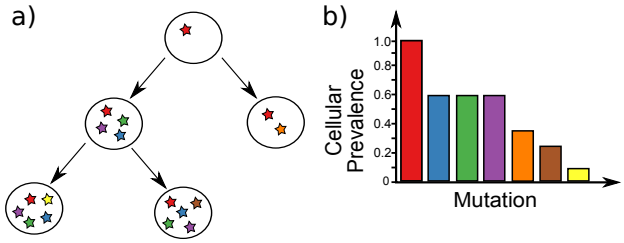
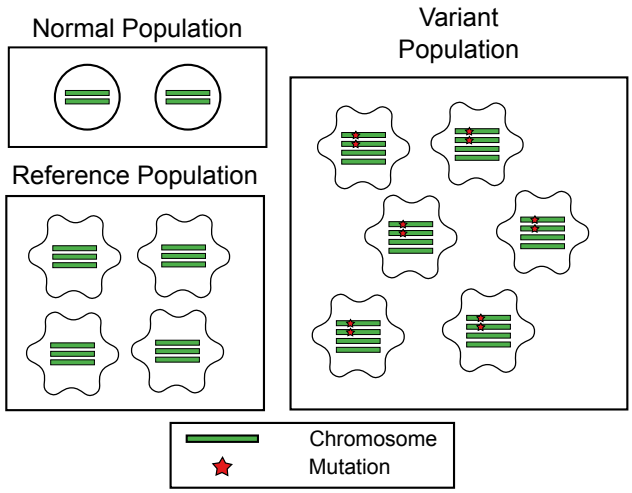
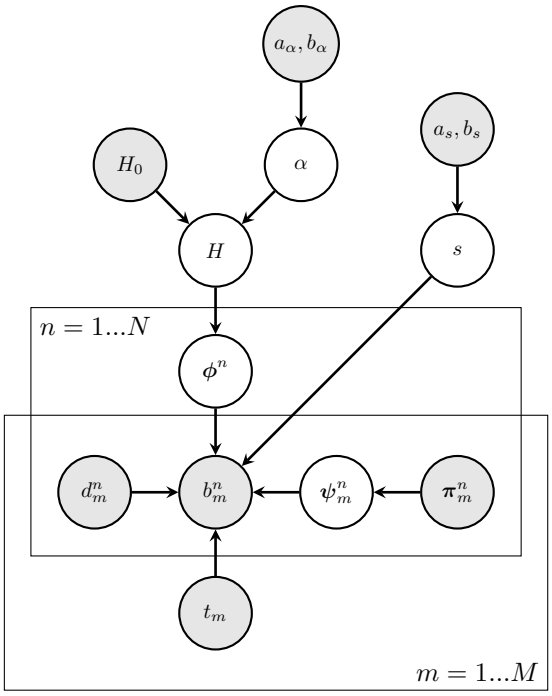
Supplementary Figures for Roth et al., PyClone: Statistical inference of clonal population structure in cancer



Supplementary Figure 1: Clonal evolution model | (**a**) A hypothetical phylogenetic tree generated by clonal expansion via the accumulation of mutations (stars). Unlike traditional phylogenetic trees internal nodes (clones) in the tree may contribute to the observed data, not just the leaf nodes. (**b**) Hypothetical observed cellular prevalences for the mutations in tree. Mutations occurring higher up the tree always have a greater cellular prevalence than their descendants (the same statement need not be true about variant allelic prevalence because of the effect of genotype). Note that the green, blue and purple mutations occur at the same cellular prevalence because they always co-occur in the clones of the tree.

