*Submitted to the Annals of Applied Statistics*

# SUPPLEMENTARY MATERIALS FOR: PHYLOGENY-BASED TUMOR SUBCLONE IDENTIFICATION USING 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 13-16: CNVs detected by SIFA from the breast cancer dataset**

**Figure 17-20: VAF for each SNV cluster identified by SIFA from the breast cancer data set**

**Figure 21-22: 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.

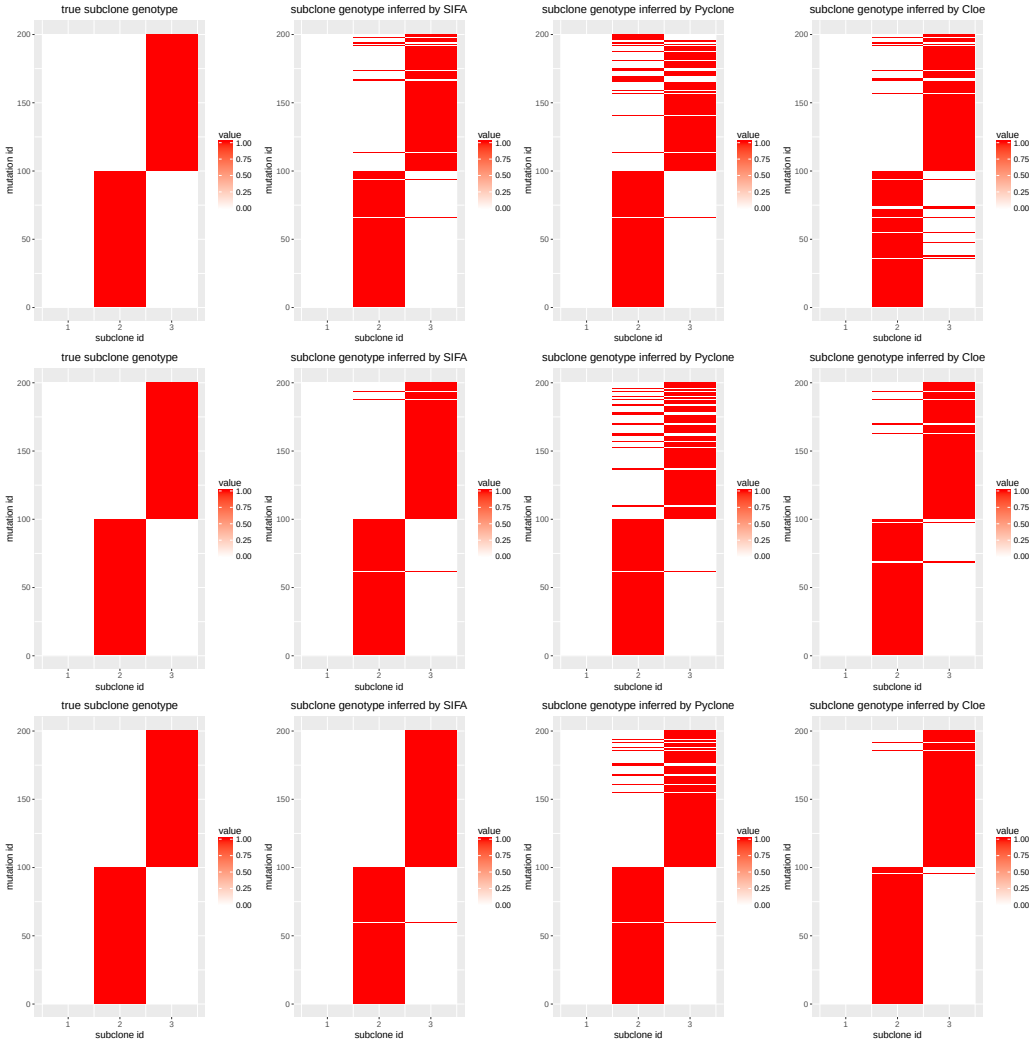


Fig 1: Results for scenarios where $K = 3$. The three rows have simulated sequencing depth $40, 60$, and $80$, respectively.

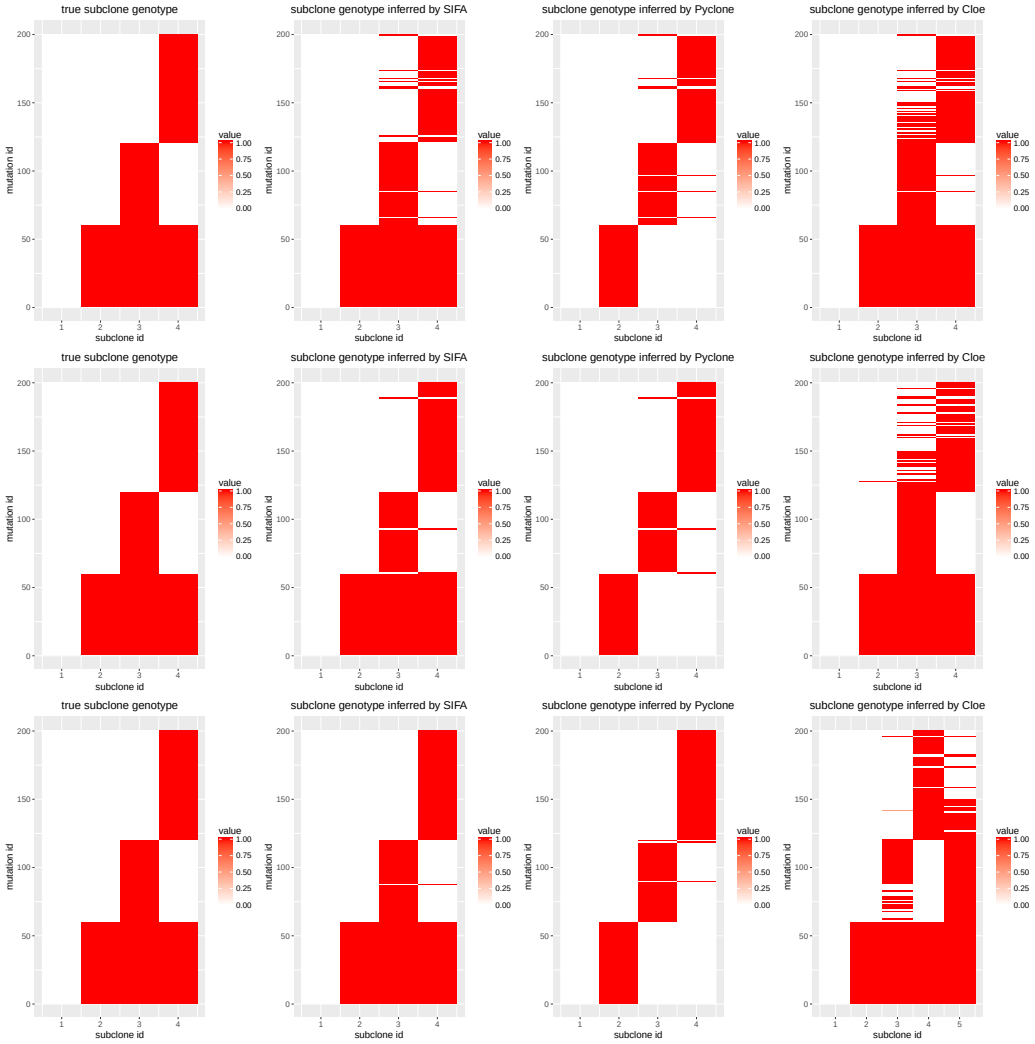


Fig 2: Results for scenarios where $K = 4$. The three rows have simulated sequencing depth $40, 60$, and $80$, respectively.

