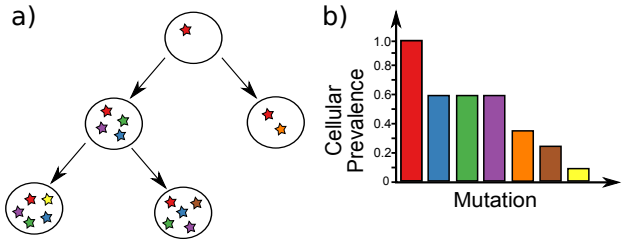
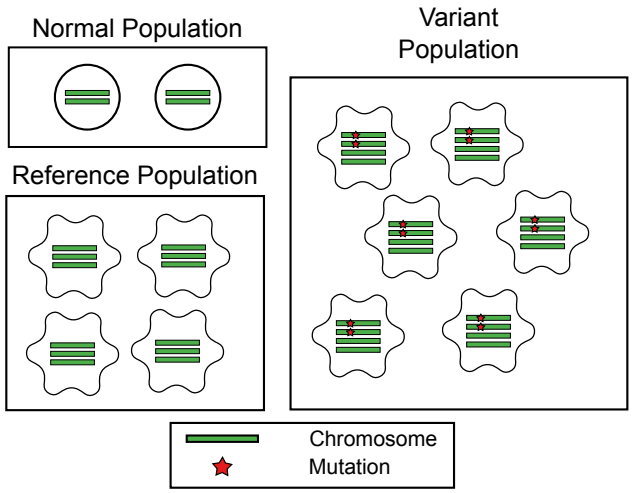
Supplementary Figures for Roth et al., PyClone: Statistical inference of clonal population structure in cancer



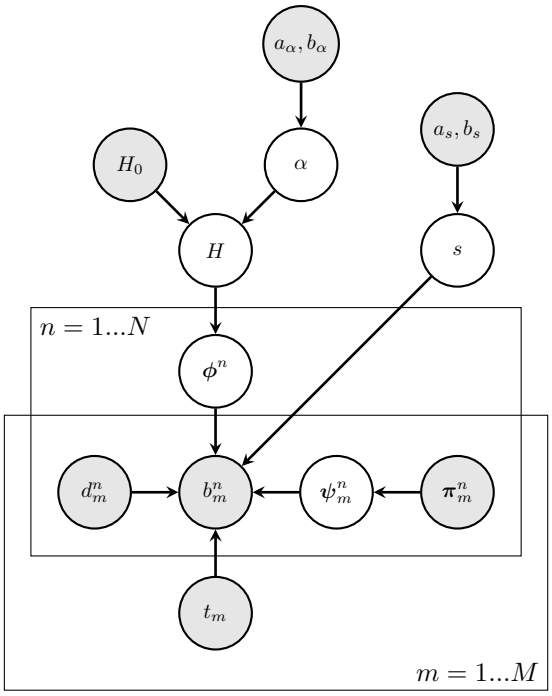
Supplementary Figure 1: Clonal evolution model | (**a**) A hypothetical phylogenetic tree generated by clonal expansion via the accumulation of mutations (stars). Unlike traditional phylogenetic trees internal nodes (clones) in the tree may contribute to the observed data, not just the leaf nodes. (**b**) 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.

Nature Methods: doi:10.1038/nmeth.2883



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 $\frac{6 \cdot 4 \cdot \frac{2}{4}}{2 \cdot 2 + 4 \cdot 3 + 6 \cdot 4} = 0.3$.

Nature Methods: doi:10.1038/nmeth.2883



$$
\begin{aligned}
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\
H_0 &= \text{Uniform}([0,1]^M) \\
H|\alpha, H_0 &\sim \text{DP}(\alpha, H_0) \\
\boldsymbol{\phi}^n|H &\sim H \\
\boldsymbol{\psi}_m^n|\boldsymbol{\pi}_m^n &\sim \text{Categorical}(\boldsymbol{\pi}_m^n) \\
\boldsymbol{\psi}_m^n &= (g_{m,\text{N}}^n, g_{m,\text{R}}^n, g_{m,\text{V}}^n)
\end{aligned}
$$

either

$$
b_m^n|d_m^n, \boldsymbol{\psi}_m^n, \phi_m^n, t_m \sim \text{Binomial}(d_m^n, \xi(\boldsymbol{\psi}_m^n, \phi_m^n, t_m))
$$

or

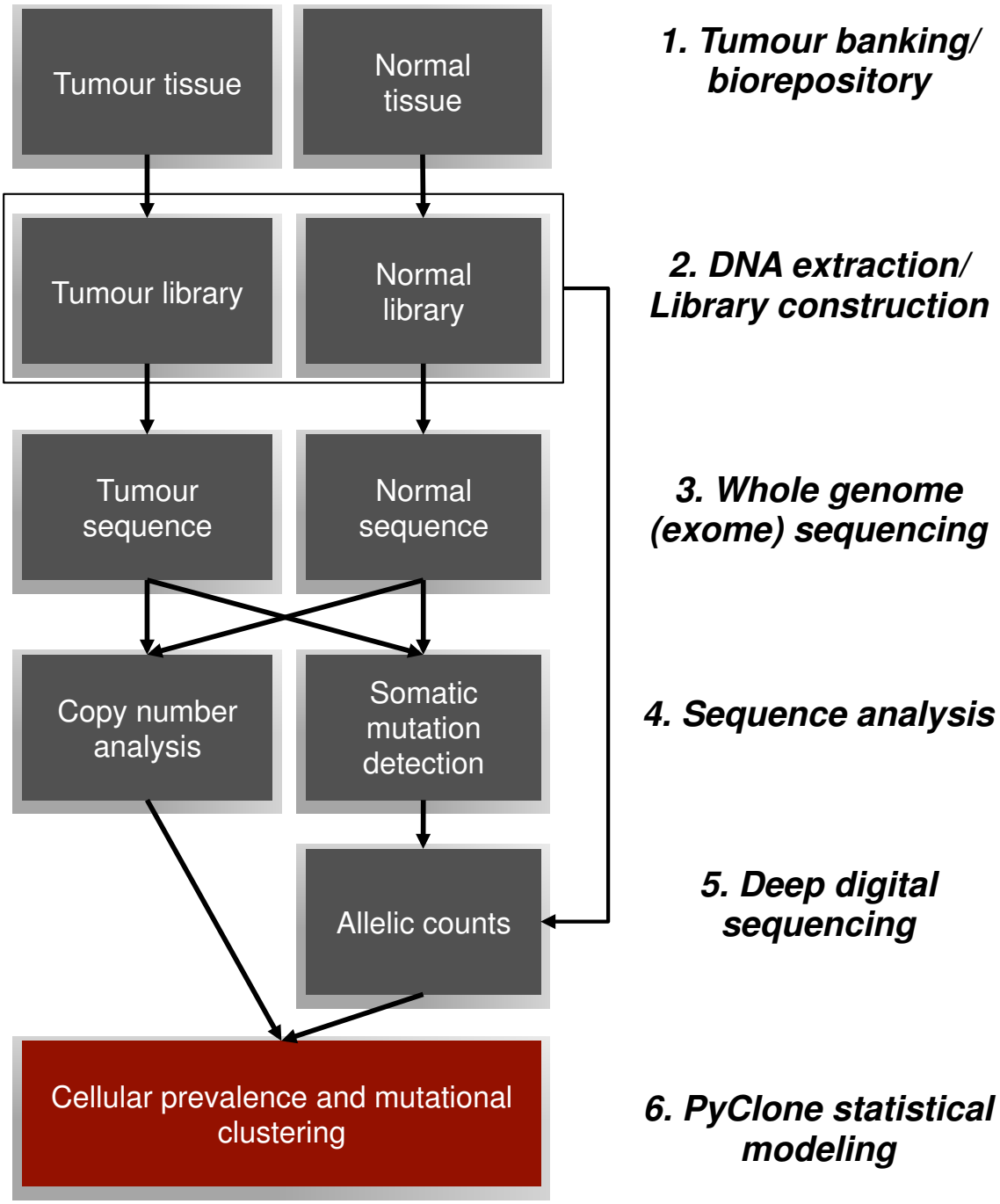
$$
\begin{aligned}
s|a, b &\sim \text{Gamma}(a_s, b_s) \\
b_m^n|d_m^n, \boldsymbol{\psi}_m^n, \phi_m^n, t_m, s &\sim \text{BetaBinomial}(d_m^n, \xi(\boldsymbol{\psi}_m^n, \phi_m^n, t_m), s)
\end{aligned}
$$

where

$$
\begin{aligned}
\xi(\boldsymbol{\psi}, \phi, t) &= \tfrac{(1-t)c(g_\text{N})}{Z}\mu(g_\text{N}) + \tfrac{t(1-\phi)c(g_\text{R})}{Z}\mu(g_\text{R}) + \\
&\quad \tfrac{t\phi c(g_\text{V})}{Z}\mu(g_\text{V}) \\
Z &= (1-t)c(g_\text{N}) + t(1-\phi)c(g_\text{R}) + \\
&\quad t\phi c(g_\text{V})
\end{aligned}
$$

