Online Methods and Supplementary information for Roth et al., PyClone: Statistical inference of tumour evolution from deep digital sequencing

Contents

1 Methods **1**
- 1.1 The PyClone model description . . . 1
- 1.2 Multiple sample modeling . . . 3
- 1.3 Overdispersion modelling . . . 3
- 1.4 Methods for eliciting PyClone priors over mutational genotypes . . . 4
- 1.5 Alternative Methods . . . 4
- 1.6 Normal cell mixing data . . . 4
- 1.7 High Grade Serous Ovarian Cancer . . . 5
- 1.8 MCMC analysis . . . 5
- 1.9 Evaluation and benchmarks . . . 5
- 1.10 Implementation and availability . . . 6

2 Supplementary Results **6**
- 2.1 Accounting for segmental copy number and mutational genotype improves performance in simulations . . . 6
- 2.2 Modeling mutational genotypes improves estimates of cellular prevalence in breast tumours . 6

3 Supplemental Methods **7**
- 3.1 Synthetic data . . . 7
- 3.2 Triple Negative Breast Cancer . . . 8

4 Discussion **8**
- 4.1 Incorporating external sources of copy number information . . . 8
- 4.2 Limitations . . . 8
- 4.3 Alternative Applications . . . 9

5 References **10**

1 Methods

1.1 The PyClone model description

{ sec:pyclone }

PyClone is a hierarchical Bayes statistical model (Figure S2). 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. 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 the probability that each pair

of mutations is clustered together. Figure S3 shows a typical experimental workflow used to produce the input data and information for setting priors.

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 (Figure S1b). 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 $g^n_{\mathsf{N}}, g^n_{\mathsf{R}}, g^n_{\mathsf{V}}$, 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 genotype with two reference alleles and one variant allele. The symbol $-$ denotes the genotype with no alleles i.e. a homozygous deletion of the locus. The vector $\boldsymbol{\psi^n} = (g^n_{\mathsf{N}}, g^n_{\mathsf{R}}, g^n_{\mathsf{V}}) \in \mathcal{G}^3$ 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. Thus, the *cellular prevalence* (the fraction of cells in the sample which contain the n^{th} mutation), is given by $t\phi^n$.

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.

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 $\boldsymbol{\psi}$ and cellular prevalence ϕ is given by:

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

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

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 $\boldsymbol{\psi}^n \in \mathcal{G}^3$ is infinite, however only a finite set of states will have non-zero prior probability which truncates the sum.

We incorporate the assumption that mutations defining a clonal genotype appear at the same cellular prevalence into the prior over $p(\phi)$ using a Dirichlet Process (DP) prior with base measure $H_0 \sim \text{Uniform}(0, 1)$. The DP is a distribution over the space of distributions and allows for automatic inference of the number of clusters in the problem [1].

Due to the presence of the DP prior, computing the exact posterior distribution is not tractable. We use an auxiliary variable sampling method [2] to perform Markov Chain Monte Carlo (MCMC) sampling from the posterior distribution, alternating between updating the cluster memberships using an auxiliary variable scheme, and re-sampling the cluster values ϕ. Cluster values in the second step 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 [3].

1.2 Multiple sample 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, $\boldsymbol{\phi}^n = (\phi_1^n \dots \phi_M^n)$ from this measure. The likelihood under this model is given by

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

For each mutation n we assign different priors, $\boldsymbol{\pi}_m^n$, for each sample allowing for 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 $\boldsymbol{t} = (t_1, \dots, t_M)$ which contains the tumour content of each sample. Using this approach the partition (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.

1.3 Overdispersion modelling

As mentioned in the introduction, next generation sequencing data are often overdispersed [4]. We implemented a version of the PyClone framework which replaces the Binomial distribution with a Beta-Binomial distribution, parameterised 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.