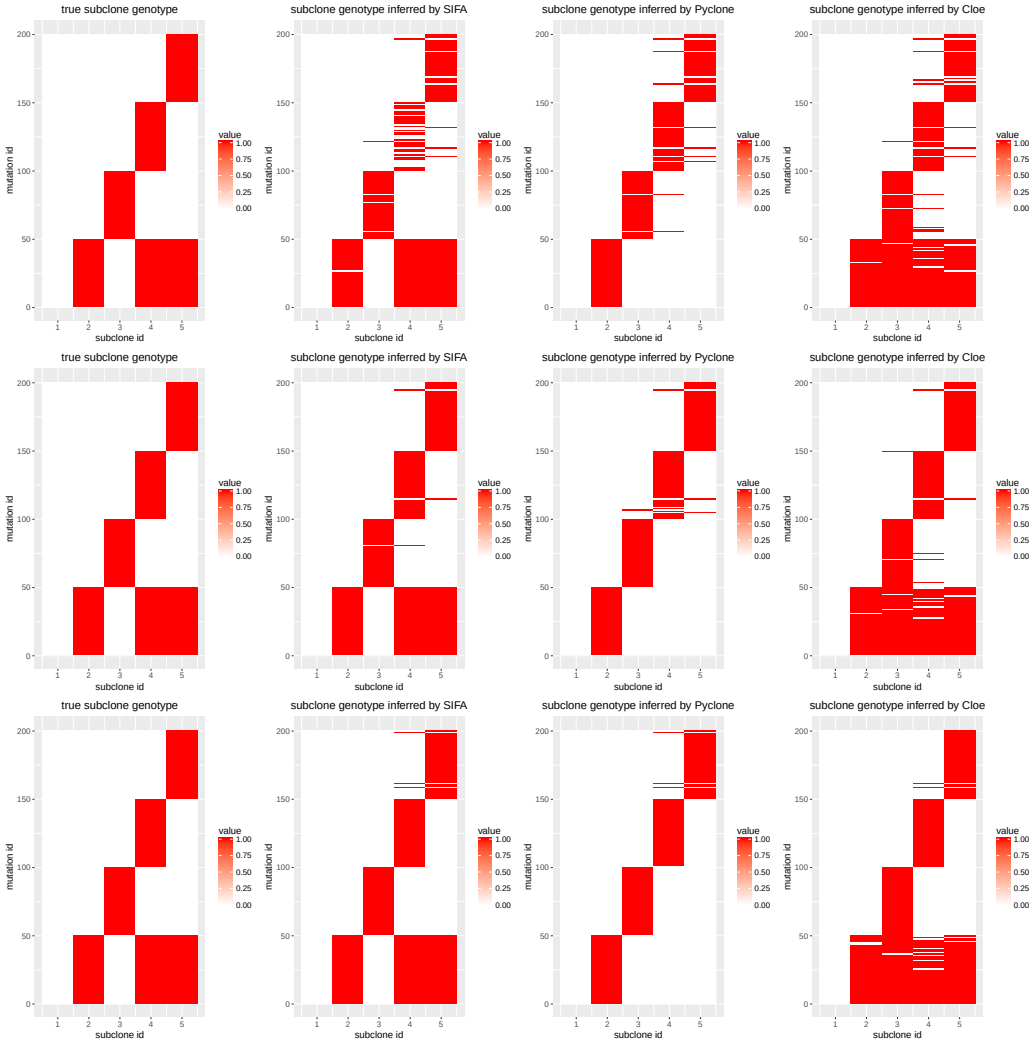


Fig 3: Results for scenarios where $K = 5$. The three rows have simulated sequencing depth $40, 60$, and $80$, respectively.

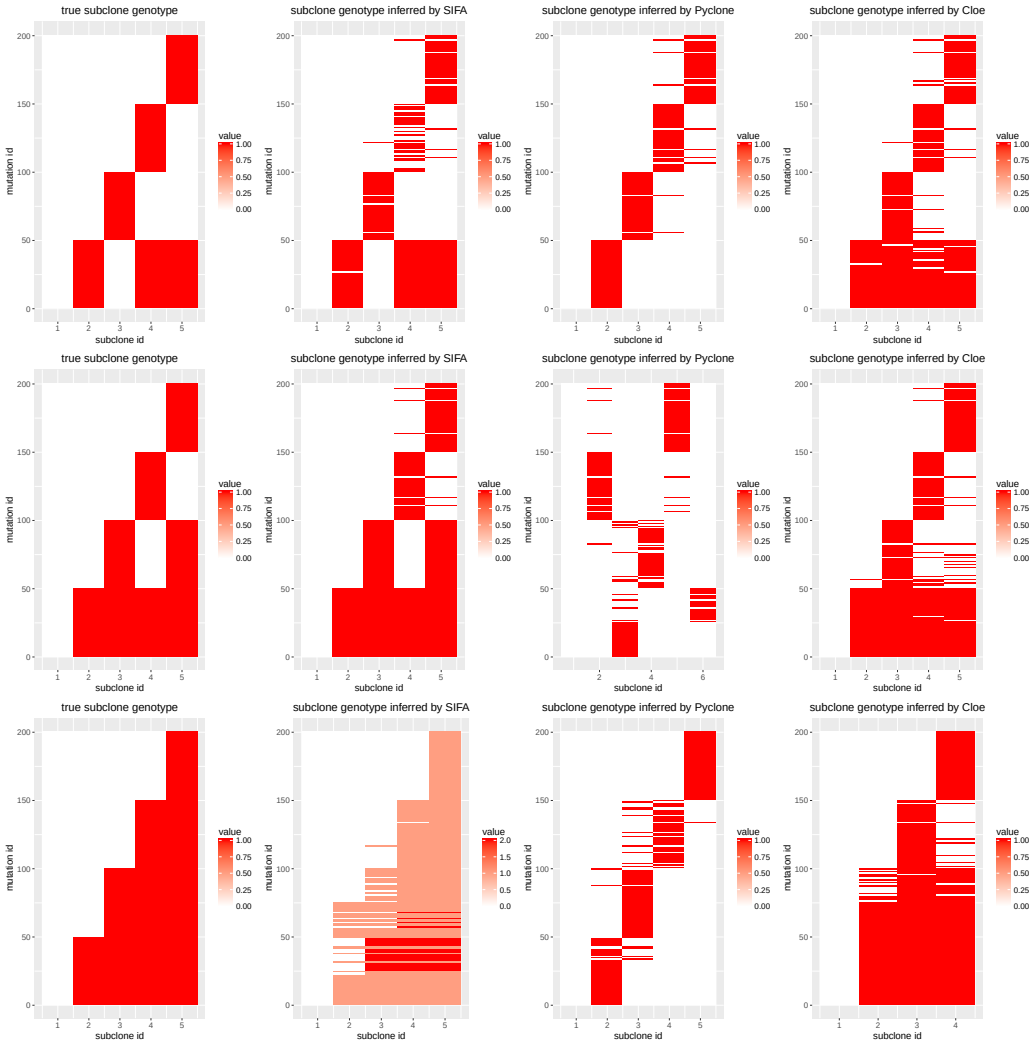


Fig 4: Results for scenarios $K = 5$ with different tree structures.

