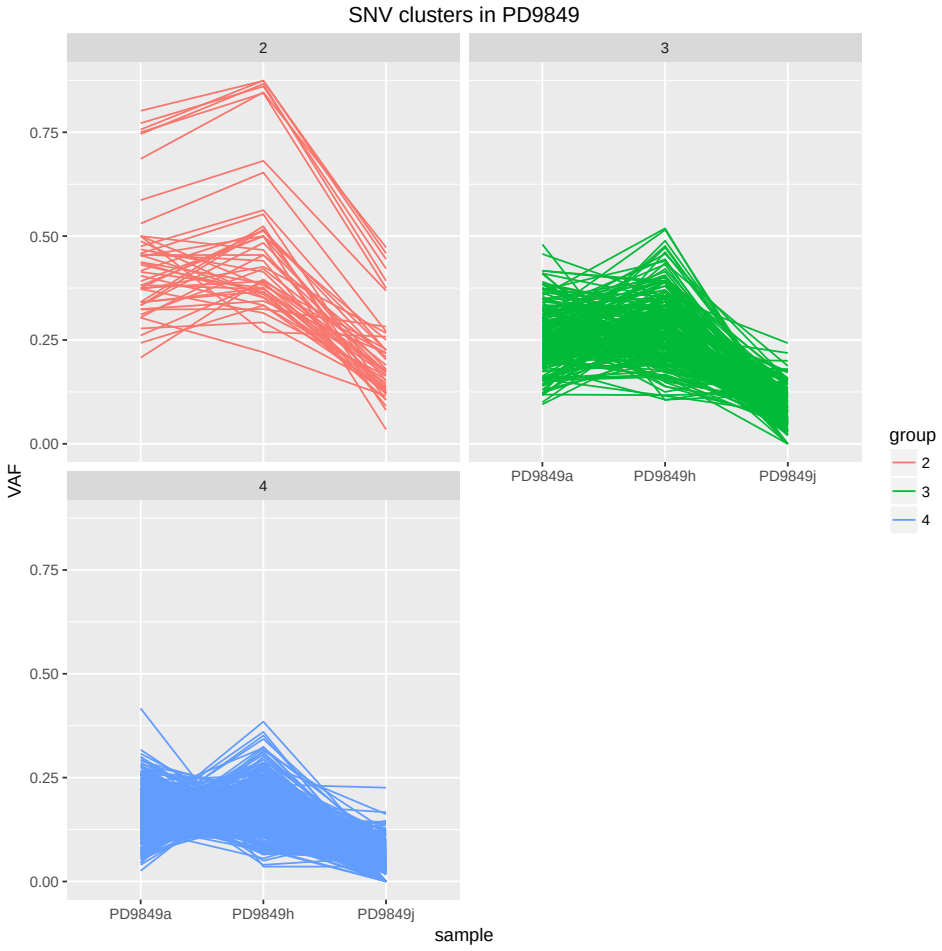

Fig 18: VAF for different SNV clusters of patient PD9849.

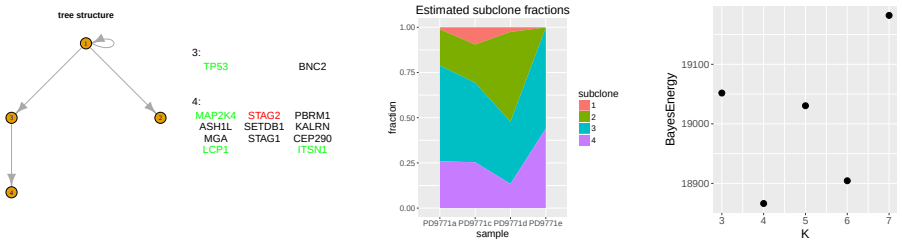

Fig 19: SIFA results for patient PD9771.

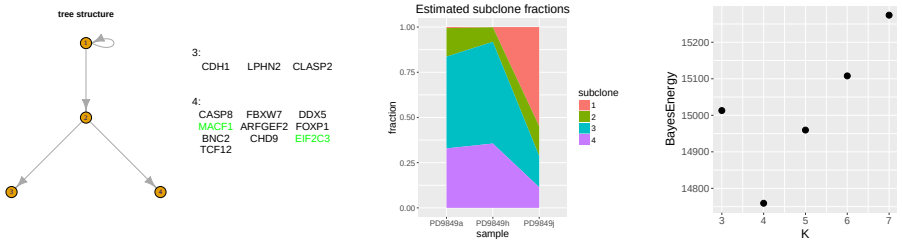

Fig 20: SIFA results for patient PD9849.