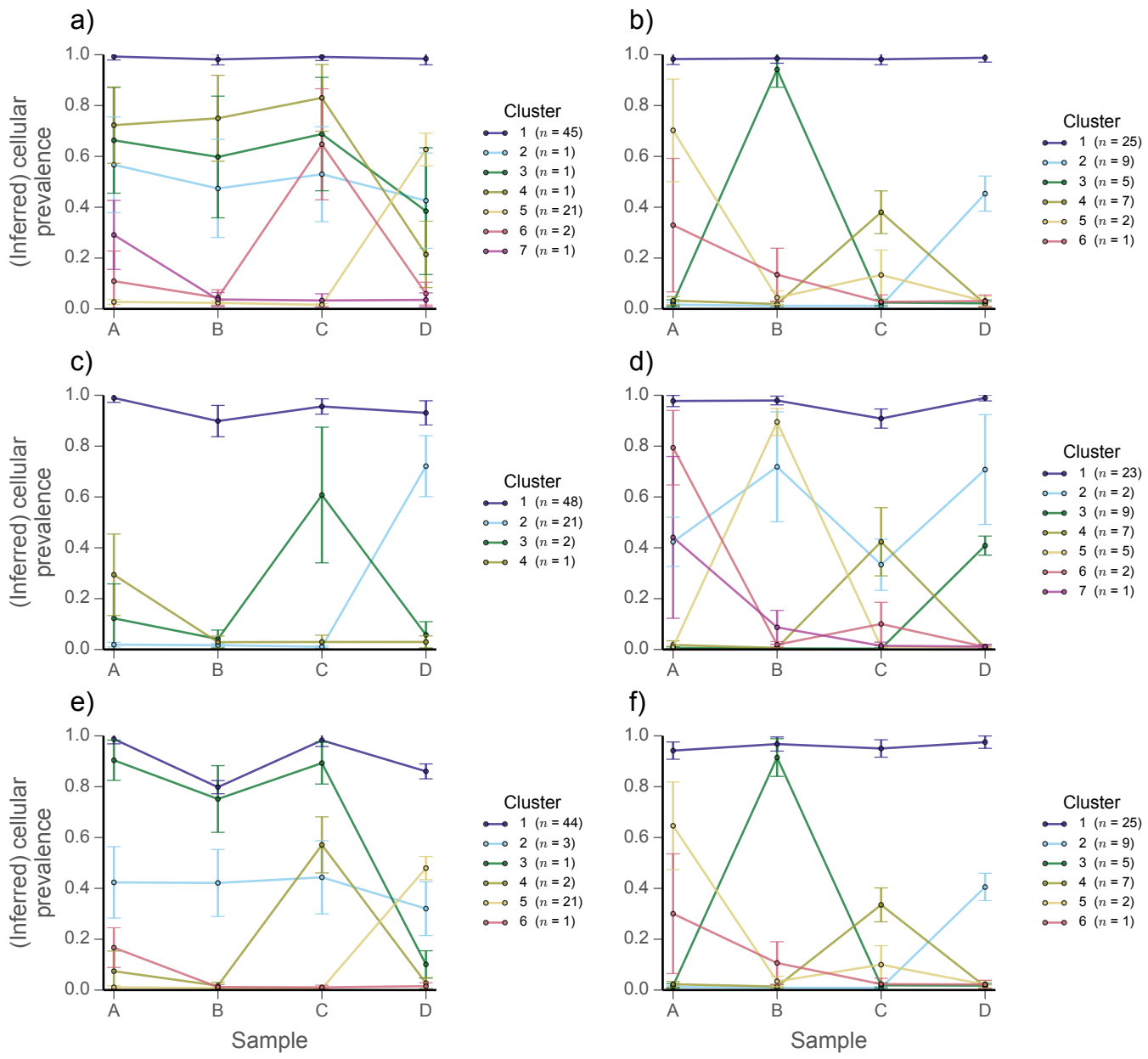
Supplementary Figure 3: Probabilistic graphical representation of the PyClone model | The model assumes the observed count data for the n^{th} mutation is dependent on the cellular prevalence of the mutation as well as the state of the normal, reference and variant populations. The cellular prevalence 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**).

Nature Methods: doi:10.1038/nmeth.2883



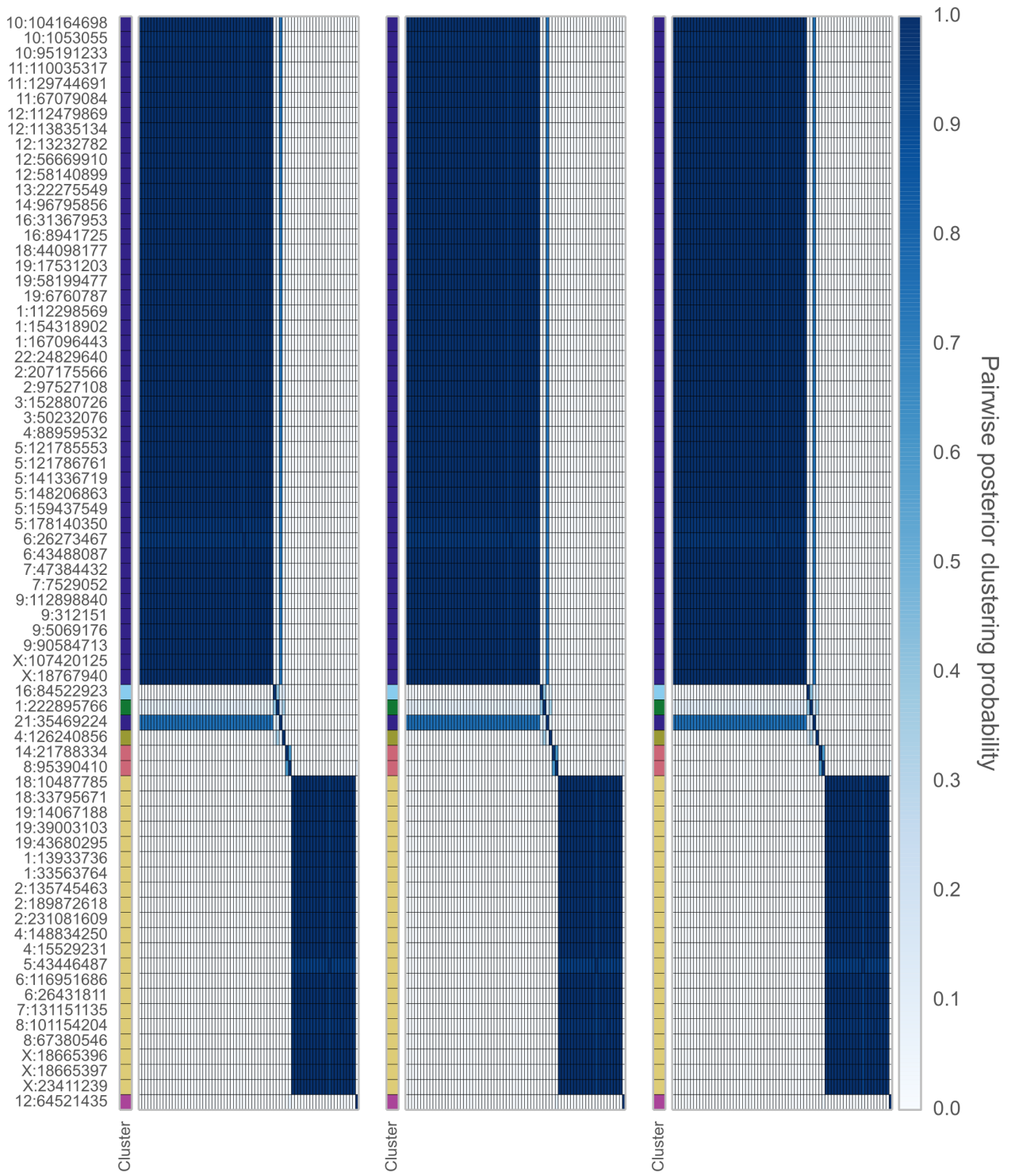
Supplementary Figure 4: Workflow for PyClone analysis | The sample is first assayed using whole genome shotgun sequencing (WGSS) or exome capture sequencing to identify putative mutations. Copy number information, which is used to inform the PyClone priors, can be derived from either sequence or array 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.

Nature Methods: doi:10.1038/nmeth.2883



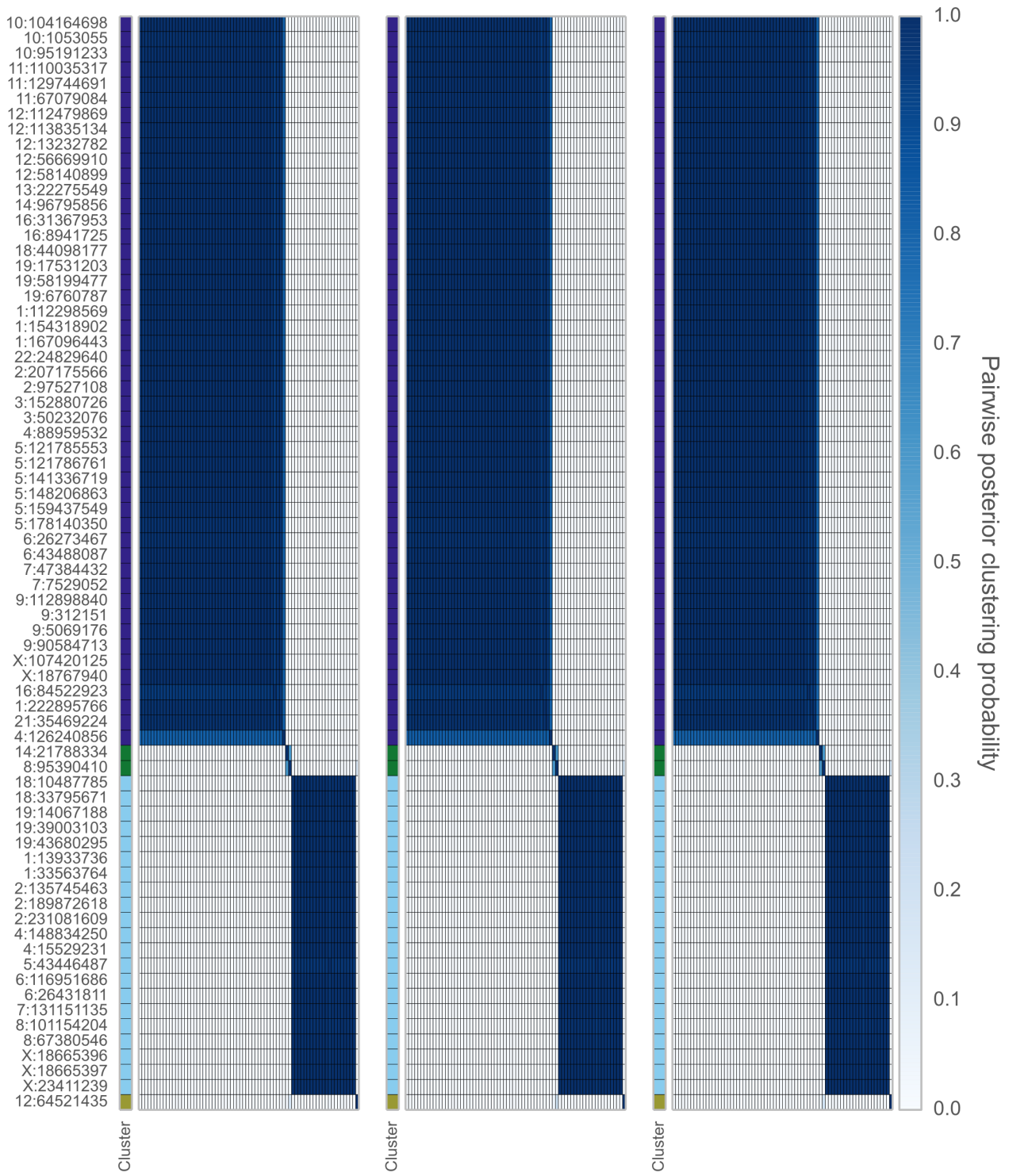
Supplementary Figure 5: PyClone results using differing copy number predictions | Predicted cellular prevalence and clustering estimates from HGSOC : case 1 using **(a)** ASCAT, **(c)** OncoSNP, **(e)** PICNIC; case 2 using **(b)** ASCAT, **(d)** OncoSNP, **(f)** PICNIC to inform PyClone using the BeBin-PCN 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.