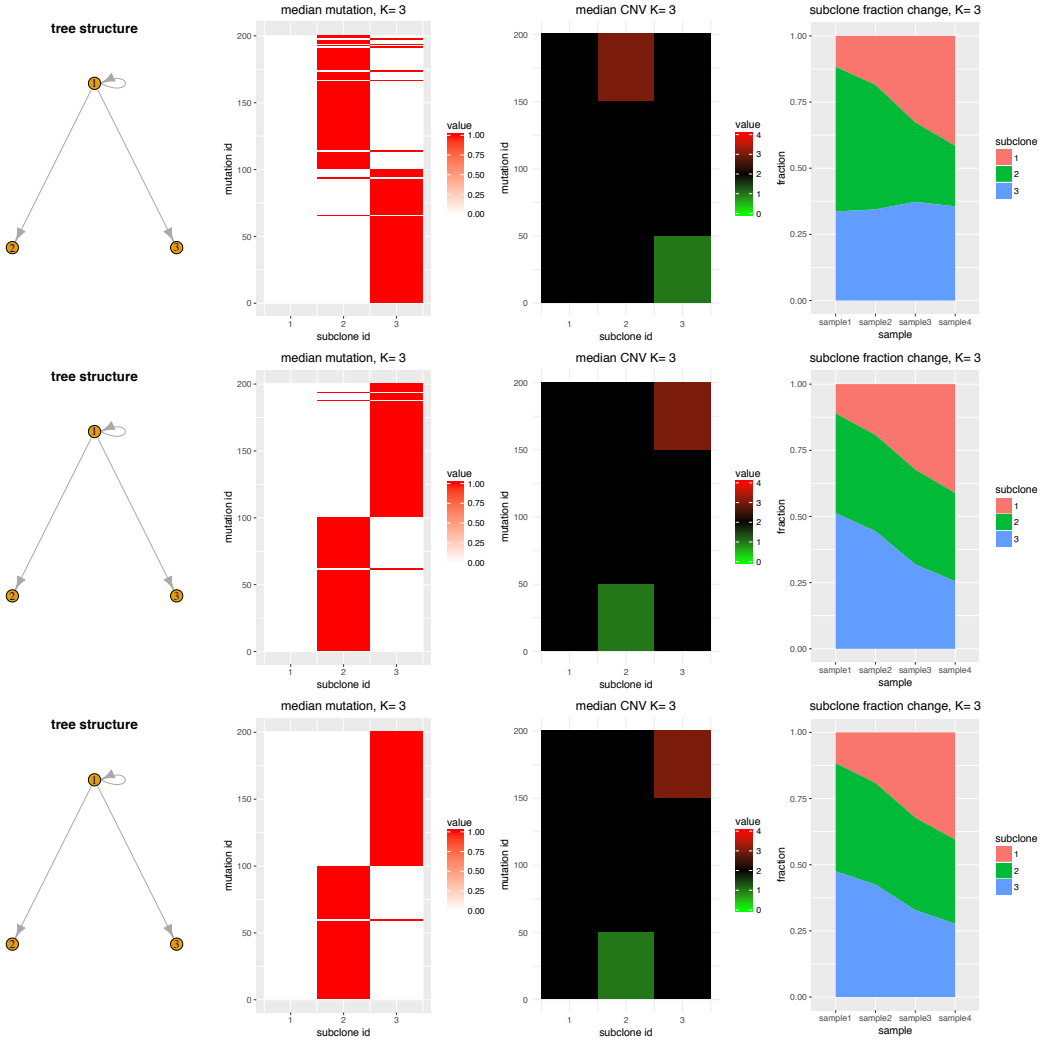


Fig 5: Results for scenarios where $K = 3$. The three rows have simulated sequencing depth $40, 60$, and $80$, respectively.

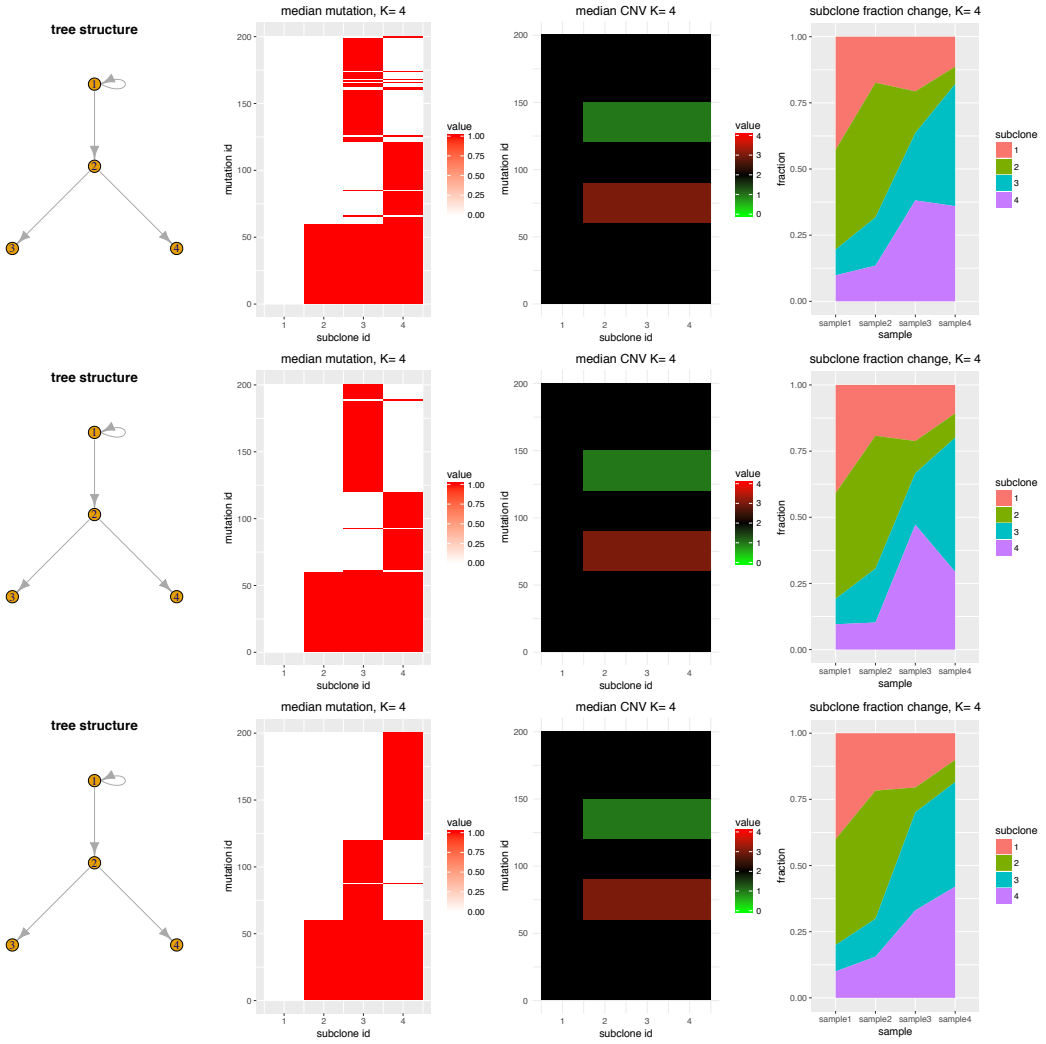


Fig 6: Results for scenarios where $K = 4$. The three rows have simulated sequencing depth $40, 60$, and $80$, respectively.

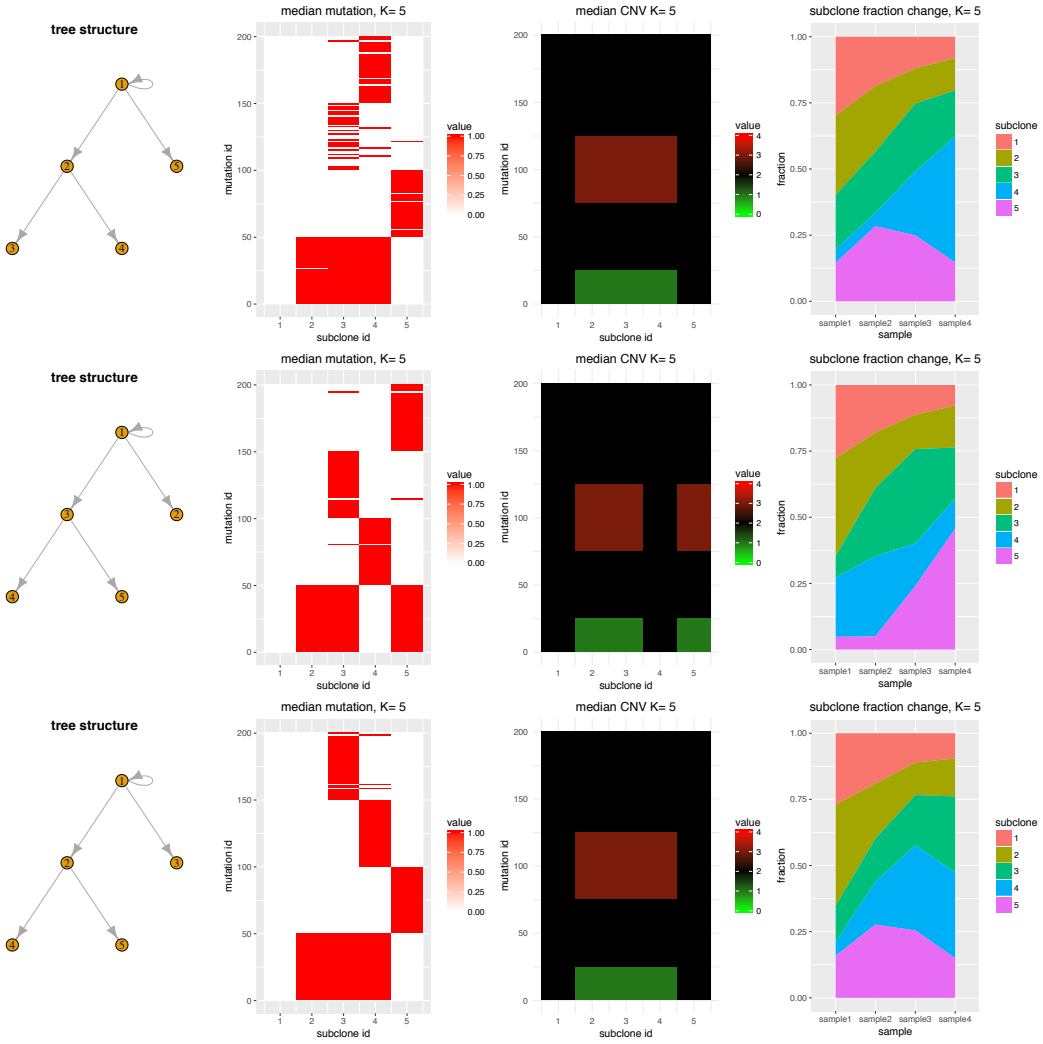


Fig 7: Results for scenarios where $K = 5$. The three rows have simulated sequencing depth $40, 60$, and $80$, respectively.