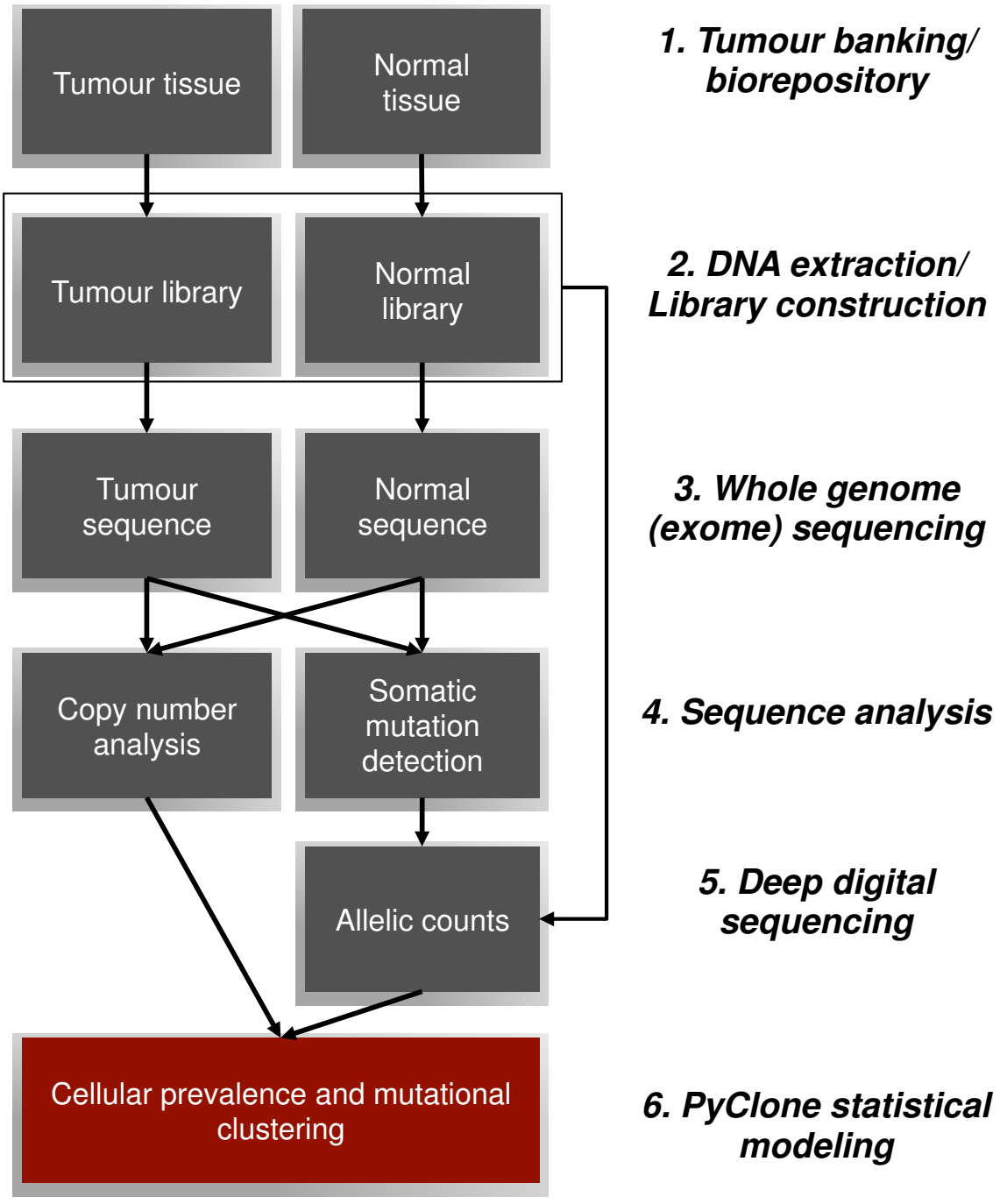


Supplementary Figure 2: PyClone population structure assumptions | Simplified structure of a sample submitted for sequencing. Here we consider the sample with respect to a single mutation (stars). With respect to this mutation we can separate the cells in the sample into three populations: the 'normal population' consists of all normal cells (circular), the 'reference population' consists of cancer cells (irregular) which do not contain the mutation and the 'variant population' consists of all cancer cells with the mutation. To simplify the model we assume all the cells within each population share the same genotype. For example all cells in the variant population in this case have the genotype AABB i.e. two copies of the reference allele, A, and two copies of the variant allele, B. Note that the fraction of cancer cells from the variant population is the cellular prevalence of the mutation which is $\frac{6}{10} = 0.6$ in this example. Due to the effect of heterogeneity and genotype the expected fraction of reads containing the variant allele (variant allelic prevalence) in this example would be $\frac{6 \cdot 4 \cdot \frac{2}{4}}{2 \cdot 2 + 4 \cdot 3 + 6 \cdot 4} = 0.3$.