1.4 Methods for eliciting PyClone priors over mutational genotypes

The PyClone model requires that we specify prior $\pi^n_{\boldsymbol{\psi}}$ on the normal, reference and variant genotypes $\boldsymbol{\psi^n} = (g^n_{\mathsf{N}}, g^n_{\mathsf{R}}, g^n_{\mathsf{V}})$ at mutation site n. A number of methods are available to profile parental and total copy number from high density genotyping arrays [5, 6], or from whole genome sequencing data [7, 8]. As segmental aneuploidies and loss of heterozygosity are accepted to be an essential part of the tumour genome landscape [9, 10], 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 implemented five prior distributions described below. For all priors discussed, we assume that $g_{\mathsf{N}} = AA$ (i.e. we assign prior probability zero to all vectors ψ^i with $g_{\mathsf{N}} \neq AA$), and we assign uniform probability over the support (in other words, priors only differ in the construction of their support).

AB prior: We assume that $g_{\mathsf{R}} = AA$ and $g_{\mathsf{V}} = AB$.

BB prior: We assume that $g_{\mathsf{R}} = AA$ and $g_{\mathsf{V}} = BB$.

No Zygosity prior (NZ): We assume that $g_{\mathsf{R}} = AA$, $c(g_{\mathsf{V}}) = \overline{c}$ and $b(g_{\mathsf{V}}) = 1$ i.e. the genotype of the variant population has the predicted copy number with exactly one mutant allele. This is similar to the approach used in [11].

Total Copy Number prior (TCN): We assume that $c(g_{\mathsf{V}}) = \overline{c}$ and $b(g_{\mathsf{V}}) \in \{1, \ldots, \overline{c}\}$ with equal probability, i.e. 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 $c(g_{\mathsf{R}}) = \overline{c}$ and $b(g_{\mathsf{R}}) = 0$.

Parental Copy Number prior (PCN): We assume that $c(g_{\mathsf{V}}) = \overline{c_1} + \overline{c_2}$ and $b(g_{\mathsf{V}}) \in \{1, \overline{c_1}, \overline{c_2}\}$ i.e. 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 $b(g_{\mathsf{V}}) \in \{\overline{c_1}, \overline{c_2}\}$ we assume $g_{\mathsf{R}} = AA$ i.e. the mutation occurs before copy number events. When $b(g_{\mathsf{V}}) = 1$ we assume g_{R} is the genotype with $c(g_{\mathsf{R}}) = \overline{c}$ and $b(g_{\mathsf{R}}) = 0$ i.e. the mutation occurs after the copy number event.

1.5 Alternative Methods

To compare our method to other clustering methods that ignore genotype, we have also implemented the following models in the PyClone software: the infinite Gaussian mixture model (IGMM), the infinite Binomial mixture model (IBMM), and the infinite Beta Binomial mixture model (IBBMM). We interpreted the mean parameter for the IGMM, the probability of success for the IBMM, and the mean parameter m for the IBBMM as the cellular prevalence of mutations. All MCMC analysis, clustering and cellular frequency inference was done as described below. PyClone also uses the PyDP package to implement DP inference for all models so that any variation in performance should be due to differing distributional assumptions, not inference methods or implementation.

1.6 Normal cell mixing data

For the idealized mixture data presented in Figure 1, MiSeq data from mixture experiments A (SRR385938), B (SRR385939), C (SRR385940) and D (SRR385941) from [12] where downloaded from the NCBI short read archive. FASTQ files were extracted from the downloaded .sra files using the `fastq-dump --split-files --clip` command from the NCBI SRA-SDK version 2.1.16. Sequences were aligned to the targeted genome using `bwasw` from the bwa 0.6.2 package. The mpileup files were generate from the aligned BAM files using the `mpileup -f REFGENOME -d 1000000 -s BAMFILE` command from the samtools 0.1.18 package. Count data was extracted for the positions of interest from the mpileup files using a custom Python script which filtered out positions with base or mapping qualities below 10. The count data was the post-processed to remove positions showing significant strand bias as determined using a Fisher exact test.