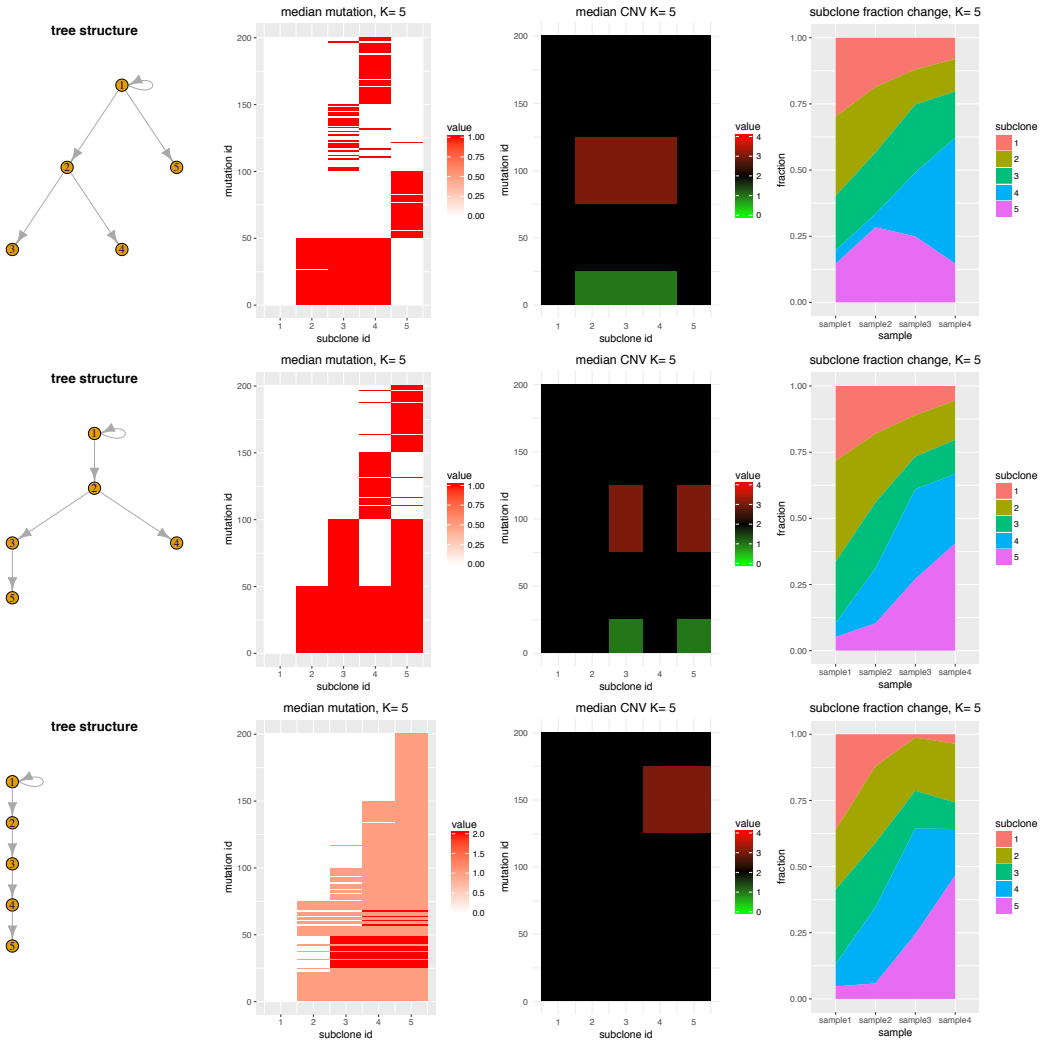


Fig 8: Results for scenarios $K = 5$ with different tree structures.

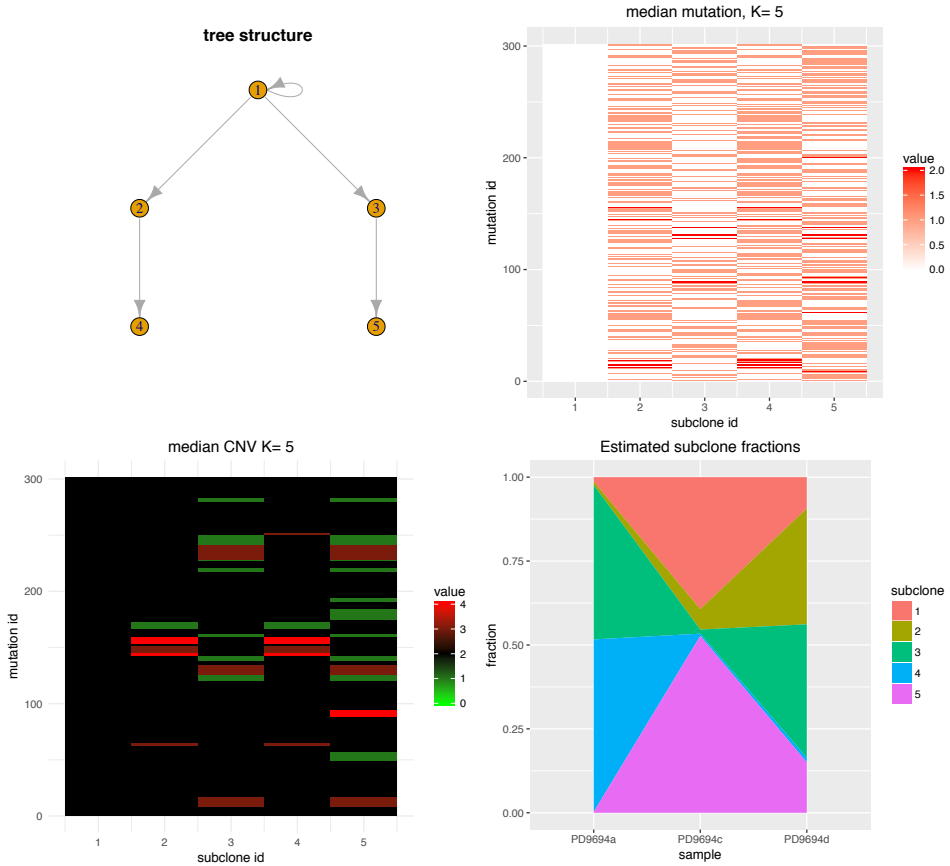


Fig 9: SIFA parameter estimations for patient PD9694.

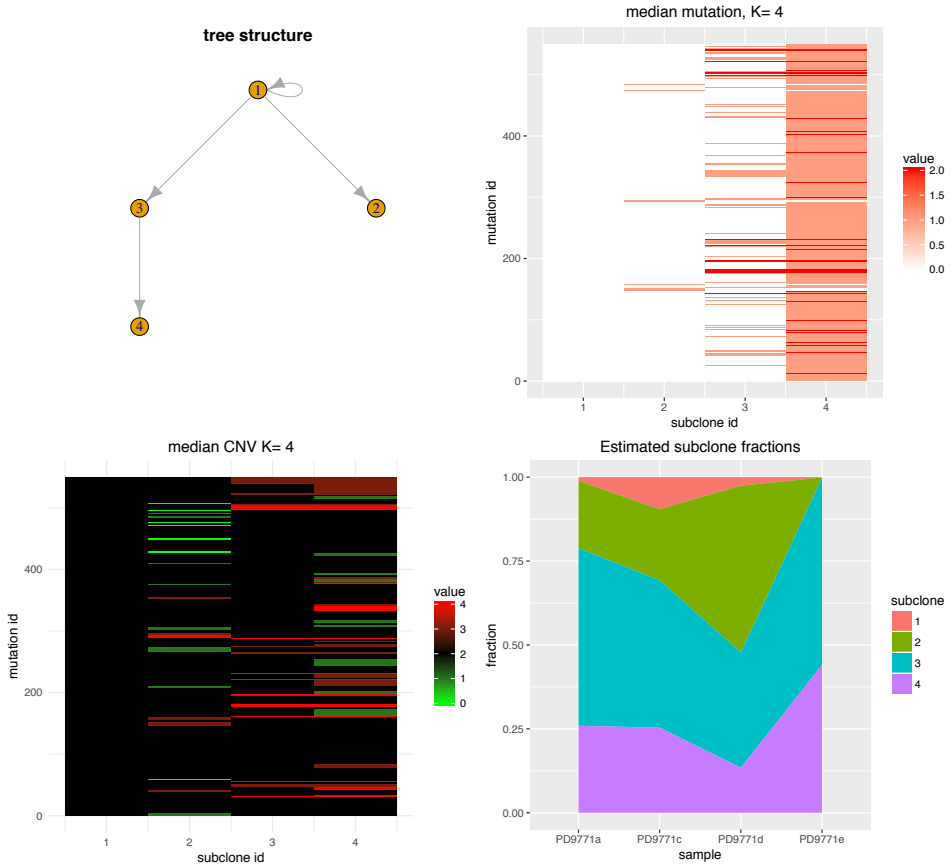


Fig 10: SIFA parameter estimations for patient PD9771.