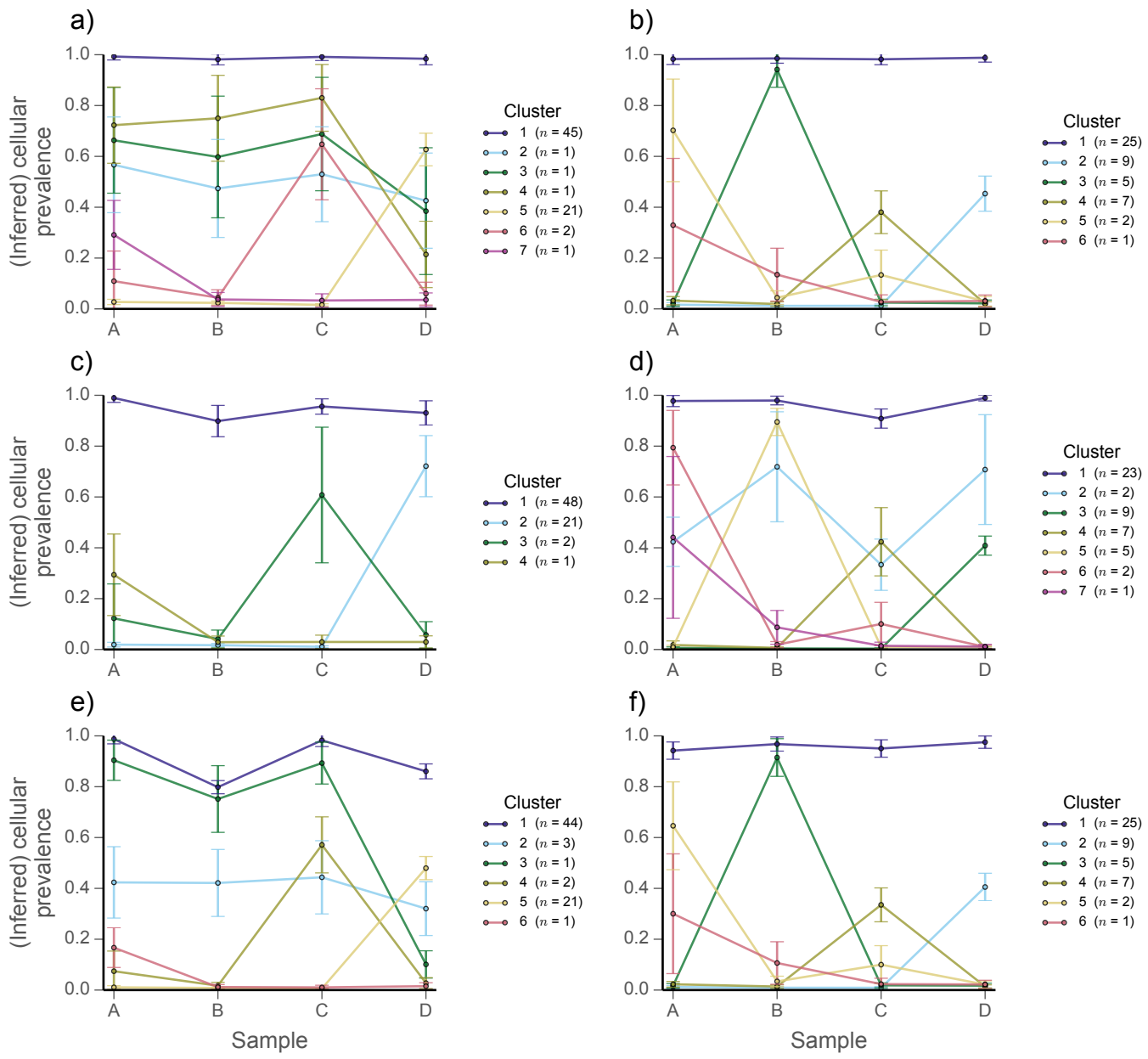


$$
\begin{aligned}
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\
H_0 &= \text{Uniform}([0,1]^M) \\
H|\alpha, H_0 &\sim \text{DP}(\alpha, H_0) \\
\boldsymbol{\phi}^n|H &\sim H \\
\boldsymbol{\psi}_m^n|\boldsymbol{\pi}_m^n &\sim \text{Categorical}(\boldsymbol{\pi}_m^n) \\
\boldsymbol{\psi}_m^n &= (g_{m,\text{N}}^n, g_{m,\text{R}}^n, g_{m,\text{V}}^n) \\
&\text{either} \\
b_m^n|d_m^n, \boldsymbol{\psi}_m^n, \phi_m^n, t_m &\sim \text{Binomial}(d_m^n, \xi(\boldsymbol{\psi}_m^n, \phi_m^n, t_m)) \\
&\text{or} \\
s|a, b &\sim \text{Gamma}(a_s, b_s) \\
b_m^n|d_m^n, \boldsymbol{\psi}_m^n, \phi_m^n, t_m, s &\sim \text{BetaBinomial}(d_m^n, \xi(\boldsymbol{\psi}_m^n, \phi_m^n, t_m), s) \\
&\text{where} \\
\xi(\boldsymbol{\psi}, \phi, t) &= \tfrac{(1-t)c(g_\text{N})}{Z}\mu(g_\text{N}) + \tfrac{t(1-\phi)c(g_\text{R})}{Z}\mu(g_\text{R}) + \\
&\quad \tfrac{t\phi c(g_\text{V})}{Z}\mu(g_\text{V}) \\
Z &= (1-t)c(g_\text{N}) + t(1-\phi)c(g_\text{R}) + \\
&\quad t\phi c(g_\text{V})
\end{aligned}
$$

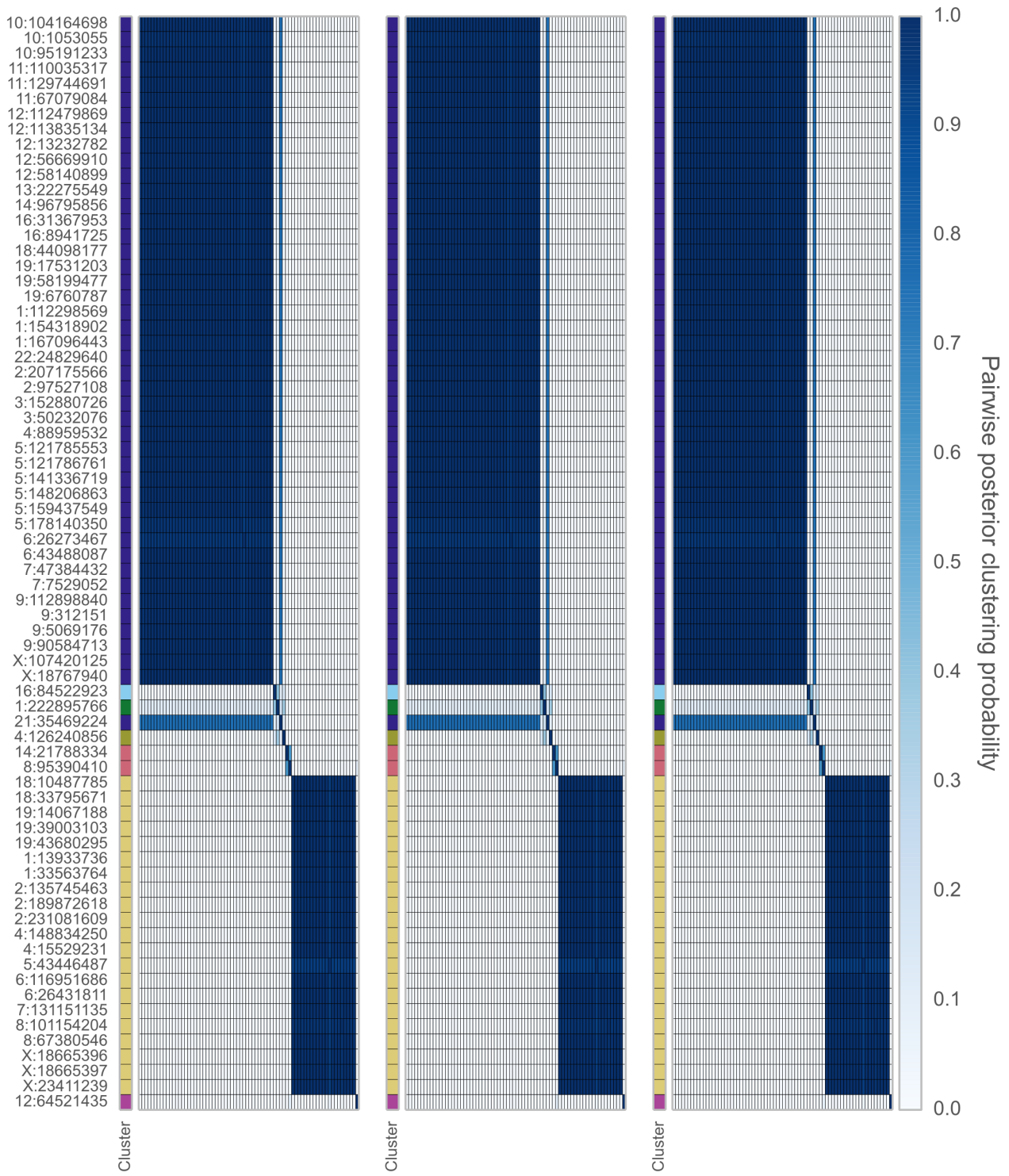
Supplementary Figure 3: Probabilistic graphical representation of the PyClone model | The model assumes the observed count data for the n^{th} mutation is dependent on the cellular prevalence of the mutation as well as the state of the normal, reference and variant populations. The cellular prevalence of mutation n across the M samples, $\boldsymbol{\phi^n}$, is drawn from a Dirichlet Process (DP) prior to allow mutations to cluster and the number of clusters to be inferred. For brevity we show the multi-sample version of PyClone which generalises the single sample case (M=1). We also show the model with either the Binomial or Beta Binomial emission densities. For all analyses conducted in this paper we set vague priors of $a_\alpha = 1, b_\alpha = 10^{-3}$ for the DP concentration parameter α and $a_s = 1, b_s = 10^{-4}$ for the Beta Binomial precision parameter s. The Gamma distributions are parametrised in terms of the shape, a, and rate, b, parameters (see **Supplementary Note**).