Primer start and stop positions for the UDT-Seq protocol were obtained from the supplemental table 2 of [12]. We downloaded hg19 from UCSC website and used primer positions to build a targeted reference alignment file which contained only the regions spanned by the primers.

Variants positions in the four cases used were previously identified in [13]. We compiled a list of positions with variants in only one of the four cases for use in our analysis. We manually removed positions which appeared to have outlying variant allelic prevalence in one of the four mixtures. The outliers were either due to incorrect annotation of the genotype in one of the four cases, or because sequencing appeared to work poorly at the target location. The position coordinates were converted from hg18 to hg19 coordinates using the UCSC liftover program.

The position selection simulated an idealized mixture of four clonal cell populations sharing no mutations and with diploid genomes providing ground truth for systematic evaluation. Because the dataset was highly imbalanced for variants with the BB genotype we randomly down sampled the BB positions to obtain 10 smaller datasets of 30 mutations with 50:50 mixtures of positions with AB and BB genotypes for each case.

The predicted genotype from [13] was used to determine the homologous copy number for the PCN as follows: for the positions predicted to have the AB genotype in the variant sample we set the major and minor copy numbers to 1; for positions with the BB genotype in the variant sample we set the major copy number to 2 and the minor copy number to 0.

Benchmarking in this experiment was measured by the ability to correctly group mutations based on the known, but held-out reference clustering (Figure 1a). Reference clustering was defined by the case harbouring the variant genotype. We did not attempt to benchmark cellular prevalence estimates for this dataset since the mixing proportions are only approximately correct due to experimental variability.

1.7 High Grade Serous Ovarian Cancer

For the two HGSOC cases (Figure 2), we used PICNIC [6] (downloaded 07/04/13) to analyse the Affymetrix SNP6.0 array data from all four samples for both cases and infer homologous copy number. We specified the priors for cellularity estimation in PICNIC based on quantitative pathology estimates. Allelic count data was obtained from Bashashati *et al.* [14]. We used the homologous copy number information to elicit priors for PyClone as discussed above, and set the tumour content for the PyClone analysis to the value predicted by PICNIC. MCMC analysis and post-processing of the trace was done as discussed above.

1.8 MCMC analysis

With the exception of the mixture of normal tissue experiments, all experiments involving the PyClone, IGMM, IBMM and IBBMM models were ran for 10,000 MCMC iterations discarding the first 1,000 samples as burnin. For the mixture of normal tissue experiments, we ran the MCMC chains for 100,000 iterations discarding the first 50,000 samples as burnin. To generate cellular frequency plots we fit Gaussian kernel density estimators to the post-burning MCMC trace using the scipy Python library. To cluster the data we formed the pairwise posterior similarity matrix (the matrix indicating how frequently any two mutations appeared in the same cluster in the post-burnin trace). We then hierarchically clustered the data using average linkage, and the resulting dendrogram was used as a guide to find a clustering which optimised the MPEAR criterion described in [15] (the code for doing this is built into PyClone).

1.9 Evaluation and benchmarks

To assess the clustering performance of the methods we computed the V-measure [16] , calculated using the scikits-learn Python package 0.11. Cellular prevalence estimates were evaluated using the mean error over MCMC samples. Statistical analysis was performed with the `aov` and `TukeyHSD` functions in the R statistical computing package using RStudio v.0.96.331.

1.10 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 0.12.0 and PyDP 0.2.0. PyDP is freely available under open source licensing (details to be completed on publication).

2 Supplementary Results

2.1 Accounting for segmental copy number and mutational genotype improves performance in simulations

{subsect:synthetic}

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 above. Benchmarking was based on V-measure (a measure of clustering accuracy between 0 and 1 where a V-measure score of 1 represents perfect clustering) and mean error in estimating cellular prevalence (where a mean error of 0 represents perfect cellular prevalence estimates). 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 Methods below.

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) (Figure S4a). The PCN and TCN methods, which account for mutational genotype, significantly outperformed all other methods (Figure S4a). 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 (Figure S4b). 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 sequencing data.