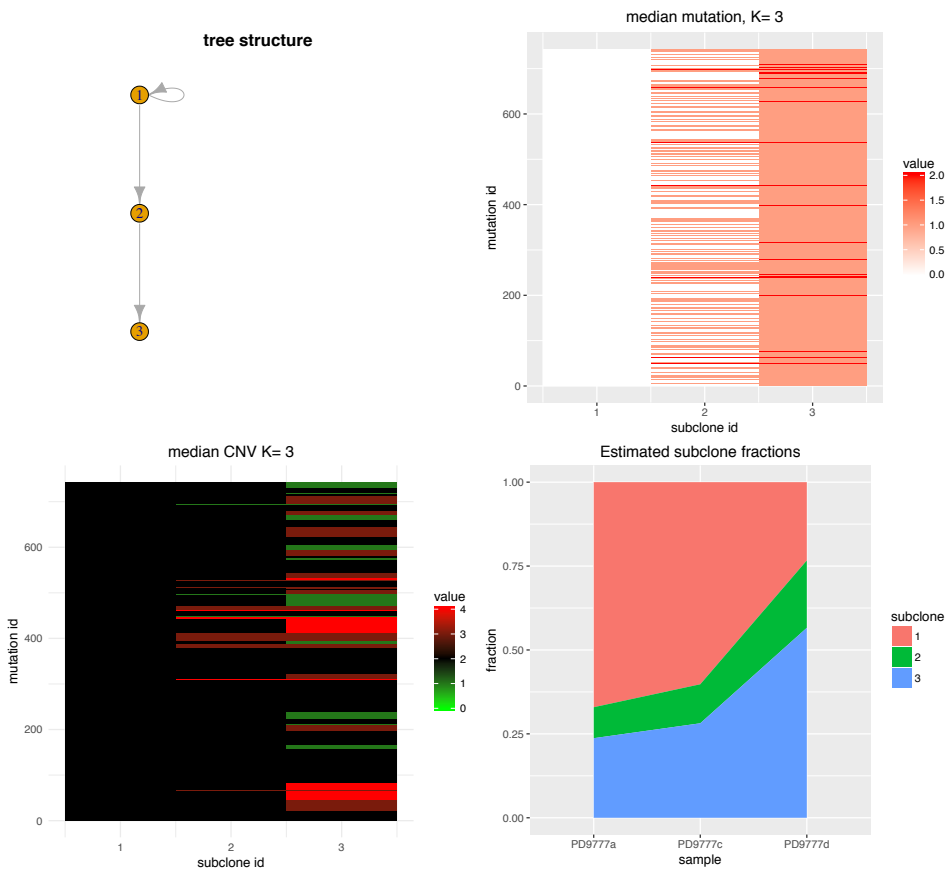


Fig 11: SIFA parameter estimations for patient PD9777.

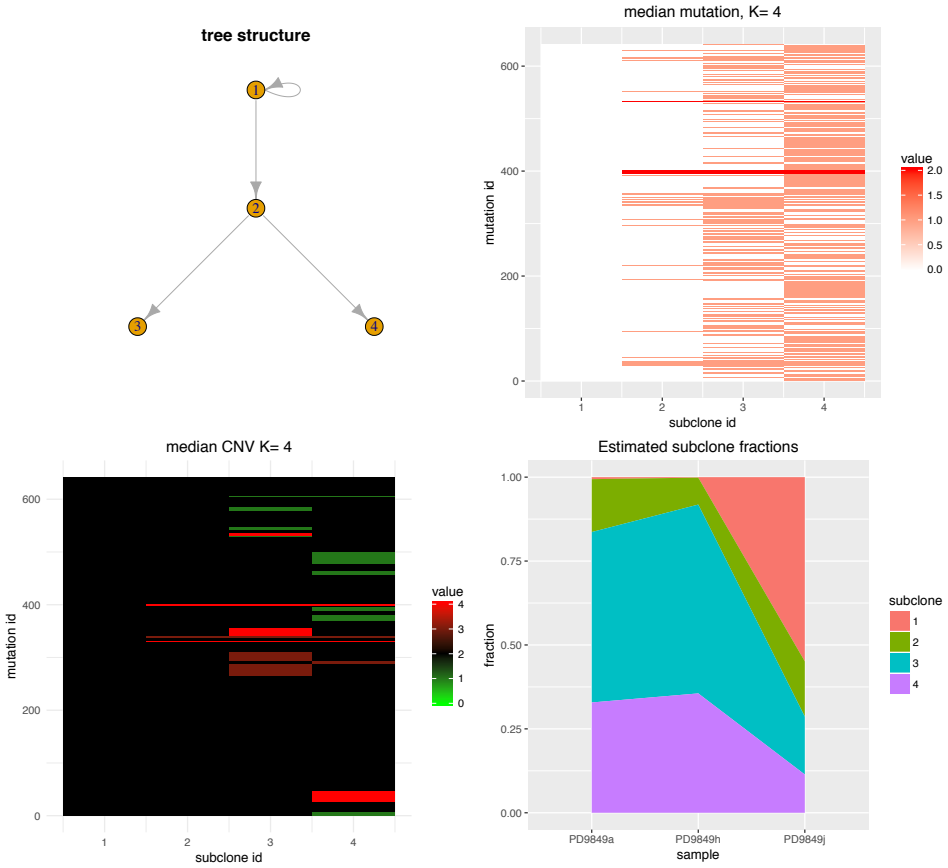


Fig 12: SIFA parameter estimations for patient PD9849.