Nature Methods: doi:10.1038/nmeth.2883



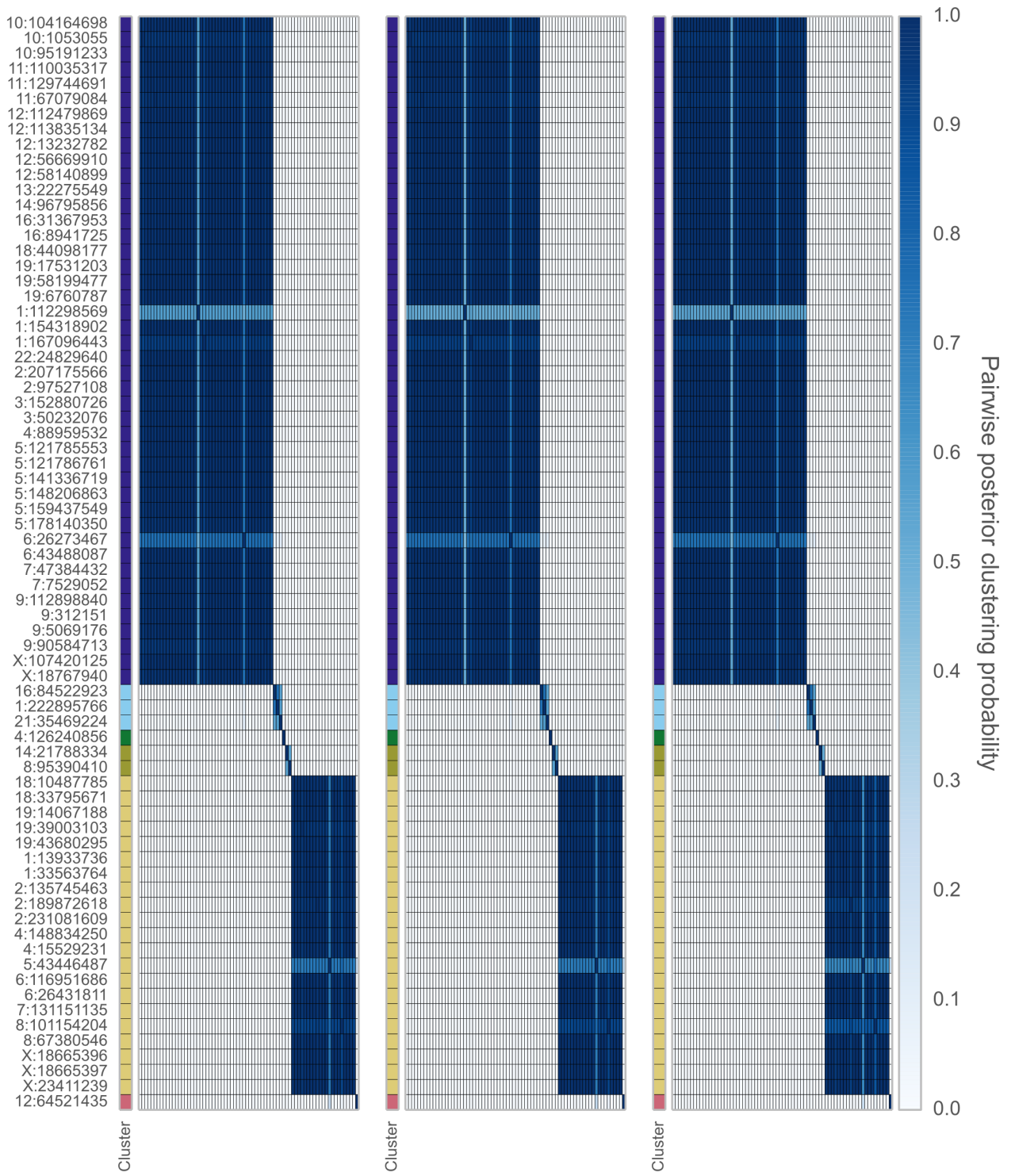
Supplementary Figure 6: HGSOC case 1 ASCAT | Posterior similarity matrices for high grade serous ovarian cancer case 1 using ASCAT for copy number prediction. Three MCMC runs from random starts are shown.

Nature Methods: doi:10.1038/nmeth.2883



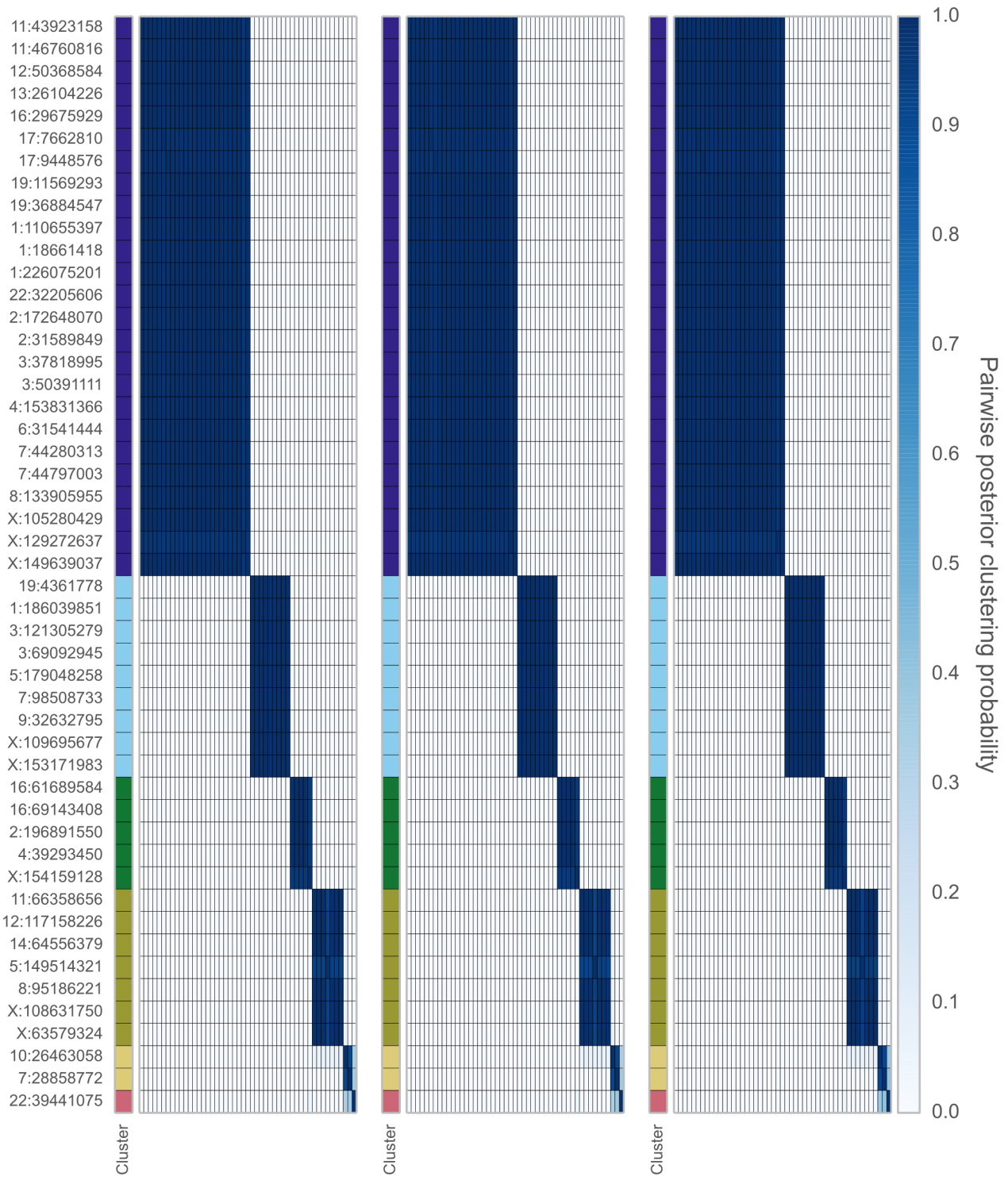
Supplementary Figure 7: HGSOC case 1 OncoSNP | Posterior similarity matrices for high grade serous ovarian cancer case 1 using OncoSNP for copy number prediction. Three MCMC runs from random starts are shown.