2.2 Modeling mutational genotypes improves estimates of cellular prevalence in breast tumours

We next sought to compare results in data derived from patient tumours. Segmental copy number changes due to karyotypic abnormalities are a hallmark of cancers, and are a dominant feature in solid epithelial malignancies [9, 10]. For example, triple negative breast cancers (TNBC) are characterized by unstable genomes with the majority of the genome subject to population-level recurrent alterations [17]. As a result, reliable inference of cellular prevalences can be challenging and consideration of copy number is essential for correct estimates. We analysed two TNBC cases: SA029 and SA223 with known and interpretable

mutations using PyClone and IBBMM using a panel of validated somatic mutations ($N = 12, 15$ respectively) originally presented in [17]. We highlight the raw variant allelic prevalence, the inferred cellular prevalence using an IBBMM, and the inferred cellular prevalence using PyClone with the PCN method and Beta-Binomial emissions for each case (Figure S5).

Based on the raw allelic frequencies or genotype-naive clustering, we would identify at least two clusters in SA029 (Figure S5a,b, Table S1). The mutation at the highest allelic prevalence (0.46 of reads) is a missense mutation in *RB1*, resulting in a p.L657P amino acid change. This mutation was clustered on its own by IBBMM (Figure S5b). Mutations in *RB1* are frequently homozygous, exhibiting a typical tumour suppressor profile of bi-allelic loss. The *RB1* locus in SA029 falls within a region of copy neutral loss of heterozygosity [7]. The remaining eleven mutations in SA029 all appear at roughly half (0.22±0.02) the allelic prevalence of the *RB1* mutation. In contrast to IBBMM, PyClone outputs a single cluster of mutations (Figure S5c), suggesting the eleven mutations are likely heterozygous, the *RB1* mutation is homozygous, and with all mutations sharing cellular prevalence close to 1.0. This simple example illustrates the fundamental importance of accounting for mutational genotype when clustering mutations in patient tumour data. Failure to account for the mutational genotype in this case would lead to spurious inference about the population structure.

The second case example is SA223. This sample contained $N = 15$ mutations, including the known missense p.H1047R 'hotspot' mutation in *PIK3CA* (Table S2). The recurrence pattern of *PIK3CA* in TNBC [17, 18] suggests it is likely to be amongst the earliest tumourigenic mutations in disease progression. The raw variant allele prevalence for *PIK3CA* (0.66) was not the maximum in the set of 15 mutations (Figure S5d). Rather, a mutation in *DIAPH1* coincident with a LOH region exhibited higher variant allele prevalences (0.85). The cellular prevalence estimates of IBBMM method (Figure S5e) sorts the mutations in descending order of prevalence as *DIAPH1*, followed by *PIK3CA*, implying *DIAPH1* arose prior to *PIK3CA* in the tumour's evolutionary history. By contrast, the PyClone estimates (Figure S5f) sort the mutations such that *PIK3CA* and *DIAPH1* were both part of the ancestral clone and occur in all cells. The copy number profiles can explain the discrepancy in allelic prevalences. The *DIAPH1* mutation is in a region of copy neutral LOH while the *PIK3CA* mutation is in a region of copy number gain. We suggest that a heterozygous *PIK3CA* mutation arose followed by a single copy gain of the mutant allele as described previously [19], providing an additive oncogenic effect as a substrate for selection. The *DIAPH1* mutation resides in a region of chromosome 5q that is frequently deleted in TNBC [7, 17]. Thus, we infer the *DIAPH1* mutation was acquired early (likely with *PIK3CA*) and was rendered homozygous by uniparental disomy.

Analysis of the mutations in SA029 and SA223 show how inference on cellular prevalence and mutational clustering is substantially affected in karyotypically disrupted genomes. Thus, interpretation of evolutionary dynamics is significantly impacted by statistical modelling assumptions that consider mutational genotype. The specific examples of PIK3CA and DIAPH1 show how both copy number amplification and deletion influence cellular prevalence estimates. Although ground truth is difficult to establish, the PyClone cellular prevalence estimates are intuitive based on the known homozygous properties of *RB1* mutation and the likelihood that *PIK3CA* is an early tumourigenic driver in TNBC aetiology.