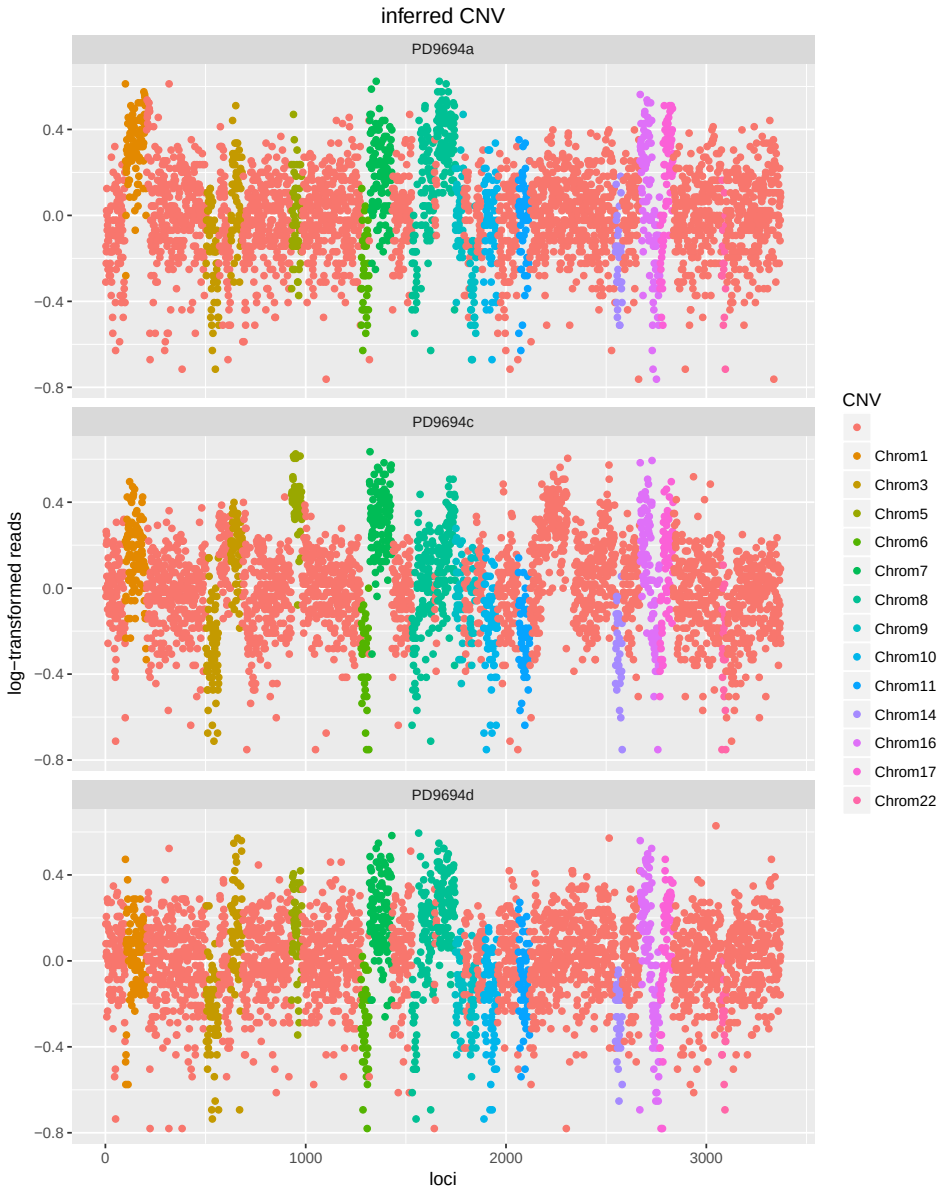


Fig 13: CNVs detected by SIFA for PD9694.

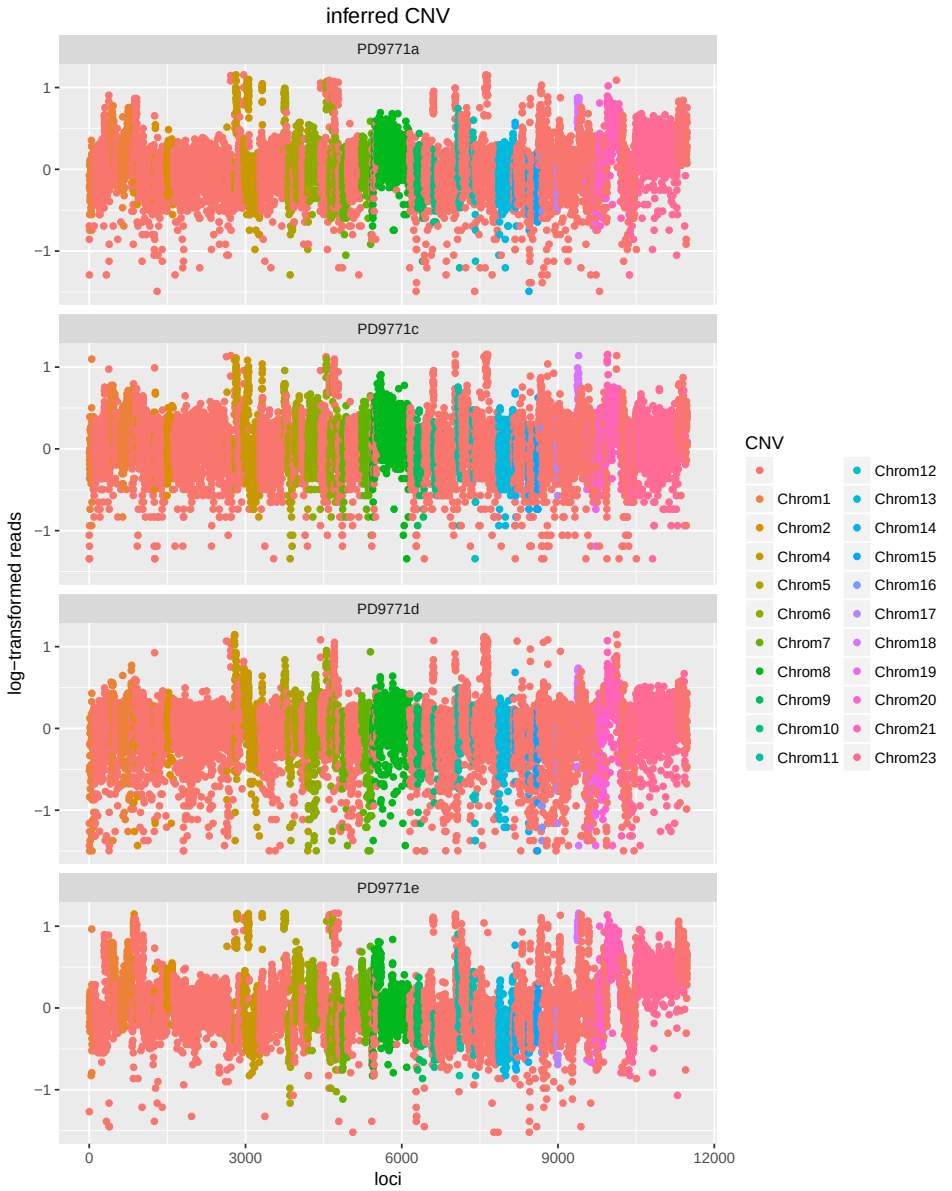


Fig 14: CNVs detected by SIFA for PD9771.

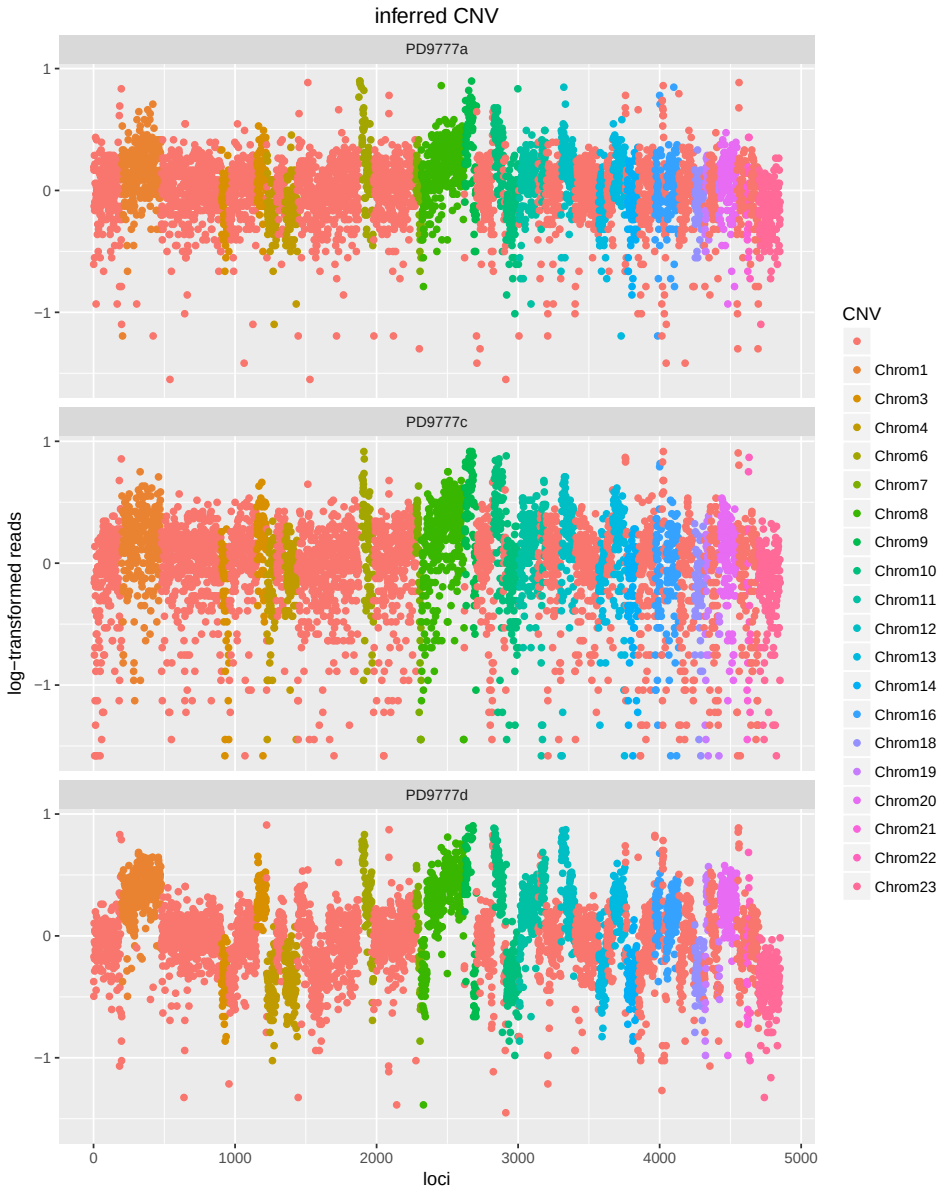


Fig 15: CNVs detected by SIFA for PD9777.