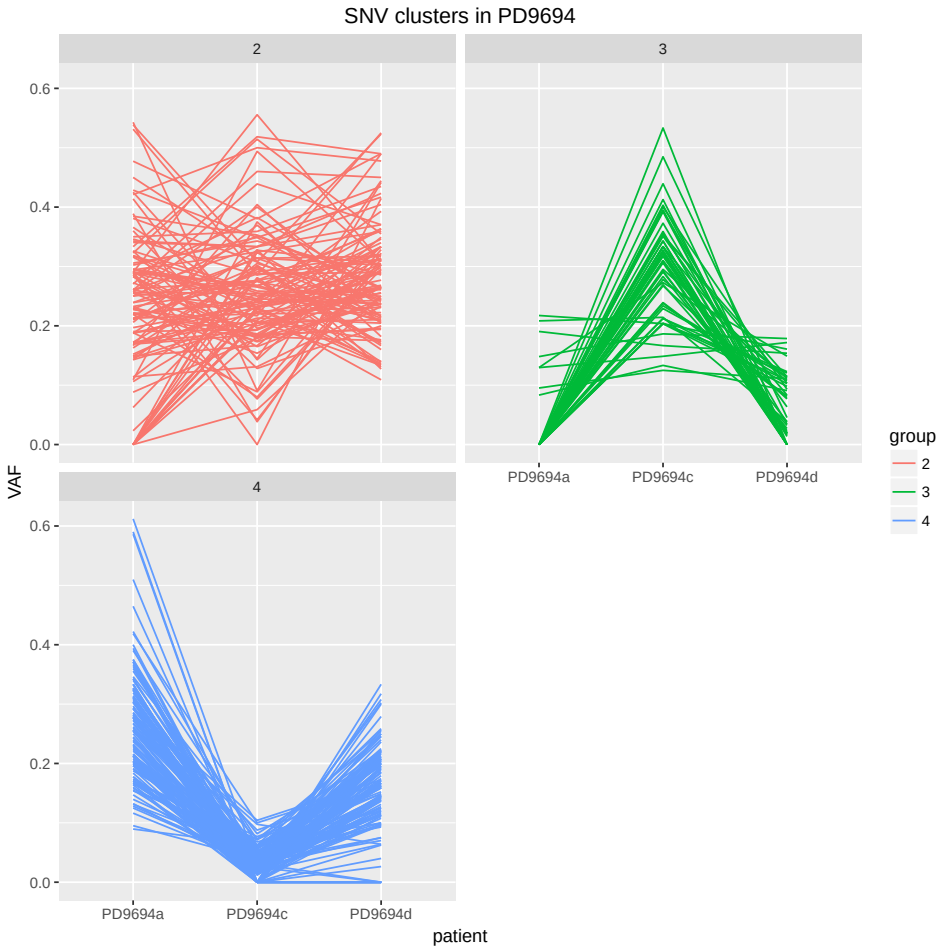

Fig 21: VAF for different SNV clusters of patient PD9694 inferred by Cloe.

Fig 22: VAF for different SNV clusters of patient PD9771 inferred by Cloe.

Fig 23: VAF for different SNV clusters of patient PD9777 inferred by Cloe.