Nature Methods: doi:10.1038/nmeth.2883



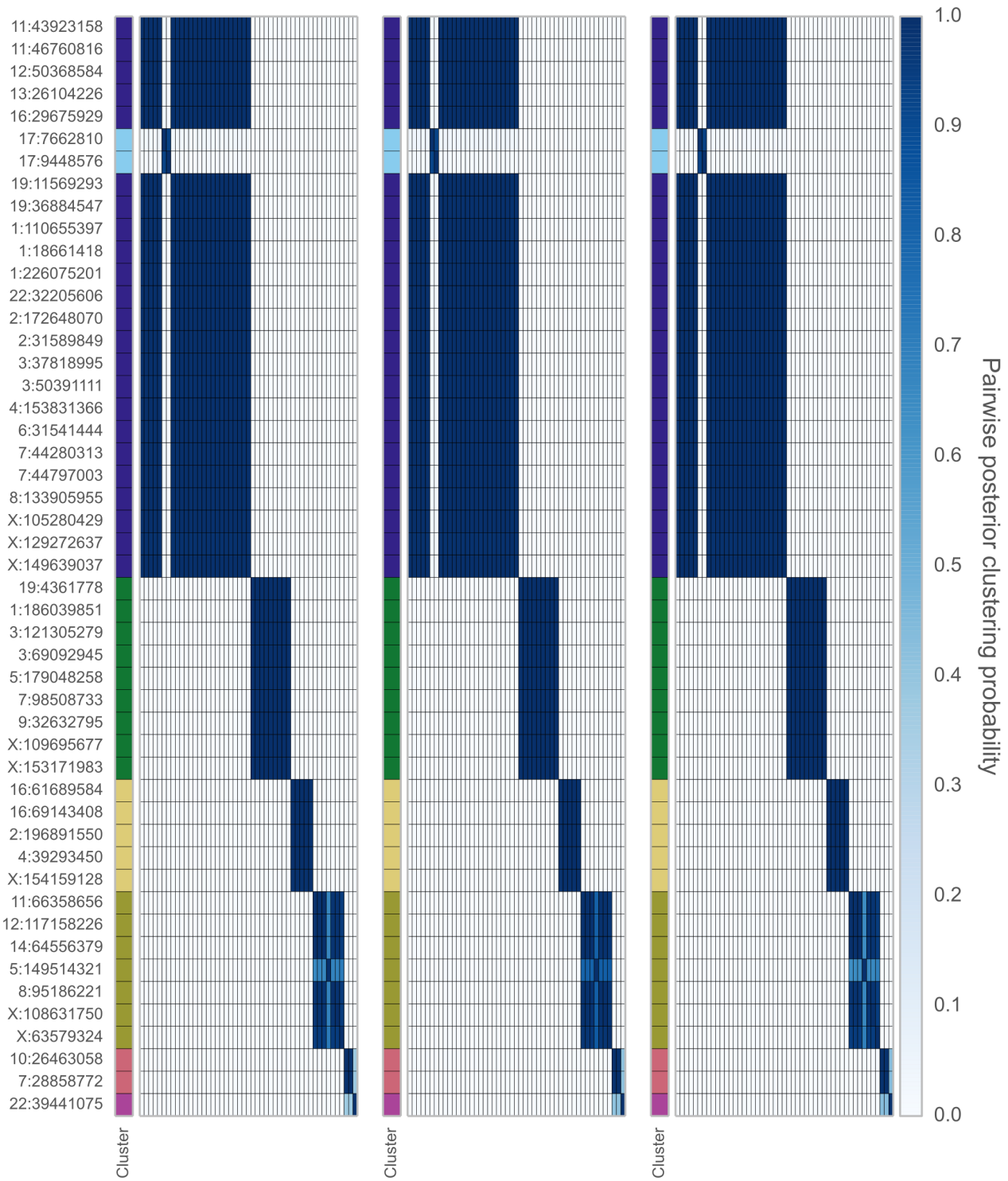
Supplementary Figure 8: HGSOC case 1 PICNIC | Posterior similarity matrices for high grade serous ovarian cancer case 1 using PICNIC for copy number prediction. Three MCMC runs from random starts are shown.

Nature Methods: doi:10.1038/nmeth.2883



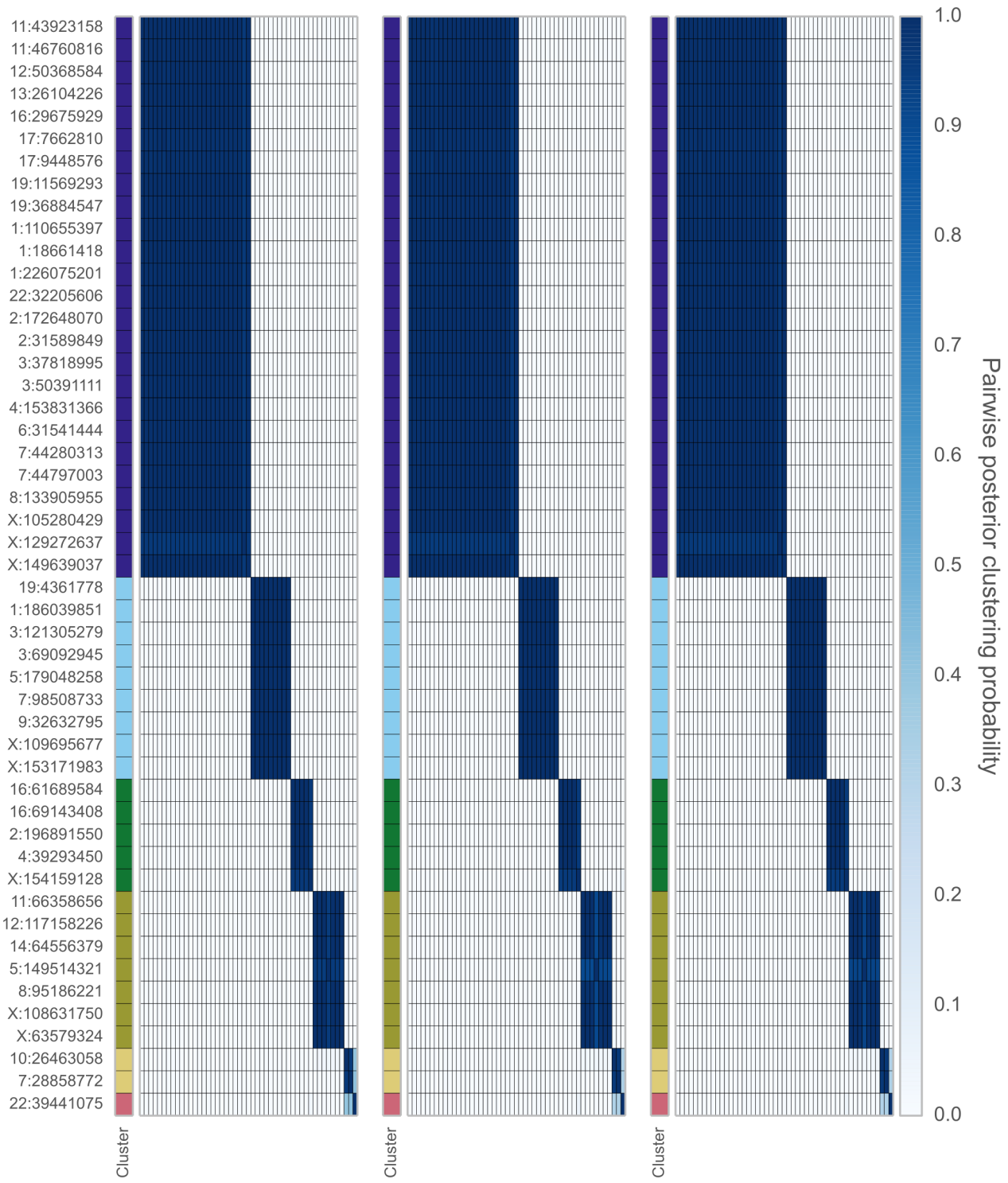
Supplementary Figure 9: HGSOC case 2 ASCAT | Posterior similarity matrices for high grade serous ovarian cancer case 2 using ASCAT for copy number prediction. Three MCMC runs from random starts are shown.

Nature Methods: doi:10.1038/nmeth.2883



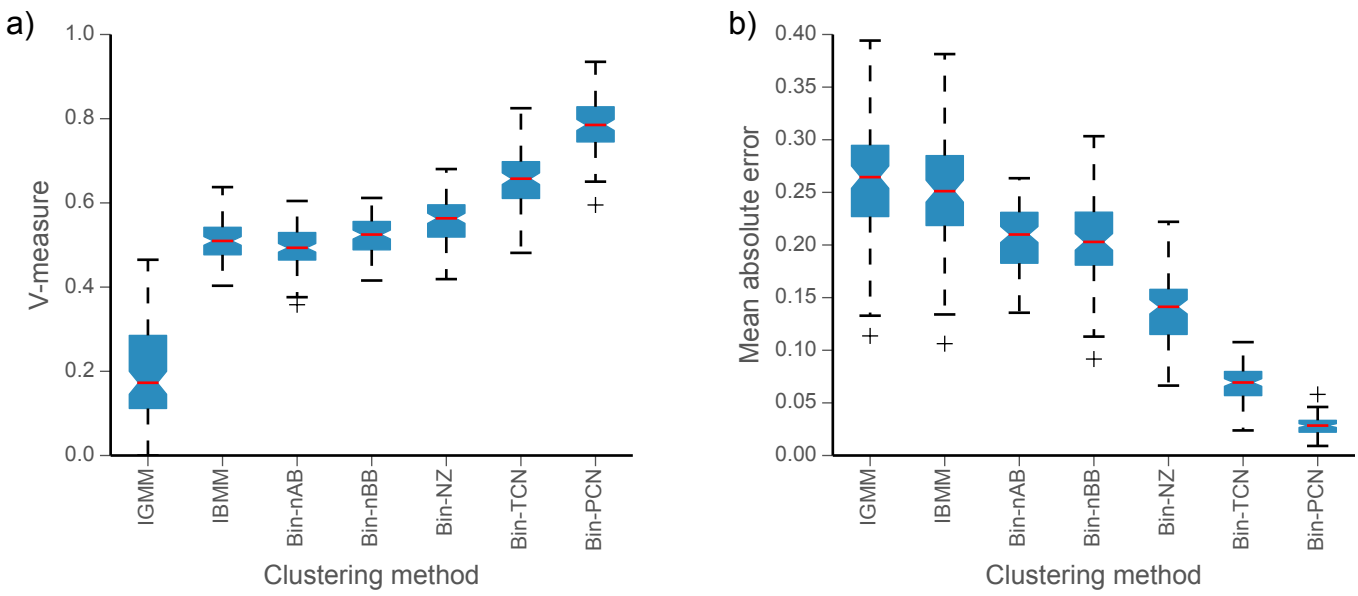
Supplementary Figure 10: HGSOC case 2 OncoSNP | Posterior similarity matrices for high grade serous ovarian cancer case 2 using OncoSNP for copy number prediction. Three MCMC runs from random starts are shown.