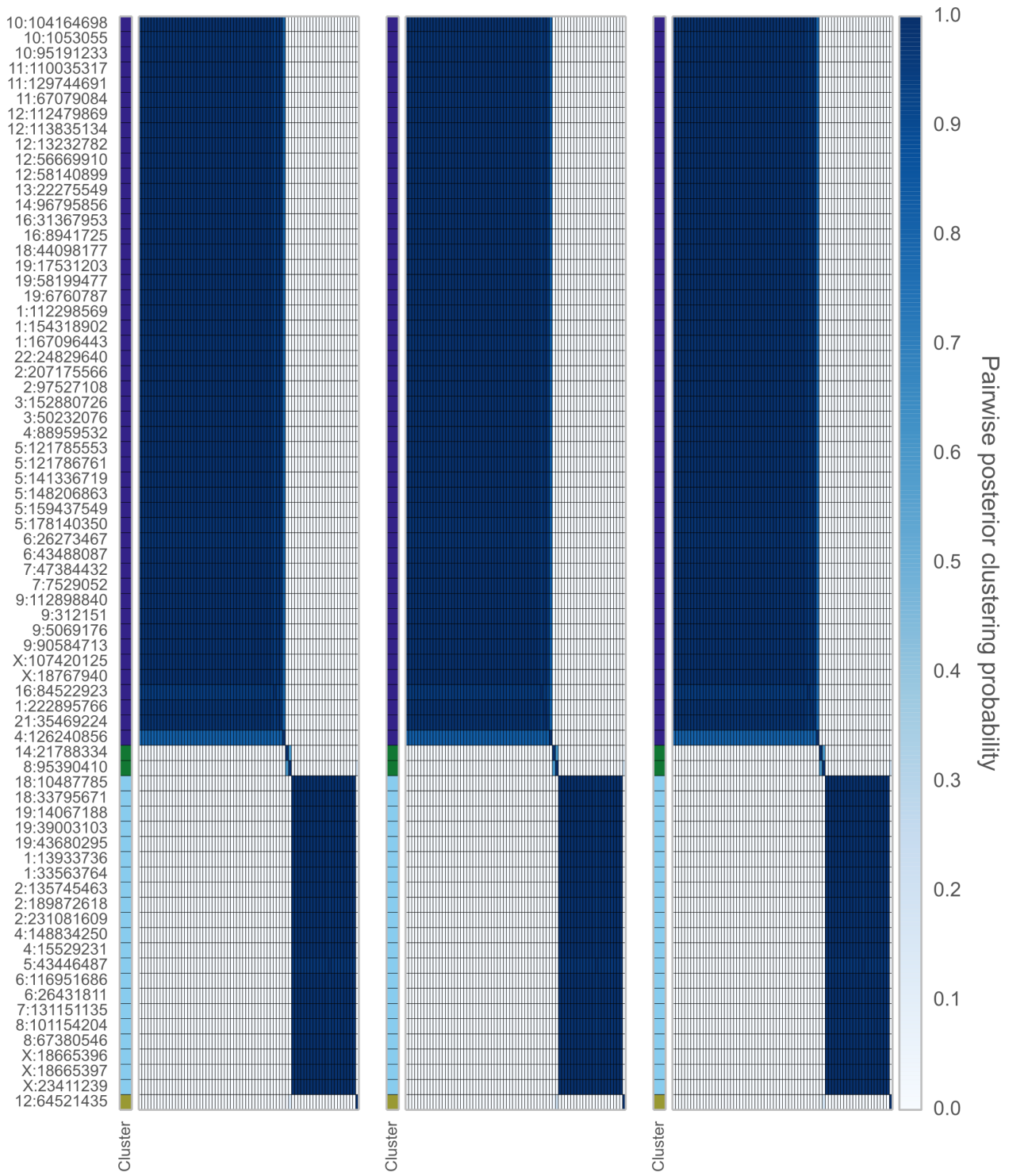
Supplementary Figure 4: Workflow for PyClone analysis | The sample is first assayed using whole genome shotgun sequencing (WGSS) or exome capture sequencing to identify putative mutations. Copy number information, which is used to inform the PyClone priors, can be derived from either sequence or array data. Putative mutations are subjected to targeted deep sequencing using either custom capture array or targeted PCR amplification. The input for the PyClone model is the allelic abundance measurements for the validated mutations from the targeted deep sequencing experiment and the prior information elicited from the copy number profiling. Additionally an estimate of tumour content derived from analysis of the array data, sequencing data, or from pathologist estimates can be supplied.



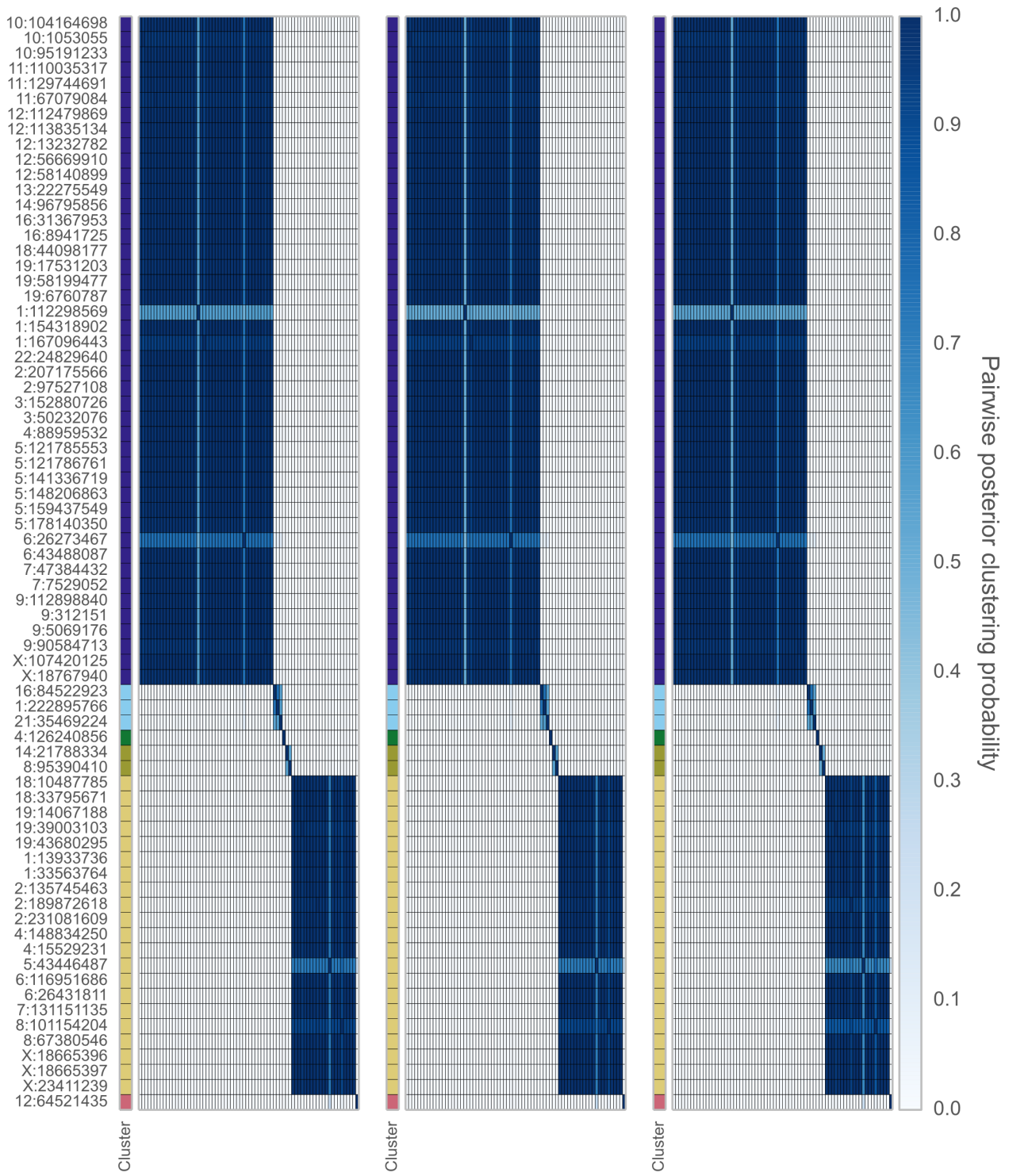
Supplementary Figure 5: PyClone results using differing copy number predictions | Predicted cellular prevalence and clustering estimates from HGSOC : case 1 using **(a)** ASCAT, **(c)** OncoSNP, **(e)** PICNIC; case 2 using **(b)** ASCAT, **(d)** OncoSNP, **(f)** PICNIC to inform PyClone using the BeBin-PCN model. Error bars indicate the mean standard deviation of MCMC cellular prevalences estimates for mutations in a cluster. n indicates the number of mutations assigned to a cluster.