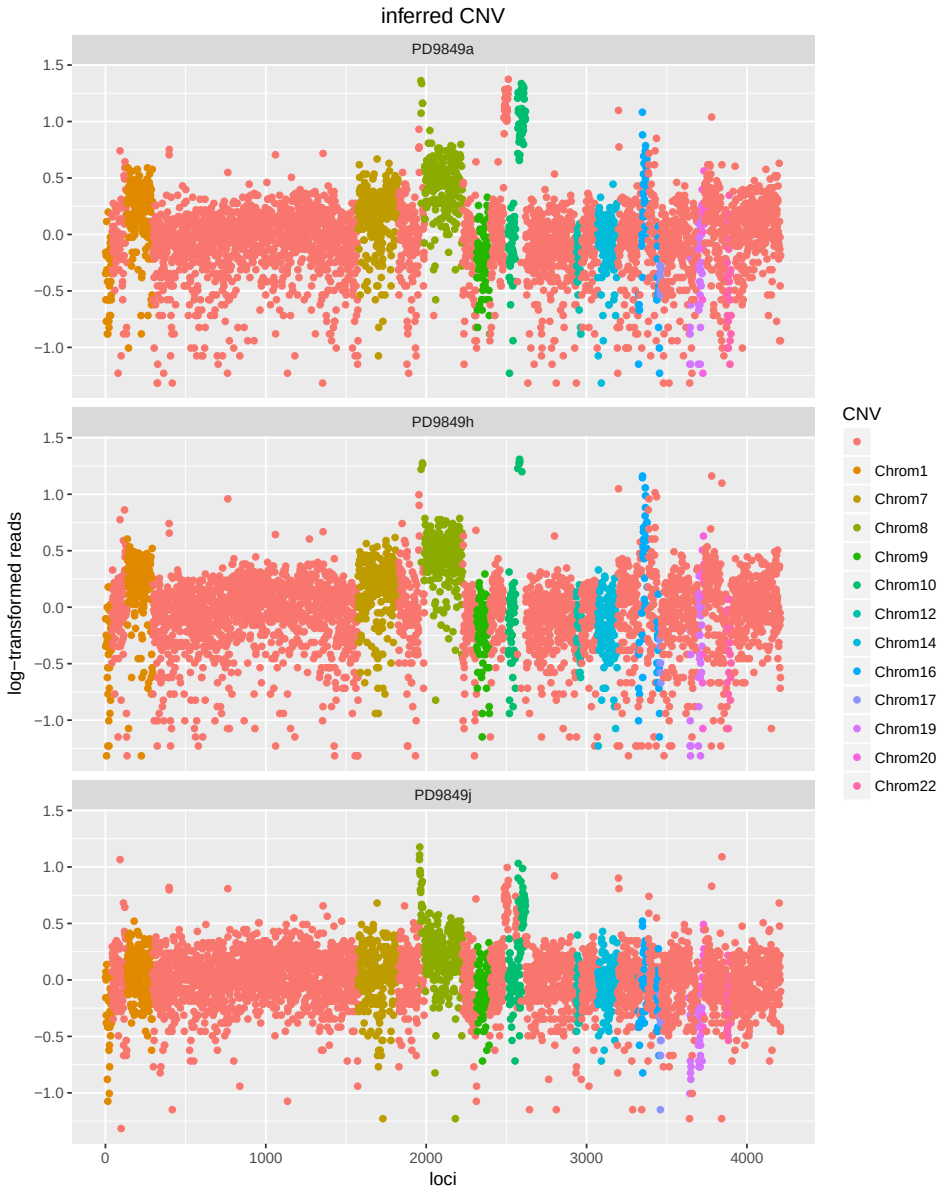


Fig 16: CNVs detected by SIFA for PD9849.

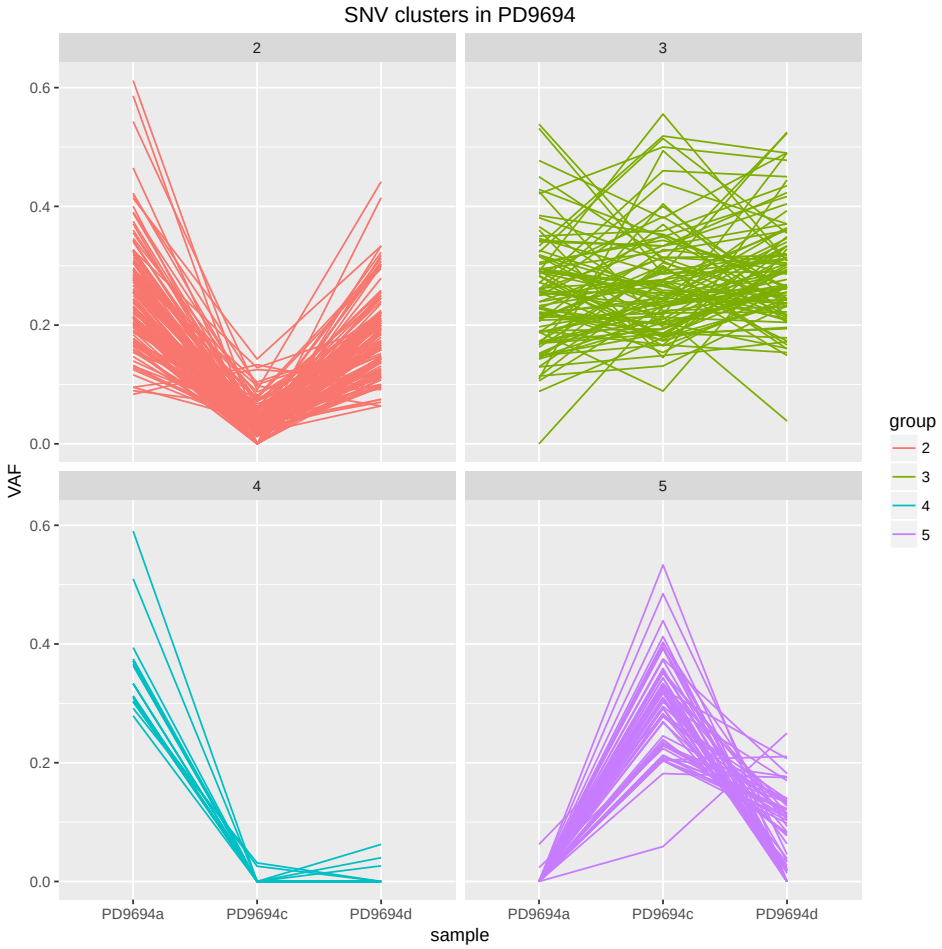


Fig 17: VAF for different SNV clusters of patient PD9694.

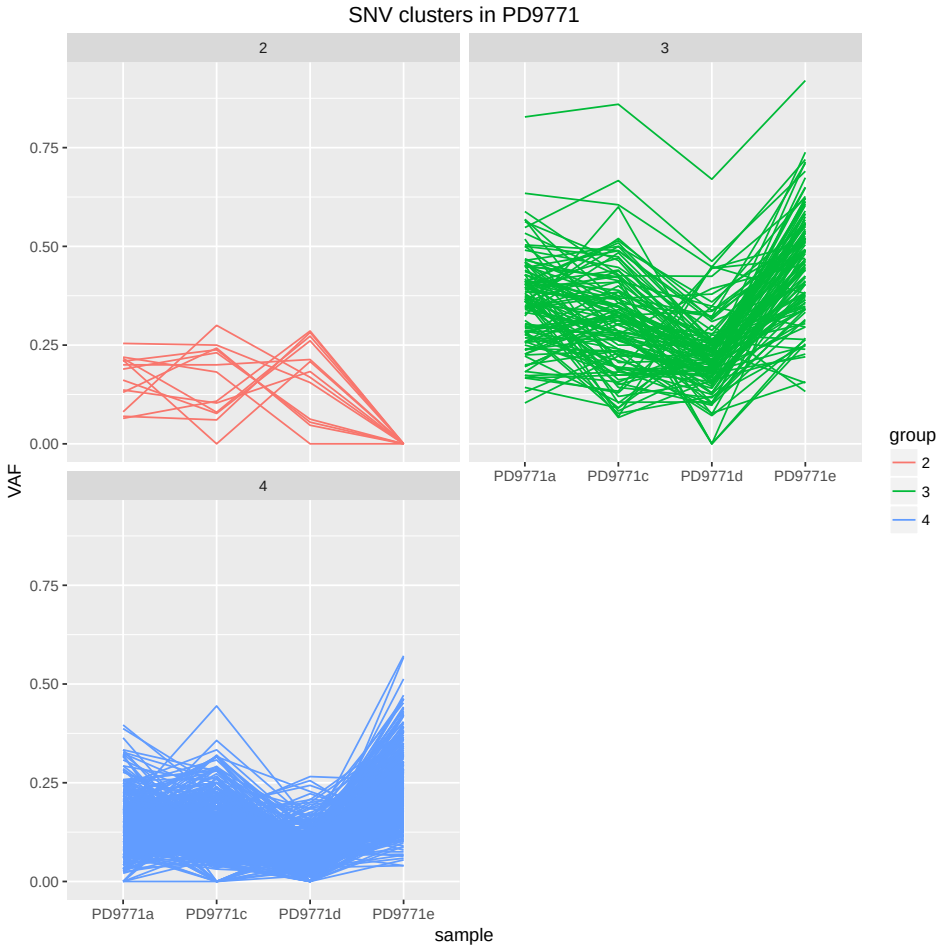


Fig 18: VAF for different SNV clusters of patient PD9771.