Nature Methods: doi:10.1038/nmeth.2883



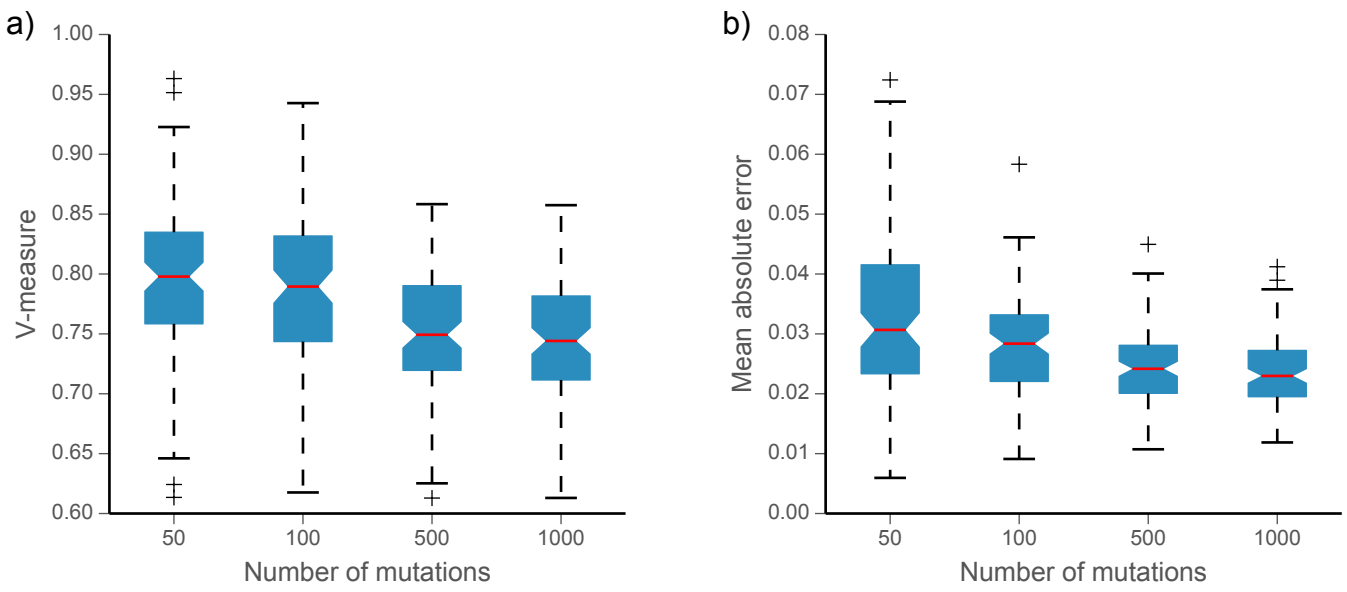
Supplementary Figure 11: HGSOC case 2 PICNIC | Posterior similarity matrices for high grade serous ovarian cancer case 2 using PICNIC for copy number prediction. Three MCMC runs from random starts are shown.

Nature Methods: doi:10.1038/nmeth.2883



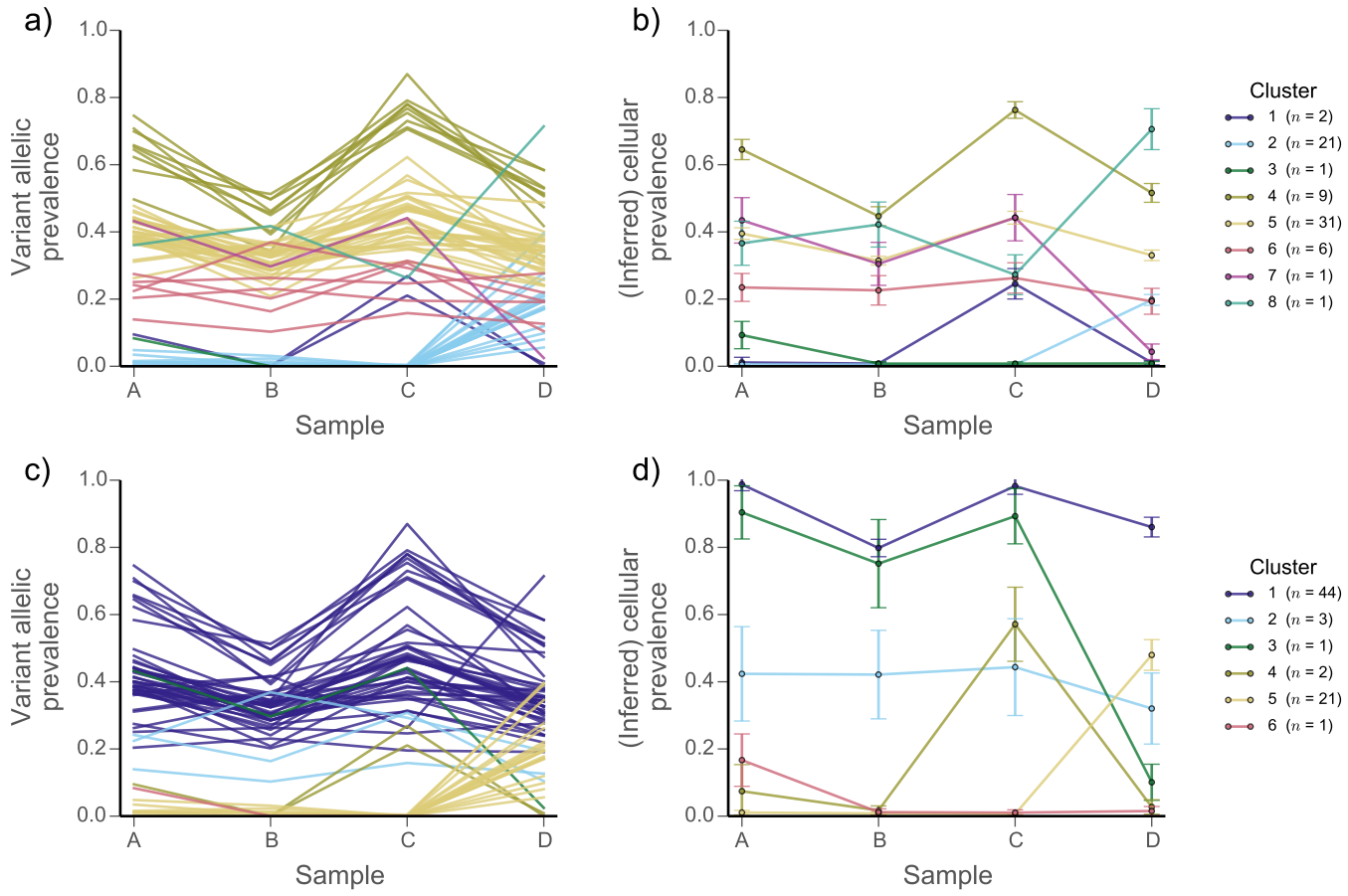
Supplementary Figure 12: Synthetic data method comparison | (**a**) Clustering performance and (**b**) estimated cellular prevalence accuracy for different methods applied to 100 synthetic data. (**a**) The accuracy of the inferred clusters is measured using the V-measure metric (y-axis). (**b**) The 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.

Nature Methods: doi:10.1038/nmeth.2883



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.

Nature Methods: doi:10.1038/nmeth.2883



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.

Nature Methods: doi:10.1038/nmeth.2883

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

Supplementary Table 1: Allelic counts, IBBMM and PyClone PCN cellular prevalence estimates for mutations in high grade serous ovarian cancer case 2. Copy number predictions where inferred using PICNIC as described in the **Online Methods**. Cellular prevalences where computed by taking the mean of the post burnin trace for the cellular prevalences for the respective methods. The standard deviation of the cellular prevalence parameter estimated from the post burnin trace is also included. Cluster ids (last two columns) were predicted from the post burnin trace using the MPEAR clustering criteria as described in the **Online Methods** and **Online Note**. Mutation ids list gene name, chromosome and chromosome coordinate. All coordinates are in the hg19 coordinate system.

Supplementary Table 2: Allelic counts, IBBMM and PyClone PCN cellular prevalence estimates for mutations in high grade serous ovarian cancer case 1. Copy number predictions where inferred using PICNIC as described in the **Online Methods**. Cellular prevalences where computed by taking the mean of the post burnin trace for the cellular prevalences for the respective methods. The standard deviation of the cellular prevalence parameter estimated from the post burnin trace is also included. Cluster ids (last two columns) were predicted from the post burnin trace using the MPEAR clustering criteria as described in the **Online Methods** and **Online Note**. Mutation ids list gene name, chromosome and chromosome coordinate. All coordinates are in the hg19 coordinate system.

Nature Methods: doi:10.1038/nmeth.2883

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

1 Simulated data results

To systematically assess the performance of different modelling strategies for mutational clustering and cellular prevalence inference, we generated 100 synthetic datasets of 100 mutations with randomly assigned copy number and mutational genotypes, grouped into eight clusters in each run. We evaluated the performance of PyClone on these data using five different strategies for specifying the state priors plus two standard clustering models (IGMM, IBMM) outlined in **Online Methods**. Benchmarking was based on V-measure and mean error in estimating cellular prevalence (see **Online Methods**). Group distributions over accuracy were assessed using ANOVA tests with pair-wise TukeyHSD tests for two-group comparisons. Adjusted p values with $q<0.05$ was used as a criteria for statistical differences. Details of synthetic data generation, the PyClone model and its variants, and statistical analysis is provided in the **Supplemental Note**.

Clustering accuracy was highest in the PyClone PCN method with a V-measure of 0.78±0.06, followed by PyClone TCN (0.65±0.07) (**Supplementary Fig. 12a**). The PCN and TCN methods, which account for mutational genotype, significantly outperformed all other methods (**Supplementary Fig. 12a**). Accounting for copy number but ignoring mutational genotype, the NZ method (0.56±0.06) was worse than the PCN and TCN methods, but performed significantly better than the AB and BB methods that assume diploid states for the variant population (0.49±0.05 and 0.52±0.04, $q<0.05$, ANOVA and TukeyHSD). The IBMM (0.51±0.05) performed similarly to the PyClone diploid methods but was significantly better than IGMM (0.20±0.11).