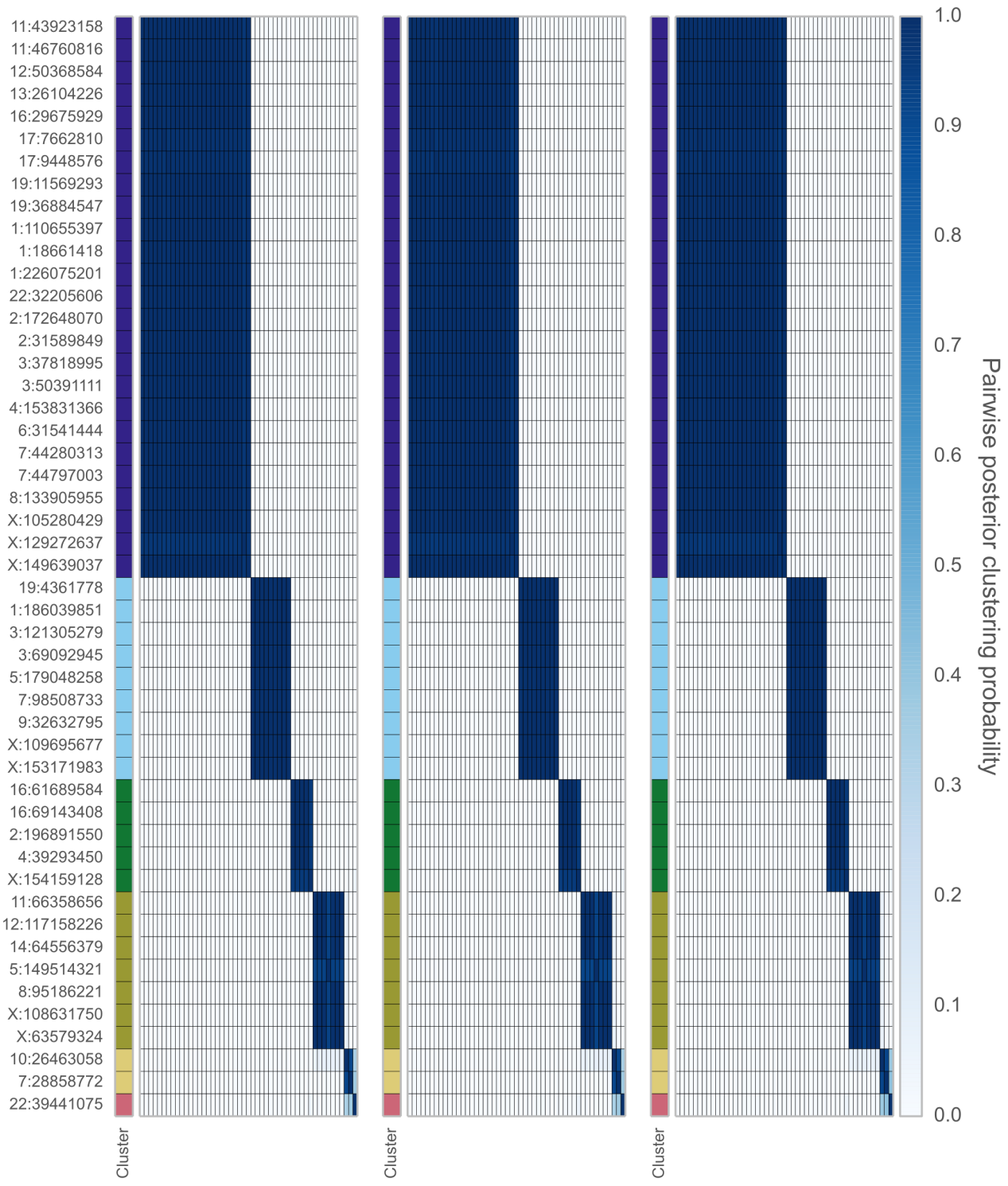
Supplementary Figure 6: HGSOC case 1 ASCAT | Posterior similarity matrices for high grade serous ovarian cancer case 1 using ASCAT for copy number prediction. Three MCMC runs from random starts are shown.