3 Supplemental Methods

3.1 Synthetic data

We generated 100 simulated datasets (Figure S4) by sampling 100 mutations for each dataset from the PyClone model with $d_i \sim \text{Poisson}(10,000)$, $t = 0.75$, and 8 clusters with cellular frequencies drawn from a $\text{Uniform}(0, 1)$ distribution. To assign genotypes to each mutation, we randomly sampled a total copy number, $\overline{c}$. 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 $\{g_{\text{N}}, g^*\}$ where $c(g^*) = \overline{c}$ and $b(g^*) = 0$. If $g_{\text{R}} = g_{\text{N}}$ 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 $g_{\text{R}} \neq g_{\text{N}}$ we set g_{V} to the genotype with one variant allele and total copy number $\overline{c}$.

3.2 Triple Negative Breast Cancer

For the two TNBC samples (Figure S5), we used PICNIC [6] (downloaded 07/04/13) to analyse the Affymetrix SNP6.0 array data for both cases and infer homologous copy number. We specified the priors for cellularity estimation in PICNIC based on the pathology estimates. Since the pathology estimates were categorical values, ”high” or ”moderate”, we assumed high was tumour content of 0.8 and moderate was tumour content of 0.5. Allelic count data for the mutations was obtained from [7]. We used the homologous copy number information to elicit priors for PyClone as discussed above, and set the tumour content for the PyClone analysis to the value predicted by PICNIC. MCMC analysis and post-processing of the trace was done as discussed above.

4 Discussion

4.1 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 [20], 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 [21]. 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 [22]. In this case it would be more likely that the reference population has the same genotype 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.

4.2 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 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 ancestral reconstruction of individual cells.

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 [23]. 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 maybe 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.

The posterior densities for the cellular frequencies can often exhibit a degree of uncertainty (Figure S5f) 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) and/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.

4.3 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.

5 References

References

[1] Ferguson, T. *Annals of Statistics* **1**, 209–230 (1973).

[2] Neal, R. *Journal of computational and graphical statistics* **9**, 249–265 (2000).

[3] 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).

[4] Heinrich, V. *et al. Nucleic Acids Res* **40**, 2426–31 (2012).

[5] Yau, C. *et al. Genome Biol* **11**, R92 (2010).

[6] Greenman, C.D. *et al. Biostatistics* **11**, 164–75 (2010).

[7] Ha, G. *et al. Genome Res* **22**, 1995–2007 (2012).

[8] Boeva, V. *et al. Bioinformatics (Oxford, England)* **28**, 423–425 (2012).

[9] Bignell, G.R. *et al. Nature* **463**, 893–898 (2010).

[10] Curtis, C. *et al. Nature* **486**, 346–52 (2012).

[11] Nik-Zainal, S. *et al. Cell* **149**, 994–1007 (2012).

[12] Harismendy, O. *et al. Genome Biol* **12**, R124 (2011).

[13] Ng, S.B. *et al. Nature* **461**, 272–6 (2009).

[14] Bashashati, A. *et al. Journal of Pathology* (2013 (accepted)).

[15] Fritsch, A. and Ickstadt, K. *Bayesian analysis* **4**, 367–391 (2009).

[16] Rosenberg, A. and Hirschberg, J. in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* volume 410 page 420 (2007).

[17] Shah, S.P. *et al. Nature* **486**, 395–9 (2012).

[18] Cancer Genome Atlas Network *Nature* **490**, 61–70 (2012).

[19] Kadota, M. *et al. Cancer research* **69**, 7357–7365 (2009).