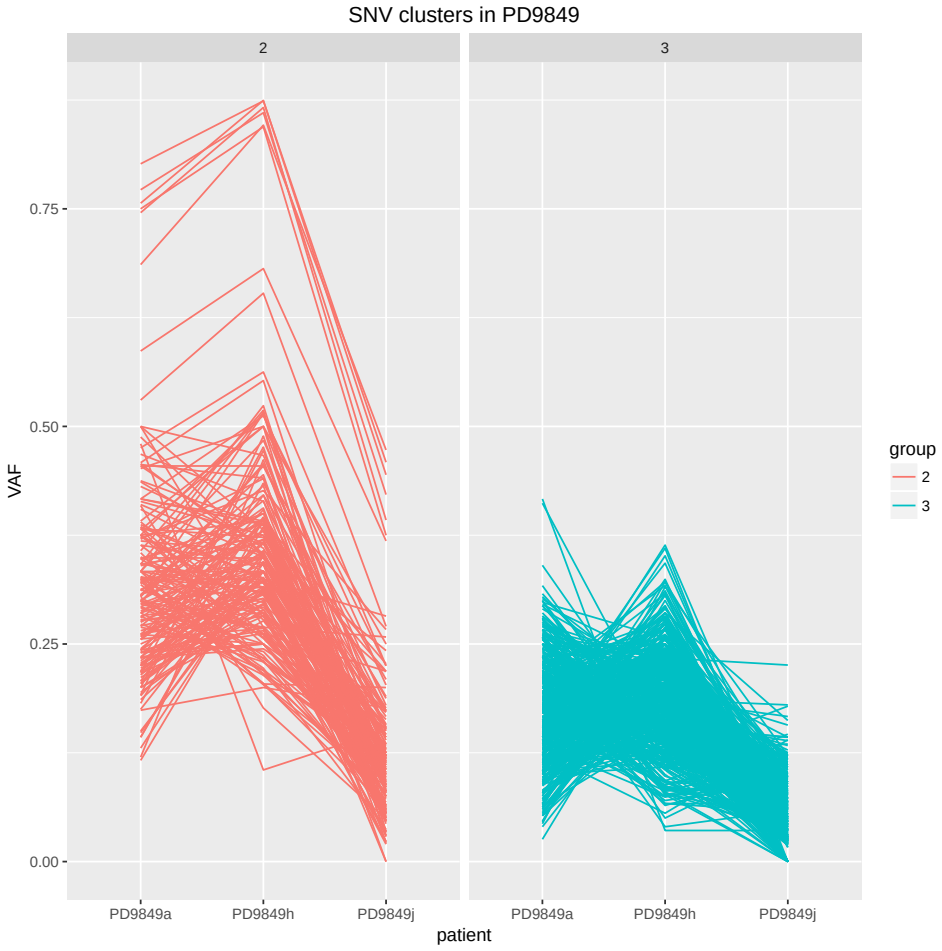

Fig 24: VAF for different SNV clusters of patient PD9849 inferred by Cloe.