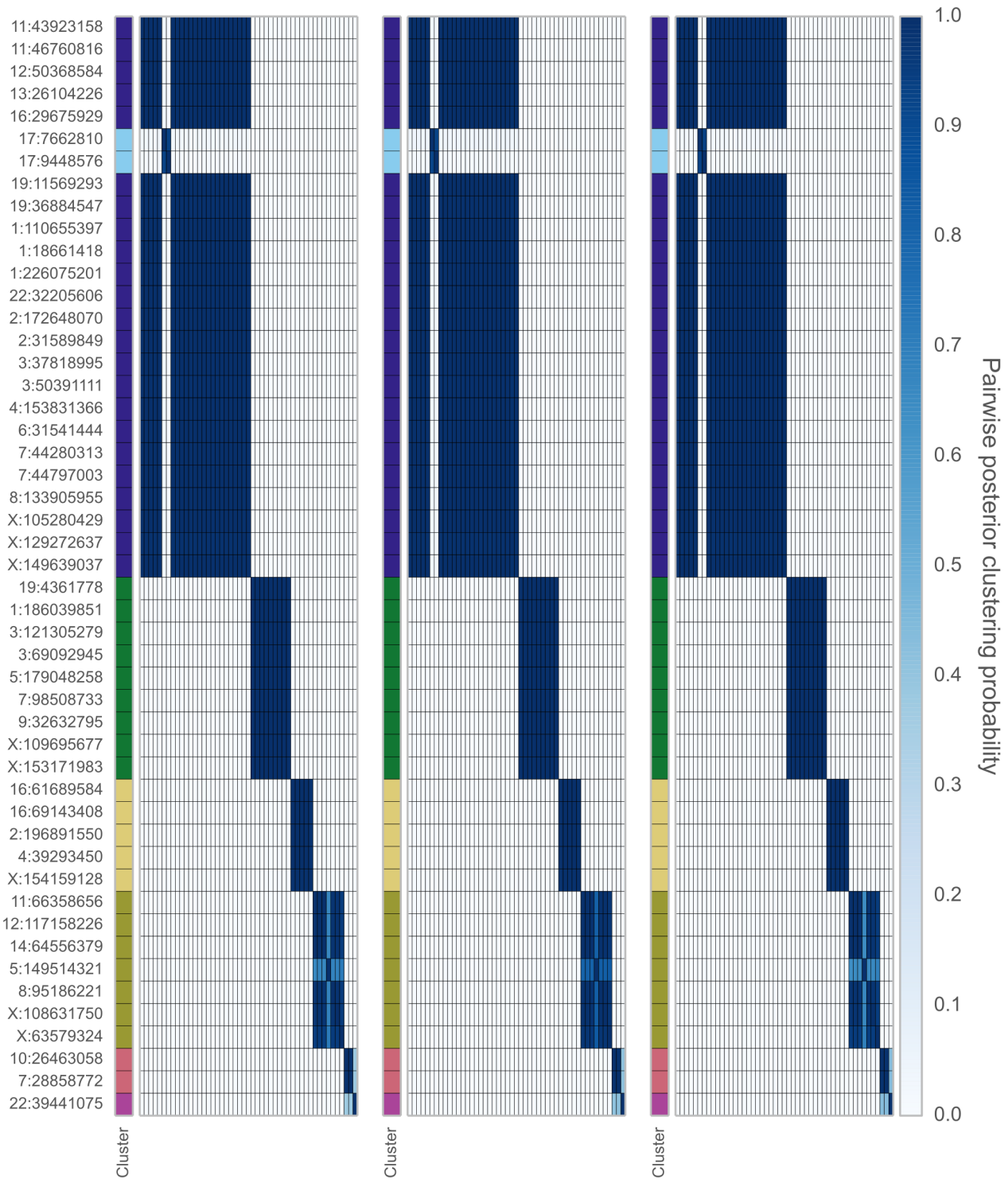
Supplementary Figure 7: HGSOC case 1 OncoSNP | Posterior similarity matrices for high grade serous ovarian cancer case 1 using OncoSNP for copy number prediction. Three MCMC runs from random starts are shown.



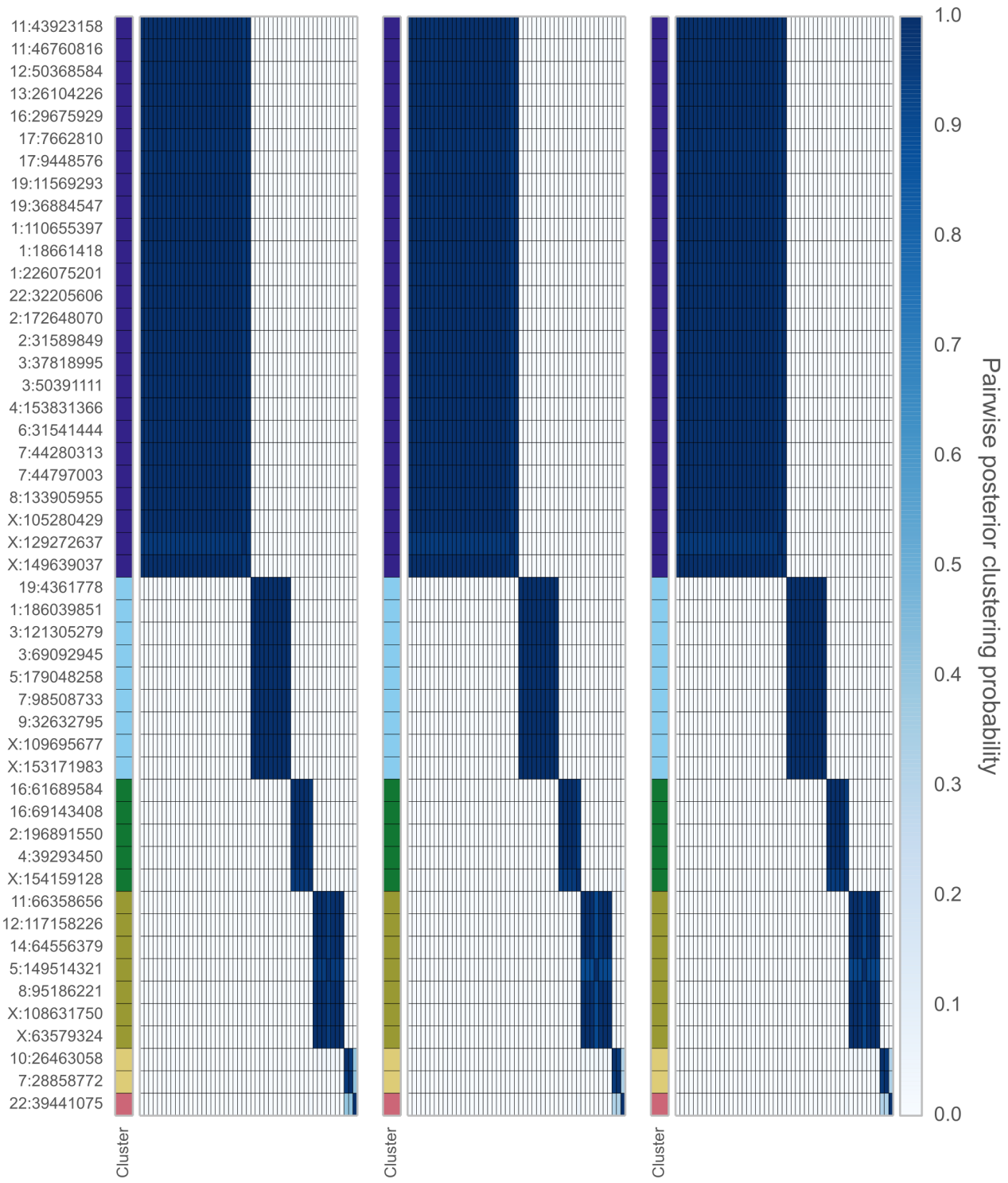
Supplementary Figure 8: HGSOC case 1 PICNIC | Posterior similarity matrices for high grade serous ovarian cancer case 1 using PICNIC for copy number prediction. Three MCMC runs from random starts are shown.