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

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

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

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

Supplementary Figures

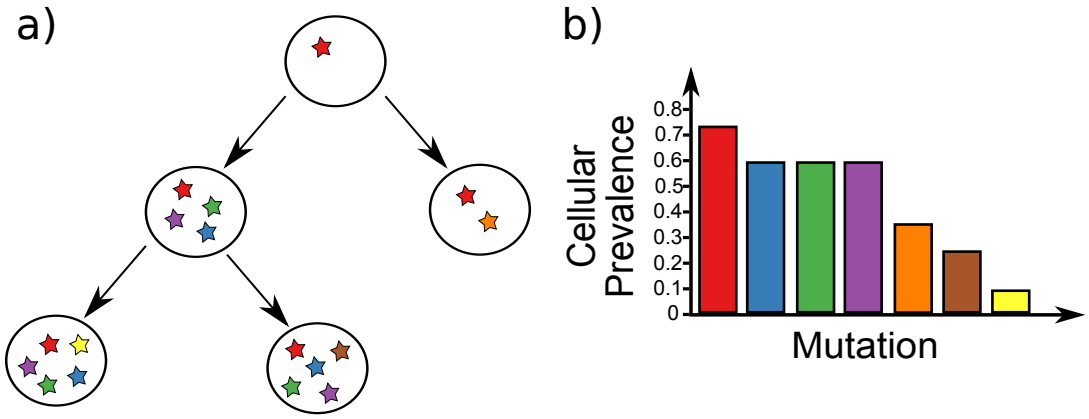
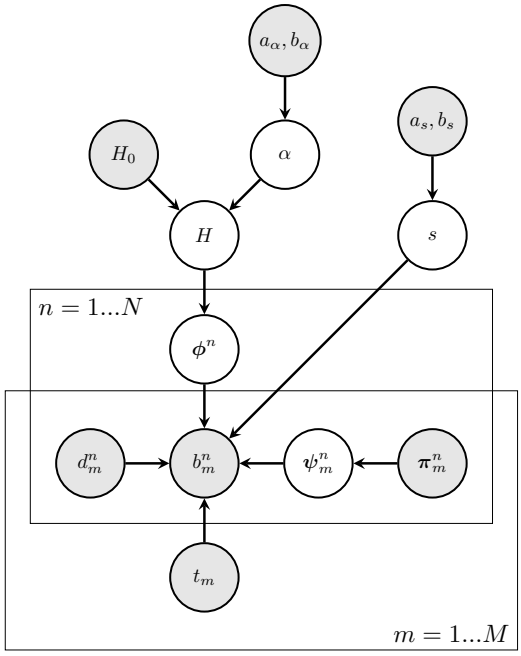


Figure S1: **a**) A hypothetical phylogenetic tree (left) generated by clonal expansion via the accumulation of mutations (stars). Unlike traditional phylogenetic trees internal nodes (clones) in the tree may contribute to the observed data, not just the leaf nodes. Hypothetical observed cellular prevalences (right) for the mutations in tree. Mutations occurring higher up the tree are 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 co-occur in the clones of the tree. In addition, the cellular prevalence of the dominant mutation (red) is below 1.0 due to the presence of contaminating normal tissue. **b**) 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 $\frac{6 \cdot 4 \cdot \frac{2}{4}}{2 \cdot 2 + 4 \cdot 3 + 6 \cdot 4} = 0.3$. {fig:clonal˙evolution}



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

Figure S2: 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, $\boldsymbol{\phi}^{\boldsymbol{n}}$, 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. {fig:pgm}