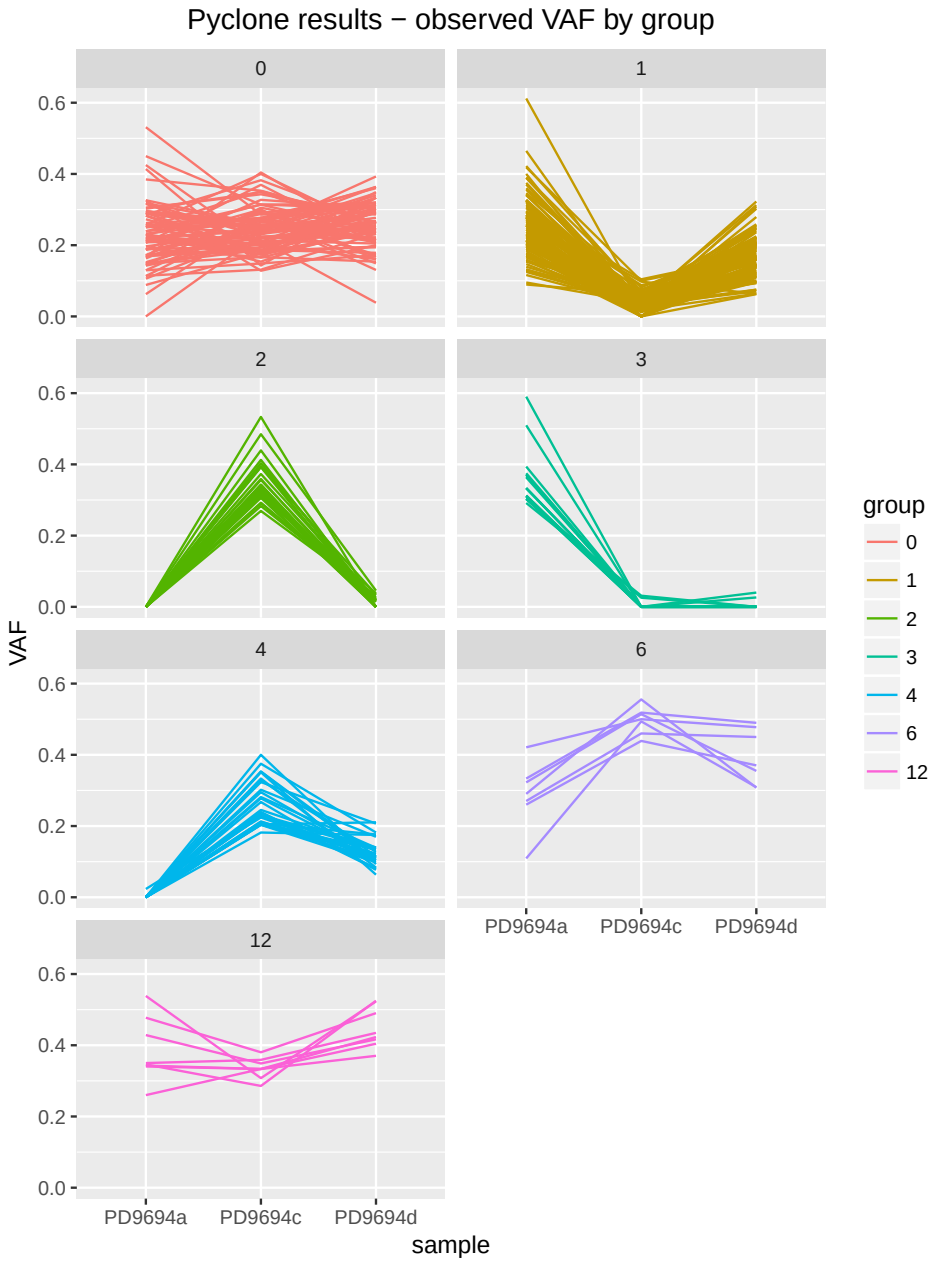

Fig 25: VAF for different SNV clusters of patient PD9694 inferred by Pyclone. Only selected major clusters (with at least 5 mutations) are presented.