Mean error on cellular prevalence estimates was also measured for each method. Similar to V-measure benchmarks, PyClone PCN was the most accurate (mean absolute error=0.03±0.01), significantly lower than all other methods (**Supplementary Fig. 12b**). TCN (0.07±0.02) and NZ (0.14±0.03) were more accurate than the AB and BB methods (0.21±0.03 and 0.20±0.04). All PyClone methods were significantly more accurate than the IBMM and IGMM methods (0.25±0.05 and 0.26±0.05). The likely reason is that the PyClone methods account for tumour content (set at 0.75 for all simulations). Of the two methods which consider mutational genotype, PCN significantly outperformed TCN. Though not surprising since the PCN method provides more informative prior information, this result suggests that given reliable parental copy number information, the PCN strategy for setting genotype priors would improve inference.

Taken together, these results systematically demonstrate the theoretical basis for estimating mutational genotype by incorporating copy number and parental allele information. This confers increased accuracy in both clustering and cellular prevalence estimates. Furthermore these results validate the basis for avoiding the use of Gaussian distributions when clustering deep digital

Nature Methods: doi:10.1038/nmeth.2883

sequencing data.

To asses how performance varies with the number of mutations, we simulated datasets with 50, 100, 500 and 1,000 mutations. 100 replicates were generated for each number of mutations using the same procedure as above. The datasets were analysed using PyClone with Bin-PCN model. We find that clustering performance deteriorates (**Supplementary Fig. 13a**) as the number of mutations analysed increases. In contrast the estimated cellular prevalence of the mutations improves as we increase the number of mutations analysed (**Supplementary Fig. 13b**). We believe that clustering performance deteriorates with more mutations because the problem of clustering larger datasets is intrinsically more difficult. The increasing accuracy of cellular prevalence estimates would suggest that mutations from the same cluster are being placed together. As we increase the number of mutations, the mean number of predicted clusters increases from 8.64 ± 2.02 with 50 mutations to 28.85 ± 13.77 with 1,000 mutations. Taken together this suggest PyClone tends to over cluster as the number of mutations increases, but the cellular prevalence of each cluster will be accurate. This differs from the over clustering of genotype naive methods, which results in mutational clusters being assigned inaccurate cellular prevalence estimates.

2 High grade serous ovarian cancer results

Case 1 mutations clustered by IBBMM (**Supplementary Fig. 14a**,**b**, **Supplementary Table 2**) resulted in mutations with the highest allelic prevalences (Cluster 4, $n = 9$ mutations) grouped together. These mutations showed a similar prevalence pattern to mutations in Cluster 5 ($n = 31$ mutations) across samples. We suggest these mutations belong to the same clone, with Cluster 4 mutations predominantly homozygous and Cluster 5 mutations predominantly heterozygous (**Supplementary Table 2**). Eight of nine mutations in IBBMM Cluster 4 fall into regions of heterozygous deletion in all four samples. By contrast, 28 of 31 mutations in Cluster 5 are in diploid heterozygous regions (**Supplementary Table 2**) in all four samples. Therefore, the difference in cellular prevalence estimates in IBBMM between Cluster 4 and 5 can be explained by the copy number of the loci spanning the mutations impacting the mutational genotype. PyClone instead groups the mutations corresponding to IBBMM Cluster 4 and Cluster 5 into one group of $n = 44$ mutations (**Supplementary Fig. 14c**,**d** - Cluster 1), with representation of both the heterozygous and homozygous loci at cellular prevalences of near 1.0.

PyClone cluster 1 in both case 1 ($n = 44$ mutations) and case 2 ($n = 25$ mutations) (**Supplementary Fig. 14d**, **Fig. 2d**) likely represent mutations comprising the ancestral clone in these tumours' aetiology. In contrast, clusters with cellular prevalences lower than 1.0 indicate putative descendant clones. We emphasize that IBBMM splits the ancestral clone into at least two clusters in both case 1 and case 2, with evidence from the copy number analysis (**Supplementary Table 1** and **2**) attributing the split based on heterozygous or homozygous mutational genotype. PyClone modelling of mutational genotypes coupled with simultaneous inference across multiple samples therefore provides a more robust approach to ascertaining clonal populations with

Nature Methods: doi:10.1038/nmeth.2883

dramatic implications for how cellular prevalence estimates of mutations are interpreted in reconstruction of evolutionary histories.

3 Stability of predictions using different copy number predictions

In order to quantify the impact that different copy number predictions have on the PyClone prediction, we analysed two HGSOC cases using three different copy number methods to predict tumour content and inform the PyClone mutational genotype priors (**Supplementary Note**). We ran three random starts of the MCMC analysis for each method to ensure the variability we observed in output was due to different inputs rather than stochastic convergence issues.

Slightly different clusterings and cellular prevalence estimates are observed (**Supplementary Fig. 5**). The difference in clusterings usually result in a single mutation being removed from a large cluster and forming a singleton cluster. The posterior pairwise clustering probabilities largely converge to the same values (**Supplementary Fig. 6 - 11**). These results suggest PyClone is somewhat sensitive to the copy number predictions used to inform the prior. However, this analysis was based on using the PCN strategy for setting mutational genotype priors. The PyClone model allows for flexible priors to be used and we are investigating approaches to combine copy number predictions from multiple methods to improve robustness. In the interim a reasonable practice maybe to remove positions which lie in regions with very different copy number predictions between methods.

Nature Methods: doi:10.1038/nmeth.2883

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

1 Incorporating external sources of copy number information

The accuracy of genotype prior information will greatly influence the results of the PyClone analysis. Vague prior information will limit the ability to accurately cluster mutations and infer cellular frequency since a large space of equally likely explanations for the observed data must be considered. This problem is not specific to our method, and represents a major challenge of the problem in general.

A principled approach to limit the search space is to restrict the variant population to have genotypes compatible with copy number information predicted on an orthogonal platform. This approach still leaves open the problems of determining which genotypes compatible with the predicted copy number are most likely, and what the copy number of the reference population is. In some cases, biological information about the mutation may help solve the first problem. For example, *EZH2* mutations in lymphoid cancers are known to be functional only if a wildtype mutation remains[1], thus we would weight the heterozygous mutations more heavily in this case. By contrast, *TP53* mutations in high grade serous ovarian cancers are nearly always homozygous[2]. As for the copy number of the reference population, it is ultimately determined by whether the mutation event precedes or follows the copy number event in the region. In some cancers, large scale alterations of the copy number architecture are believed to be early events with subsequent evolution occurring via the accumulation of point mutations[3]. In this case it would be more likely that the reference population has the same total copy number as the variant population. As we show, homologous copy number information can also be of help if we are willing to make some assumptions about the mutational process.

One area of future research is how the predictions from multiple sources could be used to inform PyClone. We saw that different software tools will produce different copy number and tumour content predictions, which will influence the PyClone results. Given that PyClone has great deal of flexibility in specifying priors, it should be possible to use multiple tools to inform the priors. A principled method to weigh the various predictions is required for this approach, ideally one that accounts for the strengths and weaknesses of different prediction tools. A closely related issue is how we can use sub-clonal copy number predictions to inform the PyClone priors.

2 Limitations

The PyClone model does not cluster cells by mutational composition, which is the traditional view of clonal heterogeneity and tumour phylogenies. Instead it clusters mutations which appear

Nature Methods: doi:10.1038/nmeth.2883

at similar cellular frequencies. In simple cases the clustering derived may be correct, however if multiple sub-clones exist at similar cellular frequencies the model will falsely cluster the associated mutations together. This is one reason we have focused on targeted deep sequencing when applying PyClone, as the chance of making this error will decrease with higher sequencing coverage. Joint analysis of multiple samples can also help address this issue as clones appearing at similar cellular prevalences in one sample may appear at different prevalences in another. Ultimately this will be resolved with the maturation of single-cell sequencing techniques that scale to allow for the reconstruction of individual cell genotypes.

A key assumption of our model is that all cells within each populations have the same genotype. This assumption is likely false in some cases as cancer cells can undergo copy number alterations and LOH events before and after the acquisition of mutations[4]. The error induced by this assumption will depend heavily on how variable the genotype of cells are at the locus of interest. In solid tumours the tissue samples prepared for deep sequencing experiments are taken from a relatively small spatial area. In this case the error from assuming the same genotype within populations may be relatively small. For liquid tumours this assumption might be much worse as cells are highly mobile. It is possible to relax this assumption but the resulting model would lead to an intractable inference problem.

We note that our assumptions are predicated on the notion of a 'perfect' and 'persistent' phylogeny whereby mutations accrue over time and persist in clones and their descendants and mutations occur exactly once. We recognize that numerous well described phenomena such as revertant mutations, recurrent mutations (occurring independently along different branches of the tree) or deletions of loci harbouring mutations would violate this assumption. Accounting for such possibilities would ultimately lead to an intractable inference problem owing to unidentifiable explanations for the observed data.

The posterior densities for the cellular frequencies can often exhibit a degree of uncertainty making interpretation difficult. Uncertainty can arise due to imprecise prior information on genotype of the mutations and to the depth of sequencing. As prior information about genotype improves and depth of sequencing increases we would expect multi-modality to become less prominent. Though uncertainty makes interpretation difficult, it is a realistic representation of the confounding factors in this problem. One of the major contributions of this work is highlighting that such uncertainty exists unless strong assumptions about the genotype of the mutations are made.

Another approach to reducing the uncertainty is to use multiple samples from patients separated in time (primary vs. relapse) or space through regional or anatomic sampling. We have shown that our model can accommodate multi-sample data in a principled way by using hierarchical Bayesian modelling where statistical strength is borrowed across datasets, dramatically decreasing uncertainty while increasing accuracy.