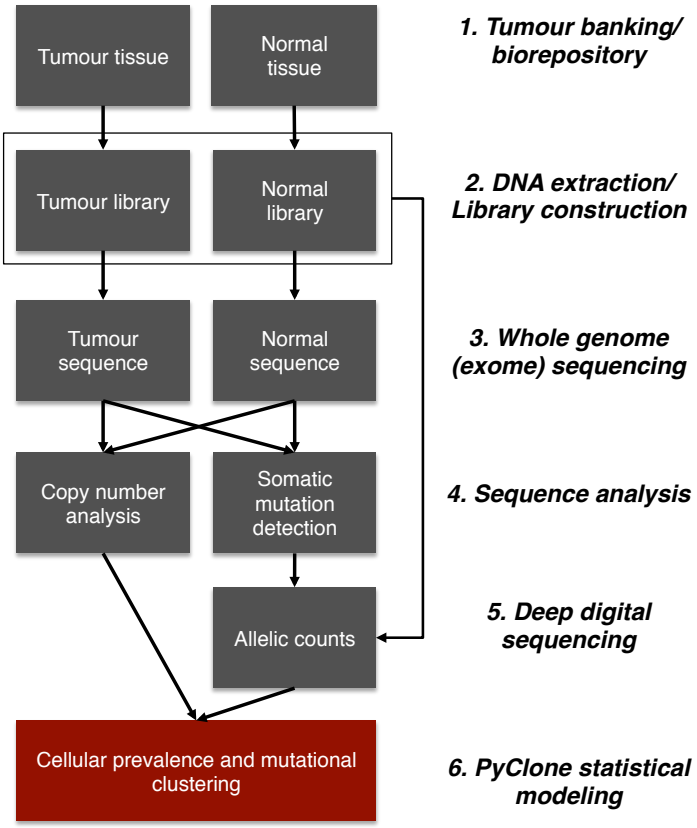


Figure S3: Workflow for PyClone analysis. The sample is first assayed using whole genome shotgun sequencing (WGSS) or exome capture sequencing to identify putative mutations. Copy number information, which is used to inform the PyClone priors, can be derived from either sequence or array data. Putative 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.

{fig:workflow}

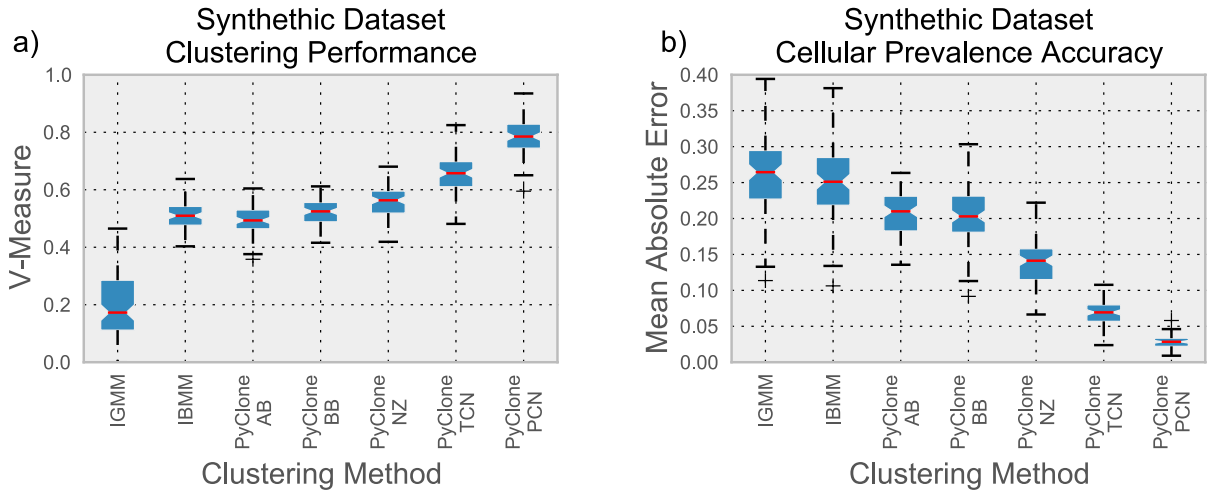


Figure S4: a) Clustering performance and b) estimated cellular prevalence accuracy for different methods applied to 100 synthetic data. In figure a) the accuracy of the inferred clusters is measured using the V-measure metric (y-axis). In figure b) the accuracy of the estimated cellular prevalence is measured as follows. For each mutation the mean of the post-burning trace was used as a point estimate. The absolute difference between this value and the true value for all 100 mutations in each dataset was computed. For each of the 100 datasets the mean value of the absolute error across mutations was taken and used to generate the boxplots.