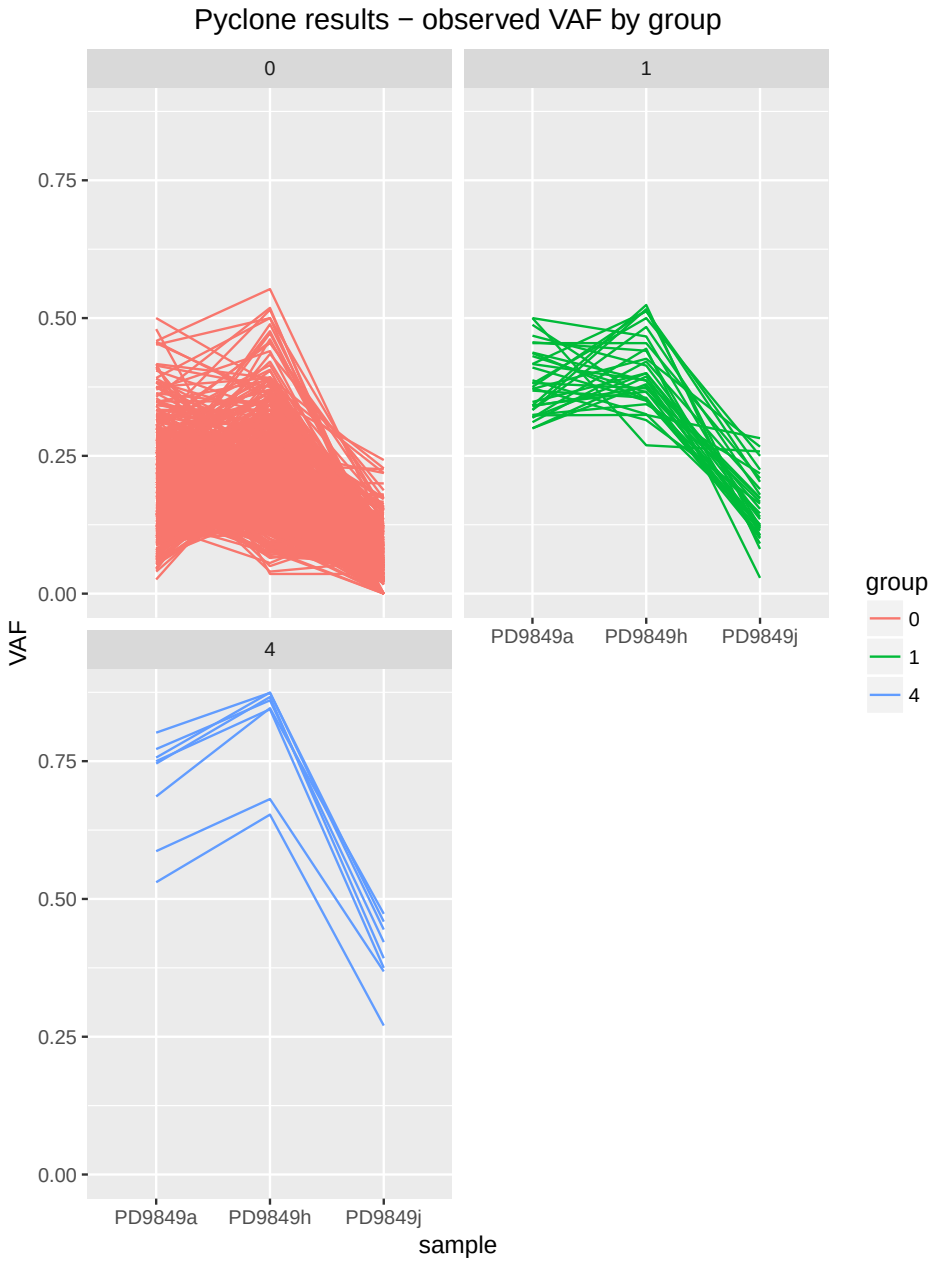

Fig 26: VAF for different SNV clusters of patient PD9849 inferred by Pyclone.

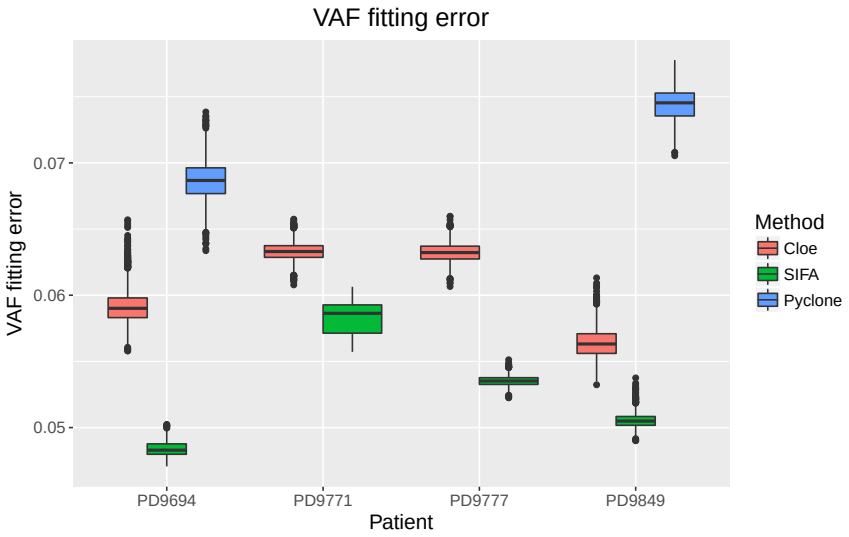

Fig 27: VAF fitting error from Cloe, SIFA, and Pyclone on the breast cancer dataset. Pyclone was only applied to patients PD9694 and PD9849, since it needed estimated major and minor allele copy numbers as input, which was missing for the other two patients.