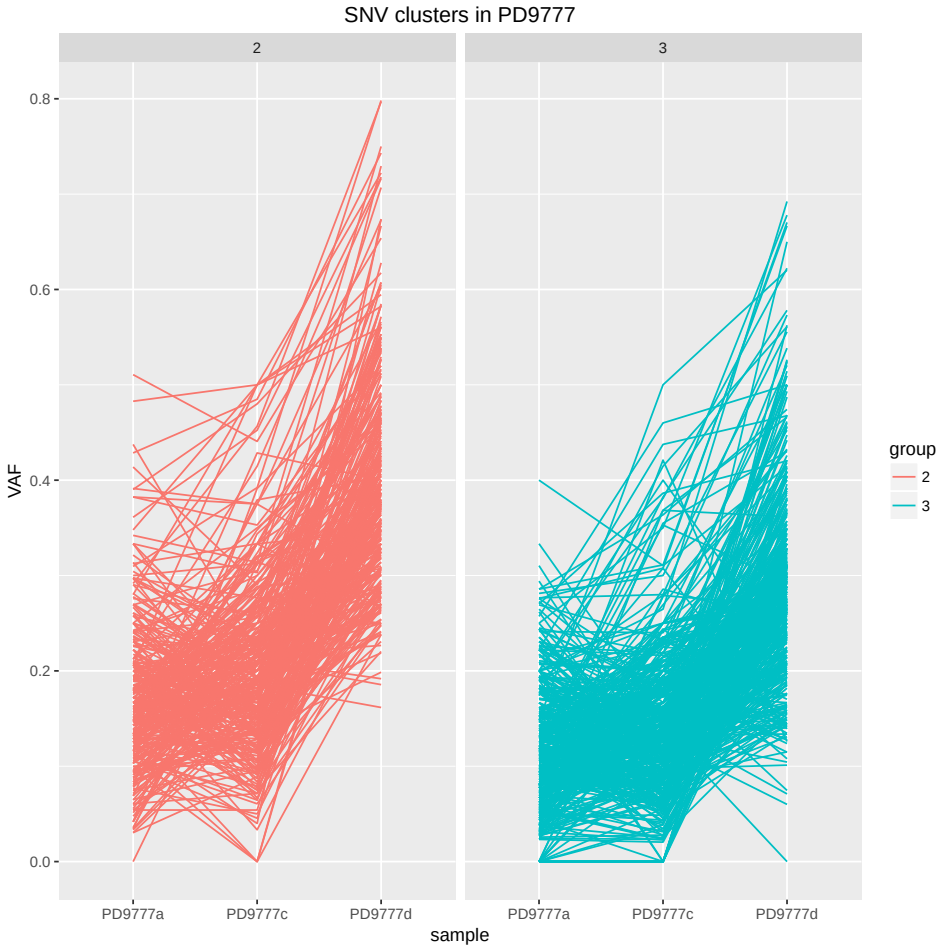


Fig 19: VAF for different SNV clusters of patient PD9777.

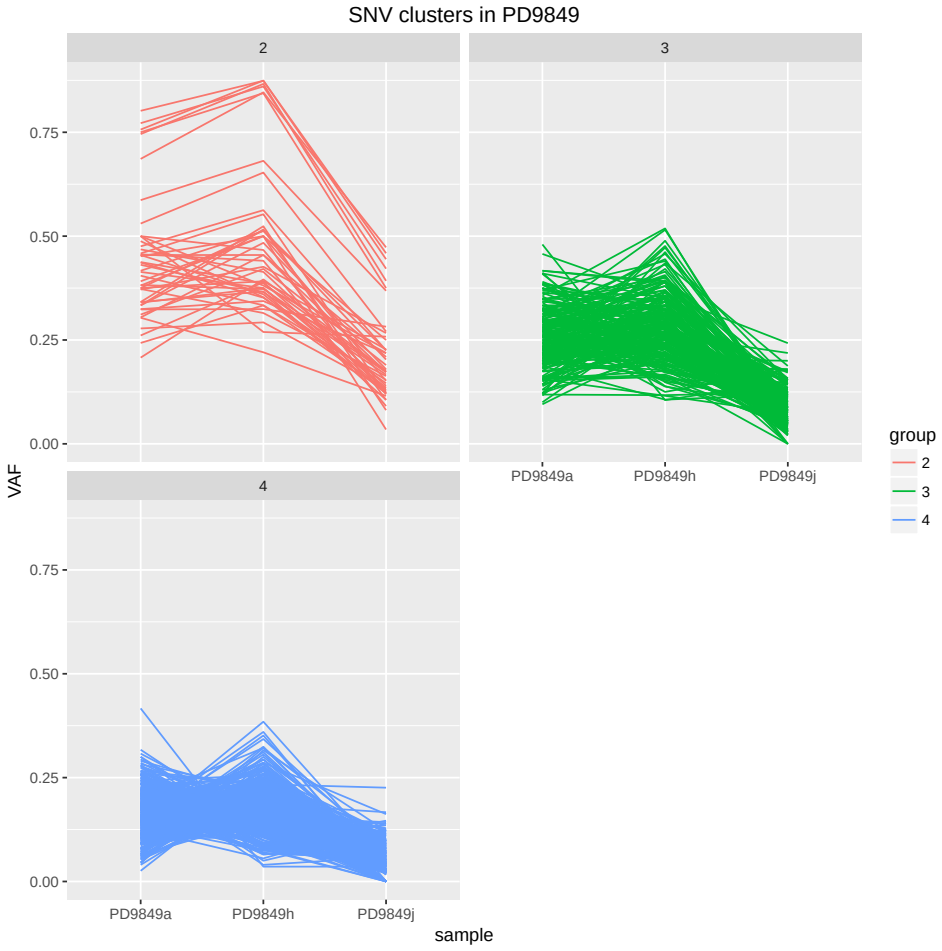


Fig 20: VAF for different SNV clusters of patient PD9849.

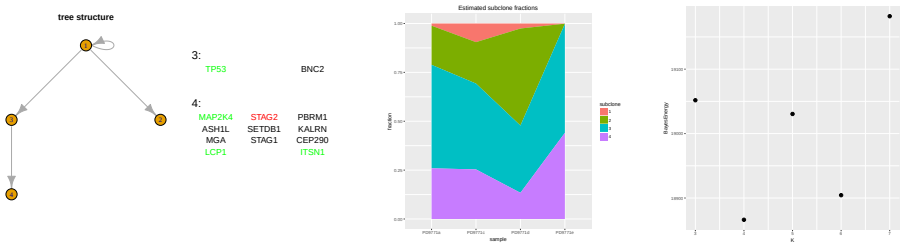


Fig 21: SIFA results for patient PD9771.

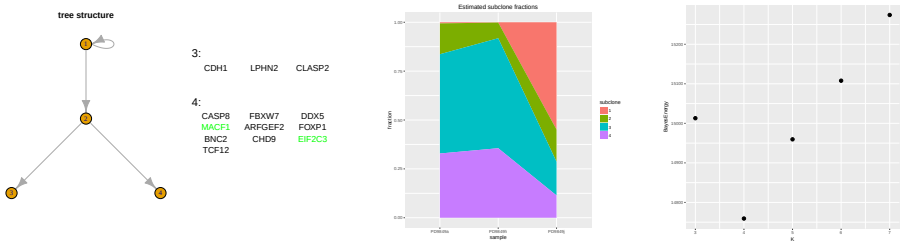


Fig 22: SIFA results for patient PD9849.