Nature Methods: doi:10.1038/nmeth.2883

3 Alternative Applications

PyClone is specifically designed for the problem of inferring the cellular prevalence of single nucleotide mutations in deeply sequenced tumour samples. However, the model is quite generic in the sense that it only assumes the sequenced sample is a heterogeneous mixture of cells which fall into three distinct sub-populations. Provided that data which accurately reflects the abundance of an alteration can be generated, we could likely apply PyClone to infer the prevalence of indels, genomic rearrangement breakpoints or methylation marks in malignant tissues. By varying the model parameters and genotype priors it would also be relatively straightforward to apply the PyClone model to infer the prevalence of somatic alterations in non-malignant tissue.

1. Yap, D.B. *et al. Blood* **117**, 2451–9 (2011).

2. Ahmed, A.A. *et al. J Pathol* **221**, 49–56 (2010).

3. Carter, S.L. *et al. Nat Biotechnol* **30**, 413–21 (2012).

4. Navin, N. *et al. Nature* **472**, 90–4 (2011).

Nature Methods: doi:10.1038/nmeth.2883

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

1 The PyClone model description

PyClone is a software package which provides tools for performing Dirichlet Process (DP) clustering of mutations. It implements several standard clustering algorithms such as the infinite Binomial mixture model (IBMM) as well as several models which account for the genotype of a mutation. In the following description we will use PyClone to refer to the collection of genotype aware clustering models.

PyClone is a hierarchical Bayes statistical model (**Supplementary Fig. 3**). Input data consists of allelic counts from a set of N deeply sequenced mutations for a given sample. Prior information is elicited from copy number estimates obtained from either genotyping arrays or whole genome sequencing. For most available tools, these estimates will represent the average copy number of a locus if the copy number of the population is heterogeneous at the locus. An optional estimate of tumour content, derived from computational methods or pathologists estimates, may also be used. The model outputs a posterior density for each mutation's cellular prevalence and a matrix containing the probability any two mutations occur in the same cluster. The model assigns two mutations to the same cluster if they occur at the same cellular prevalence in the sample(s). This is a necessary but not a sufficient condition for mutations to be present in the same clonal population. To obtain a flat clustering of the mutations from the matrix of pairwise probabilities we construct a dendrogram and find the cut point that optimises the MPEAR criterion[1] which is discussed in the **Online Methods**. **Supplementary Fig. 4** shows a typical experimental workflow used to produce the allelic count data and tumour content estimates which are inputs for a PyClone analysis. The same workflow also shows how copy number information is generated which is then used to elicit priors for possible mutational genotypes which are also required as input for a PyClone analysis. For more details on how copy number information is used to elicit priors see section 4.

The model divides the sample into three sub-populations with respect to mutation $n \in \{1, \ldots, N\}$: the normal (non-malignant) population, the reference and the variant cancer cell populations (**Supplementary Fig. 2**). The reference population consists of all cancer cells which are wildtype for the n^{th} mutation. The variant population consists of all cancer cells with at least one variant allele of the n^{th} mutation. To simplify inference of the model parameters we assume that within each sub-population, the mutational genotype at site n is the same for all cells in that sub-population. But importantly, we allow the mutational genotypes to vary across populations. We introduce a collection of categorical random variables $g^n_{\mathrm{N}}, g^n_{\mathrm{R}}, g^n_{\mathrm{V}}$, each taking values in $\mathcal{G} = \{-, A, B, AA, AB, BB, AAA, AAB, \ldots\}$, denoting the genotype of the normal, reference and variant populations with respect to mutation n. For example, the genotype AAB refers to the geno-

Nature Methods: doi:10.1038/nmeth.2883

type with two reference alleles and one variant allele. The symbol − denotes the genotype with no alleles, in other words a homozygous deletion of the locus. The vector $\boldsymbol{\psi^n} = (g^n_{\mathrm{N}}, g^n_{\mathrm{R}}, g^n_{\mathrm{V}}) \in \mathcal{G}^3$ represents the state for the n^{th} mutation, while $\boldsymbol{\pi}^n$ is a vector of prior probabilities over all possible states, $\boldsymbol{\psi^n}$, of the n^{th} mutation.

The fraction of cancer cells (tumour content) is t, with fraction of normal cells $1 - t$. We fix t prior to inference, assuming estimates from orthogonal assays such as WGSS, micro-arrays or histopathology. We define the fraction of cancer cells from the variant population ϕ^n, and correspondingly $1 - \phi^n$ as the fraction of cancer cells from the reference population. With this formulation the *cellular prevalence*, the fraction of cancer cells harbouring a mutation, is given by ϕ^n. The *cellular prevalence* is a fundamental quantity for examining population dynamics across multiple samples as it is not affected by the tumour content of a sample. As a result the *cellular prevalence* allows us to track changes in the population structure across samples (with regards to SNVs) without the confounding effect of contaminating normal cells.

For a genotype $g \in \mathcal{G}$, $c(g) : \mathcal{G} \mapsto \mathbb{N}$ returns the copy number of the genotype, for example $c(AAB) = 3$. We define $b(g) : \mathcal{G} \mapsto \mathbb{N}$, which returns the number of variant alleles in the genotype, for example $b(AAB) = 1$. If $b(g) \neq 0$ and $b(g) \neq c(g)$ we assume that the probability of sampling a variant allele from a cell with genotype g is given by $\mu(g) = \frac{b(g)}{c(g)}$. In the case where $b(g) = 0$ we assume $\mu(g) = \epsilon$, where ϵ is the probability of erroneously observing a B allele when the true allele sequenced was A. We make this modification to allow for the effect of sequencing error. Similarly we define $\mu(g) = 1 - \epsilon$ when $b(g) = c(g)$. The definition of $\mu(g)$ assumes the probability of a sequencing error is independent of the sequenced allele. Because of this assumption we do not account for sequencing errors for other genotypes since these errors should cancel on average and the expected fraction of B alleles should stay the same as the error free case.

We assume that the sequenced reads are independently sampled from an infinite pool of DNA fragments. Thus the probability of sampling a read covering a given locus from a sub-population is proportional to the prevalence of the sub-population and the copy number of the locus in cells in from that population. Therefore, the probability of sampling a read containing the variant allele covering a mutation with state $\boldsymbol{\psi} = (g_{\mathrm{N}}, g_{\mathrm{R}}, g_{\mathrm{V}})$ and cellular prevalence ϕ is given by:

$$\begin{aligned} \xi(\boldsymbol{\psi}, \phi, t) &= \frac{(1-t)c(g_{\mathrm{N}})}{Z}\mu(g_{\mathrm{N}}) + \frac{t(1-\phi)c(g_{\mathrm{R}})}{Z}\mu(g_{\mathrm{R}}) + \\ &\quad \frac{t\phi c(g_{\mathrm{V}})}{Z}\mu(g_{\mathrm{V}}) \\ Z &= (1-t)c(g_{\mathrm{N}}) + t(1-\phi)c(g_{\mathrm{R}}) + t\phi c(g_{\mathrm{V}}) \end{aligned}$$

We let b^n denote the number of reads observed with the B allele, with d^n total reads covering the locus, where the n^{th} mutation has occurred. It is straightforward to show that b^n follows a

Nature Methods: doi:10.1038/nmeth.2883

Binomial distribution with parameters d^n and $\xi(\boldsymbol{\psi}_n, \phi_n, t)$. This assertion follows from the fact that the sum of n Bernoulli random variables with parameter p follows a Binomial distribution with parameters n, p [2].

The posterior distribution of the prevalences $\phi = (\phi^1, \ldots, \phi^N)$ is then given by

$$\begin{aligned}
&p(\phi|b^n, d^n, \boldsymbol{\pi}^n, t) \\
&\quad \propto p(\phi) \prod_{n=1}^{N} p(b^n|\phi^n, d^n, \boldsymbol{\pi}^n, t) \\
&\quad = p(\phi) \prod_{n=1}^{N} \sum_{\boldsymbol{\psi}^n \in \mathcal{G}^3} p(b^n|\phi^n, d^n, \boldsymbol{\psi}^n, t) p(\boldsymbol{\psi}^n|\boldsymbol{\pi}^n) \\
&\quad = p(\phi) \prod_{n=1}^{N} \sum_{\boldsymbol{\psi}^n \in \mathcal{G}^3} \text{Binomial}(b^n|d^n, \xi(\boldsymbol{\psi}^n, \phi^n, t)) \pi^n_{\boldsymbol{\psi}^n}.
\end{aligned}$$

In principle, the sum over $\boldsymbol{\psi}^n \in \mathcal{G}^3$ could be infinite. In practice we cannot enumerate an infinite set of states. Thus we must specify a finite set of states which will have non-zero prior probability which in turn truncates the sum.

Mutations from the same clonal population should appear at the same cellular prevalence. To account for this we specify a DP prior with base measure $H_0 \sim \text{Uniform}(0, 1)$ for $p(\phi)$ which allows mutations to share the same cellular prevalence [3]. If we were to directly use a continuous distribution such as Uniform(0, 1) as a prior for $p(\phi)$, the cellular prevalence of all mutations would be different with probability one. The DP prior converts this continuous distribution into a discrete distribution with an infinite number of point masses. Since the DP distribution is discrete, it gives a non-zero prior probability to mutations sharing the same cellular prevalence. Though each mutation samples its own value of ϕ^n from the DP, the fact that ϕ^n can be identical induces a clustering of the data.

Due to the presence of the DP prior, computing the exact posterior distribution is not tractable. We use an auxiliary variable sampling method [4] to perform Markov Chain Monte Carlo (MCMC) sampling from the posterior distribution. First, the sampler iterates over each mutation choosing a new value of ϕ^n from $p(\phi)$. The mutations may either choose a value of ϕ^n used by other mutations, effectively joining a cluster, or choose an unused value of ϕ^n, starting a new cluster. After this step the values of each cluster are resampled using a Metropolis-Hastings step with the base measure H_0 as the proposal distribution. The concentration parameter, α, in the DP is sampled using the method described in West *et al.*[5]. This method places a Gamma distribution on α which leads to simple a Gibbs resampling step. The Gamma distribution prior is parametrised in terms of the shape a and rate b parameters. The density for this prior is given by

Nature Methods: doi:10.1038/nmeth.2883

$$p(\alpha|a,b) = \frac{b^a \alpha^{a-1} \exp(-b\alpha)}{\Gamma(a)}$$

where $\Gamma(x)$ is the gamma function. The mean and variance of this distribution are given by

$$\begin{aligned} \mathbb{E}(\alpha) &= \frac{a}{b} \\ Var(\alpha) &= \frac{a}{b^2} \end{aligned}$$

We typically use values of $a = 1.0$ and $b = 10^{-3}$ so that the variance of the prior distribution on α is 10^6, which is extremely vague.