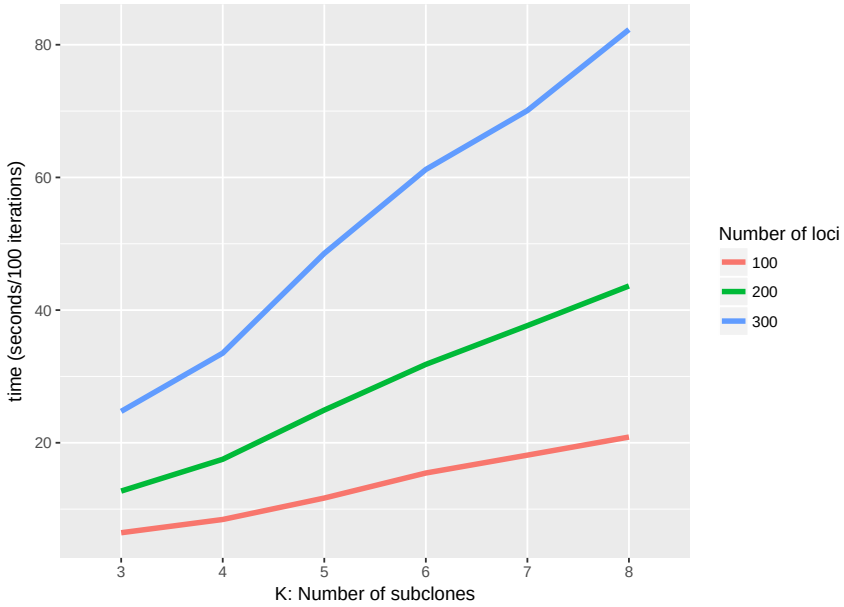

Fig 28: Computing speed of SIFA under different data sizes and number of subclones. Number of Markov chains is set to 8.

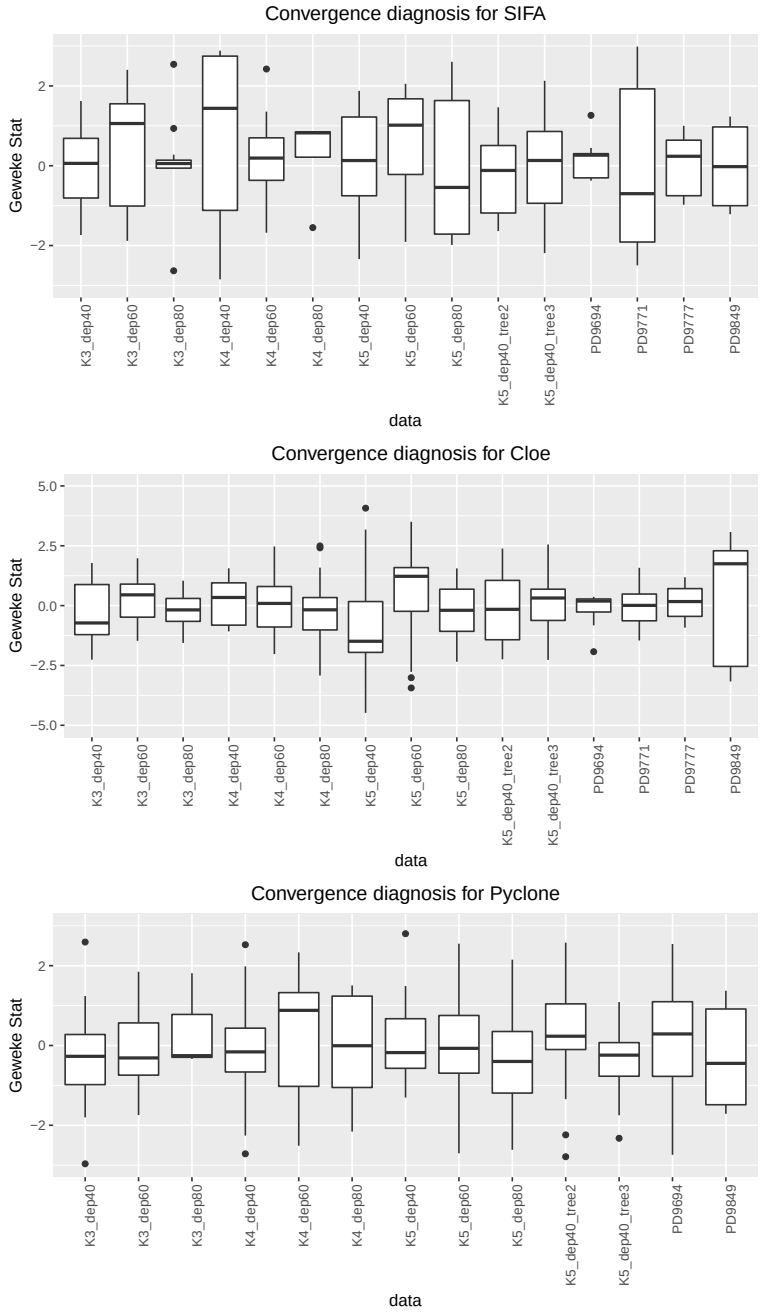

Fig 29: Geweke's Statistics for testing convergence of Bayesian samples. Top, middle,and bottom panels present the statistics for SIFA, Cloe, and Pyclone, respectively.