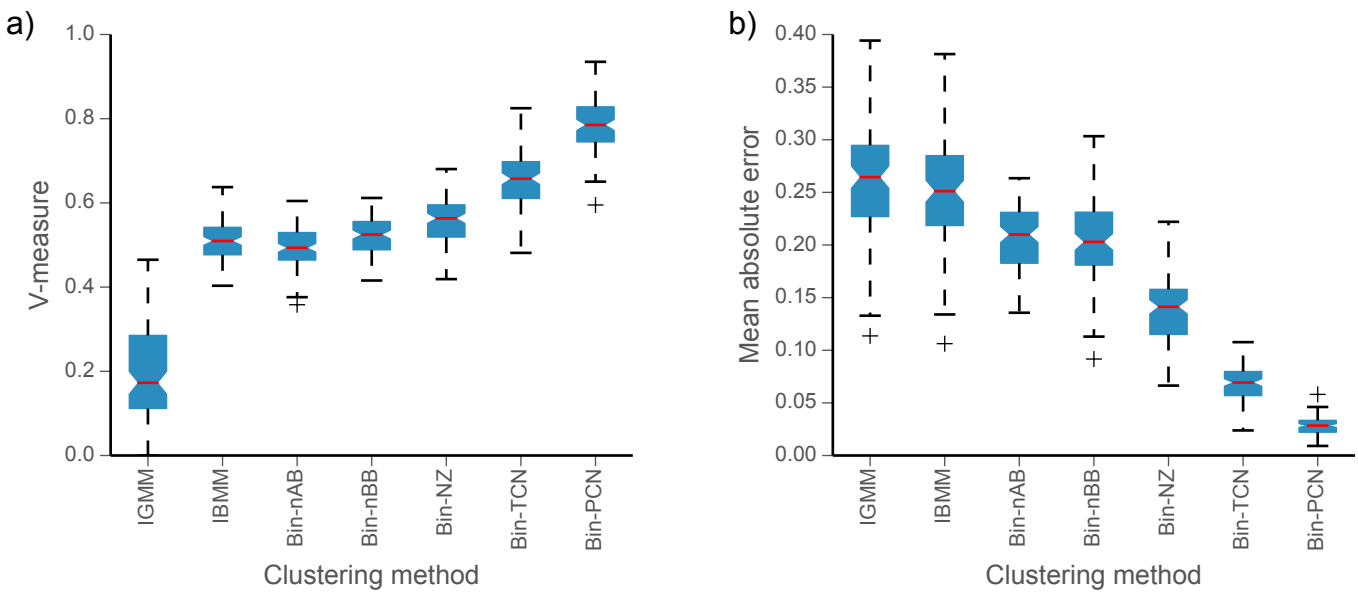
Supplementary Figure 9: HGSOC case 2 ASCAT | Posterior similarity matrices for high grade serous ovarian cancer case 2 using ASCAT for copy number prediction. Three MCMC runs from random starts are shown.



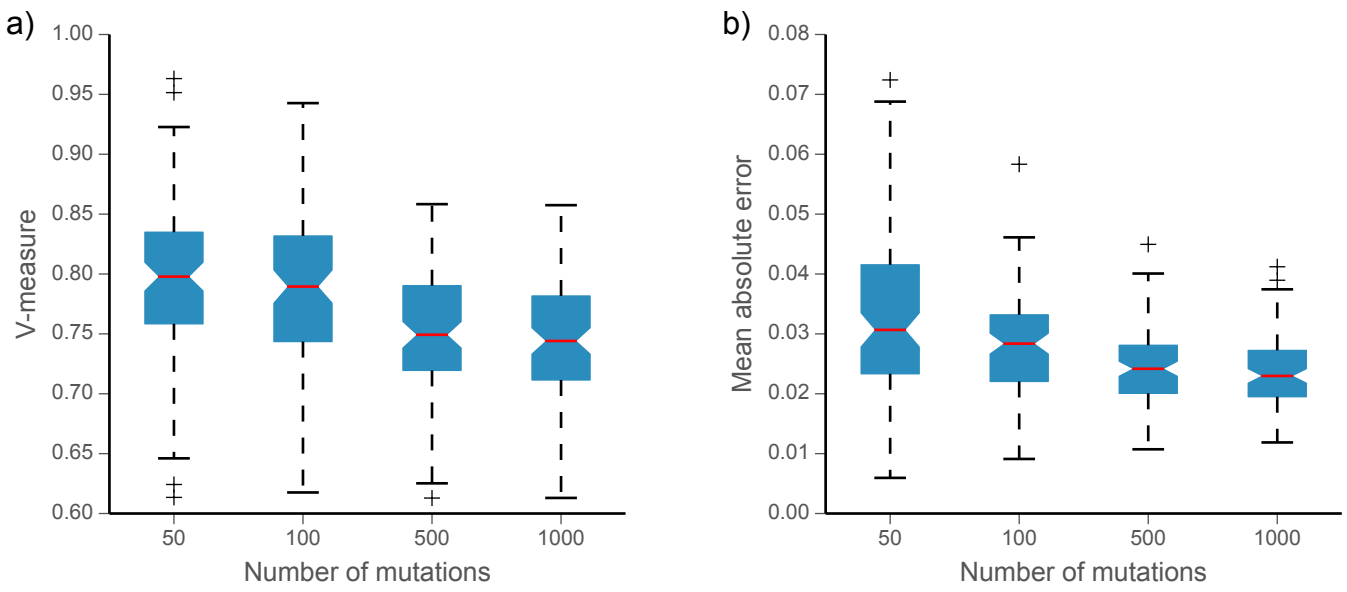
Supplementary Figure 10: HGSOC case 2 OncoSNP | Posterior similarity matrices for high grade serous ovarian cancer case 2 using OncoSNP for copy number prediction. Three MCMC runs from random starts are shown.