To initialise the sampler we assign all mutations to separate clusters. As a result the computational complexity of the first pass of the sampler is $O(N^2)$, where N is the number of mutations. Subsequent iterations have computational complexity $O(NK)$ where K is the number of active clusters.

2 Multiple samples modeling

Increasingly, common experimental designs acquire deep digital sequencing across spatial or temporal axes, examining shifts in prevalence as a marker of selection. As these measurements are not independent (derivative clones are related phylogenetically), we assume M samplings from the same cancer can share statistical strength to improve clustering performance. We substitute the univariate base measure in H_0 with a multivariate base measure; for concreteness we use the uniform distribution over $[0,1]^M$. The Dirichlet process then samples a discrete multivariate measure H over the clusters and each data point draws a vector of, $\boldsymbol{\phi}^n = (\phi_1^n \ldots \phi_M^n)$ from this measure. The likelihood under this model is given by

$$\begin{aligned} &p(\boldsymbol{\phi}|\boldsymbol{b^n}, \boldsymbol{d^n}, \boldsymbol{\pi}^n, \boldsymbol{t}) \propto \\ &p(\boldsymbol{\phi}) \prod_{m=1}^{M} \prod_{n=1}^{N} \sum_{\boldsymbol{\psi}_m^n \in \mathcal{G}^3} p(b_m^n | d_m^n, \phi_m^n, \boldsymbol{\psi}_m^n, t_m) p(\boldsymbol{\psi}_m^n | \boldsymbol{\pi}_m^n) \end{aligned}$$

Nature Methods: doi:10.1038/nmeth.2883

For each mutation n we assign different priors, $\boldsymbol{\pi}_m^n$, for each sample, allowing for the genotypes of the reference and variant populations to change between samples; for example if the samples came from a regional samples in a tumour mass, primary tumour and distant metastasis, or pre- and post- chemotherapy. We also introduce the vector $\boldsymbol{t} = (t_1, \ldots, t_M)$ which contains the tumour content of each sample. Using this approach the clustering of mutations is shared across all samples but the cellular prevalence of each mutation is still free to vary in the M samples. Thus the final output of the model will be a single posterior similarity matrix for all mutations and $N \times M$ posterior densities (one per mutation per sample) for the cellular prevalences of each mutation.

3 Addressing overdispersion

Next generation sequencing data are often overdispersed [6]. We implemented a version of the PyClone framework which replaces the Binomial distribution with a Beta-Binomial distribution, parametrised in terms of the mean and precision. The density is given by

$$p(b|d, m, s) = \binom{d}{b} \frac{B(b + sm, d - b + s(1 - m))}{B(sm, s(1 - m))}$$

where B is the Beta function. We set $m = \xi(\boldsymbol{\psi}^n, \phi^n, t)$ and to reduce the number of parameters which need to be estimated we share the same value s across all data points, and when applicable all samples.

4 Methods for eliciting PyClone priors over mutational genotypes

The genotype aware models implemented in the PyClone package requires that we specify prior $\pi_{\boldsymbol{\psi}}^n$ for the state of the sample at the n^{th} mutation. The state is defined by the normal, reference and variant genotypes and is denoted by the state vector $\boldsymbol{\psi^n} = (g_{\mathrm{N}}^n, g_{\mathrm{R}}^n, g_{\mathrm{V}}^n)$. A number of methods are available to profile parental (allele specific) and total copy number from high density genotyping arrays [7–9], or from whole genome sequencing data [10,11]. As segmental aneuploidies and loss of heterozygosity are accepted to be an essential part of the tumour genome landscape [12,13], it has become routine to assay the genome architecture in conjunction with mutational analysis. To explore the impact different prior assumptions have on performance, we consider a range of strategies for setting the prior probabilities over states. We denote the total copy number by $\overline{c}$, and the copy number of each homologous chromosome by $\overline{c_1}, \overline{c_2}$. In what follows we assume that correct copy number information is available for $\overline{c}$, $\overline{c_1}$, and $\overline{c_2}$.

We consider five strategies for eliciting prior distributions. For all priors discussed we assume that $g_{\mathrm{N}} = AA$ (in other words we assign prior probability zero to all vectors ψ^i with $g_{\mathrm{N}} \neq AA$). We

Nature Methods: doi:10.1038/nmeth.2883

assign uniform probability over the support. In other words, priors only differ in which states are assigned non-zero probability. All states with non-zero probability receive equal weight.

- **AB prior:** We assume that $g_R = AA$ and $g_V = AB$. Intuitively this means each mutation is assumed to be diploid and heterozygous.
- **BB prior:** We assume that $g_R = AA$ and $g_V = BB$. Intuitively this means each mutation is assumed to be diploid and homozygous.
- **No Zygosity prior (NZ):** We assume that $g_R = AA$, $c(g_V) = \overline{c}$ and $b(g_V) = 1$. In other words the genotype of the variant population has the predicted copy number with exactly one mutant allele. This is similar to the approach used in [14].
- **Total Copy Number prior (TCN):** We assume that $c(g_V) = \overline{c}$ and $b(g_V) \in \{1, \ldots, \overline{c}\}$. In other words the genotype of the variant population has the predicted copy number and at least one variant allele. We assume, with equal probability, that g_R is either AA or the genotype with $c(g_R) = \overline{c}$ and $b(g_R) = 0$. Intuitively this means the genotype of the variant population at the locus has the predicted total copy number and we consider the possibility that any number of copies (> 0) of the locus contains the mutant allele. We consider states where the reference population has the AA genotype or the genotype with the predicted copy number and all A's.
- **Parental Copy Number prior (PCN):** We assume that $c(g_V) = \overline{c}$ and $b(g_V) \in \{1, \overline{c_1}, \overline{c_2}\}$. In other words the genotype of the variant population has the genotype with the predicted copy number and one variant allele, or as many variant alleles as one of the parental copy numbers. When $b(g_V) \in \{\overline{c_1}, \overline{c_2}\}$ we assume $g_R = g_N$, in other words the mutation occurs before copy number events. When $b(g_V) = 1$ we assume g_R is the genotype with $c(g_R) = \overline{c}$ and $b(g_R) = 0$, in other words the mutation occurs after the copy number event. Intuitively this means each mutant locus has the predicted total copy number. We then consider if the mutation occurred before the copy number event, in which case the number of copies with the mutant allele should match one of the predicted parental copy numbers. Alternatively if the mutation occurs after the copy number event we assume only a single copy of the locus contains the mutant allele. This scheme assumes that a point mutation only occurs once. If more then one copy of the mutant allele is present in the variant population genotype, this occurred because the mutation preceded any copy number changes and was subsequently amplified.

5 Generation of synthetic data

To generate synthetic data for **Supplementary Figs. 12** and **13**, we sampled from the PyClone model with a Binomial emission letting $d_i \sim \text{Poisson}(10,000)$, $t = 0.75$, and using 8 clusters with cellular frequencies drawn from a $\text{Uniform}(0, 1)$ distribution. To assign genotypes to each

Nature Methods: doi:10.1038/nmeth.2883

mutation, we randomly sampled a total copy number, $\overline{c} \in \{1, \ldots, 5\}$. We sampled another value c^* uniformly from the set $\{0, 1, ..., \overline{c}\}$ and set the major copy number, $\overline{c}_1$, to $\max\{c^*, \overline{c} - c^*\}$ and the minor copy number, $\overline{c}_2$, to $\overline{c} - \overline{c}_1$. We randomly sample g_R from the set $\{g_N, g^*\}$ where $c(g^*) = \overline{c}$ and $b(g^*) = 0$. If $g_R = g_N$ then we assumed the mutation occurred early so that g_V had either $\overline{c}_1$ or $\overline{c}_2$ B alleles and total copy number $\overline{c}$. If $g_R \neq g_N$ we set g_V to the genotype with one variant allele and total copy number $\overline{c}$. For **Supplementary Fig. 12** we generated 100 simulated datasets by sampling 100 mutations for each dataset. For **Supplementary Fig. 13** we generated 400 datasets, 100 datasets with 50, 100, 500 and 1,000 mutations.

6 Implementation and availability

The code implementing all methods plus plotting and clustering is included in the PyClone software package. PyClone is implemented in the Python programming language. All analyses were performed using PyClone 0.12.4 and PyDP 0.2.1. PyDP is freely available under open source licensing. PyClone is freely available for academic use at `http://compbio.bccrc.ca/software/pyclone/` .

1. Fritsch, A. and Ickstadt, K. *Bayesian analysis* **4**, 367–391 (2009).
2. Ross, S.M. *Simulation* Elsevier third edition (2002).
3. Ferguson, T. *Annals of Statistics* **1**, 209–230 (1973).
4. Neal, R. *Journal of computational and graphical statistics* **9**, 249–265 (2000).
5. West, M. and Escobar, M. *Hierarchical priors and mixture models, with application in regression and density estimation* Institute of Statistics and Decision Sciences, Duke University (1993).
6. Heinrich, V. *et al. Nucleic Acids Res* **40**, 2426–31 (2012).
7. Yau, C. *et al. Genome Biol* **11**, R92 (2010).
8. Greenman, C.D. *et al. Biostatistics* **11**, 164–75 (2010).
9. Loo, P.V. *et al. Proceedings of the National Academy of Sciences of the United States of America* **107**, 16910–16915 (2010).
10. Ha, G. *et al. Genome Res* **22**, 1995–2007 (2012).
11. Boeva, V. *et al. Bioinformatics (Oxford, England)* **28**, 423–425 (2012).
12. Bignell, G.R. *et al. Nature* **463**, 893–898 (2010).
13. Curtis, C. *et al. Nature* **486**, 346–52 (2012).
14. Nik-Zainal, S. *et al. Cell* **149**, 994–1007 (2012).

Nature Methods: doi:10.1038/nmeth.2883