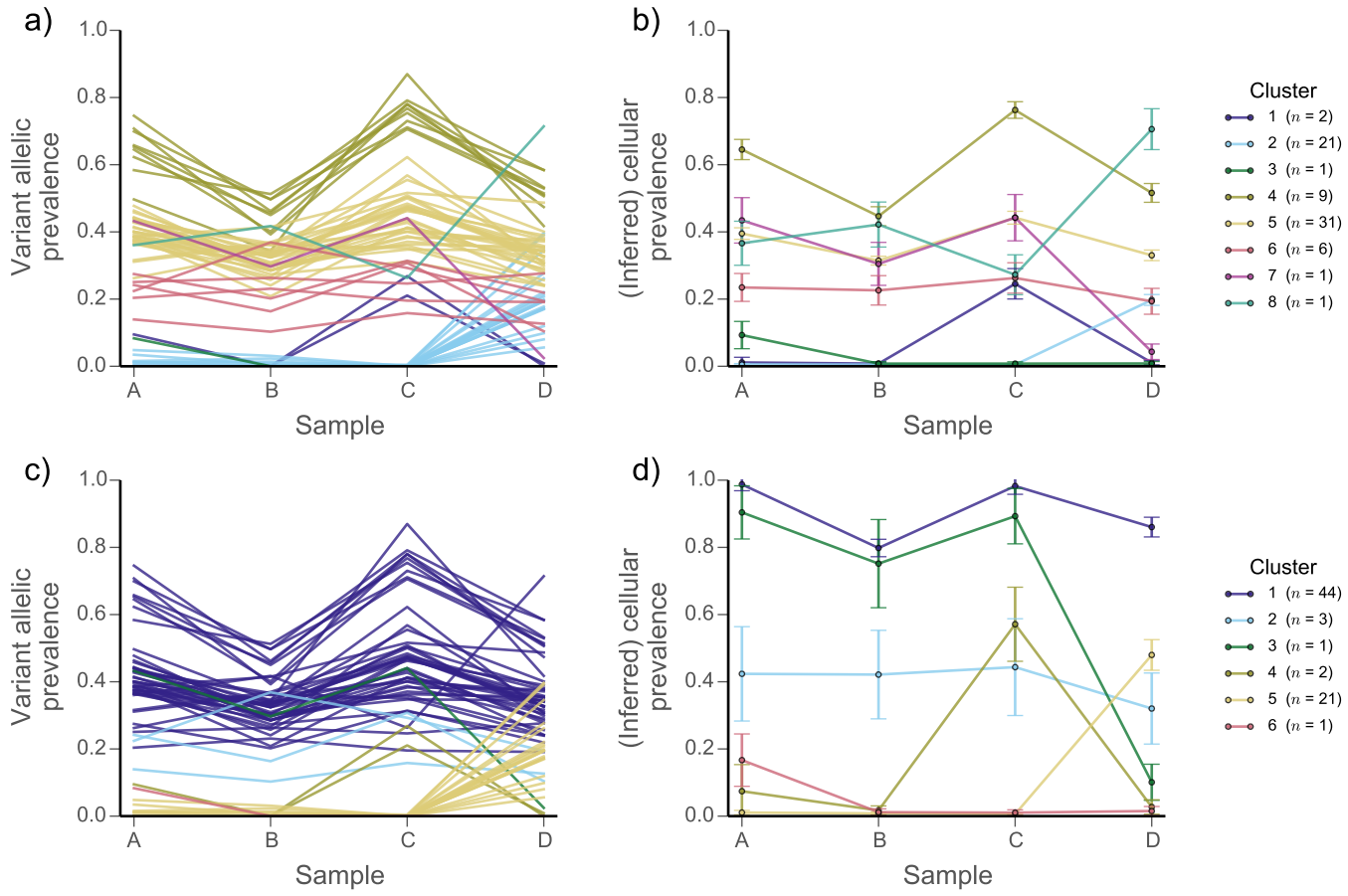
Supplementary Figure 11: HGSOC case 2 PICNIC | Posterior similarity matrices for high grade serous ovarian cancer case 2 using PICNIC for copy number prediction. Three MCMC runs from random starts are shown.



Supplementary Figure 12: Synthetic data method comparison | (**a**) Clustering performance and (**b**) estimated cellular prevalence accuracy for different methods applied to 100 synthetic data. (**a**) The accuracy of the inferred clusters is measured using the V-measure metric (y-axis). (**b**) The accuracy of inferred cellular prevalence is measured by computing the difference between the mean posterior value inferred from MCMC sampling and true value (see **Online Methods**). Whiskers indicate 1.5 the interquartile range, the red bars indicate the median, and boxes represent the interquartile range.



Supplementary Figure 13: Synthetic performance varying number of mutations | (**a**) Clustering performance and (**b**) estimated cellular prevalence accuracy for the PyClone BeBin-PCN model as function of the number of mutations. (**a**) The accuracy of the inferred clusters is measured using the V-measure metric (y-axis). (**b**) The accuracy of inferred cellular prevalence is measured by computing the difference between the mean posterior value inferred from MCMC sampling and true value (see **Online Methods**). Whiskers indicate 1.5 the interquartile range, the red bars indicate the median, and boxes represent the interquartile range.



Supplementary Figure 14: HGSOC case 1 | Joint analysis of multiple samples from high grade serous ovarian cancer (HGSOC) case 1. The variant allelic prevalence for each mutation color coded by predicted cluster using the (**a**) IBBMM and (**c**) PyClone with BeBin-PCN model to jointly analyse the four samples. The inferred cellular prevalence for each cluster using the (**b**) IBBMM and (**d**) BeBin-PCN methods. As in **Fig. 1** the cellular prevalence of the cluster is the mean value of the cellular prevalence of mutations in the cluster. Error bars indicate the mean standard deviation of MCMC cellular prevalences estimates for mutations in a cluster. The number of mutations n in each cluster is shown in the legend in parentheses.