{fig:synthetic}

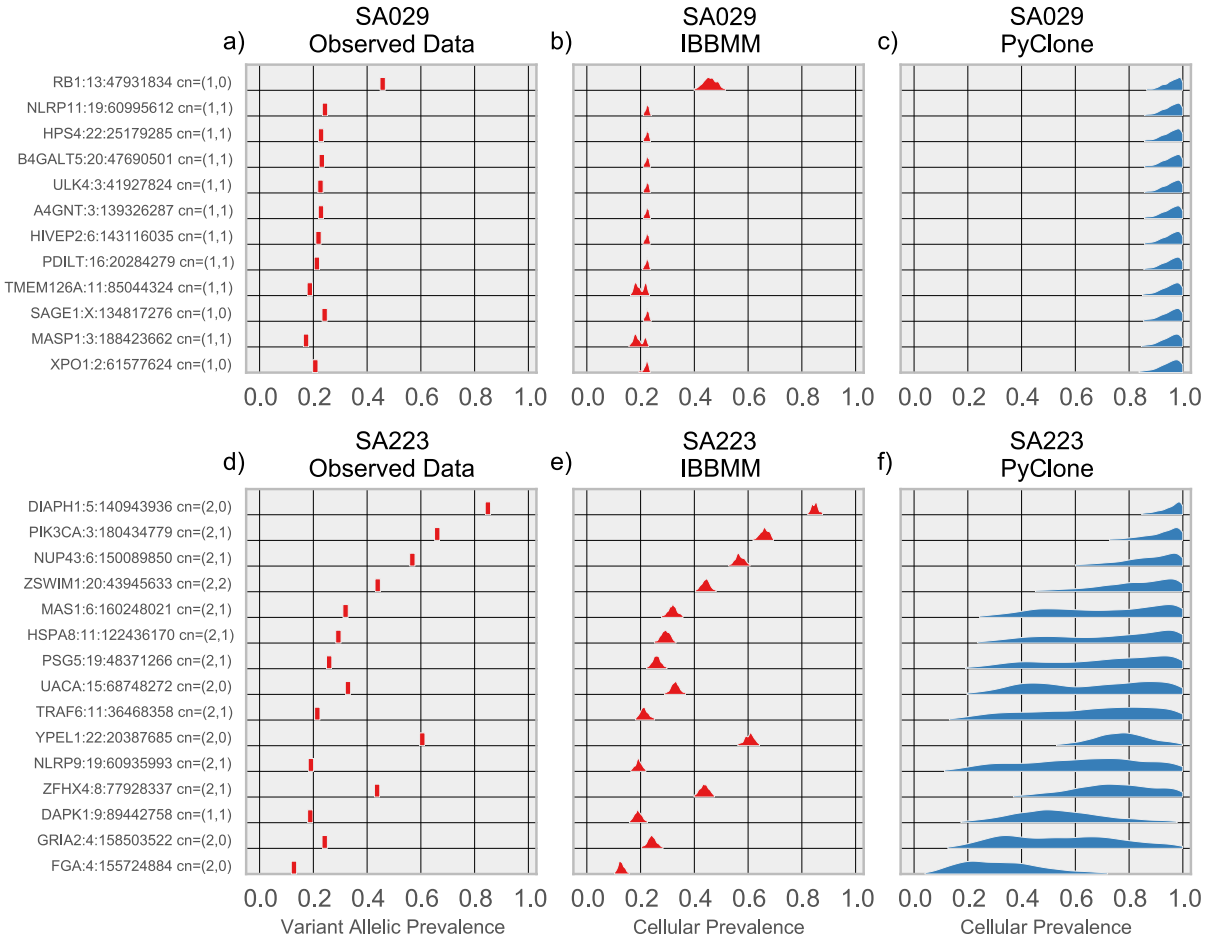


Figure S5: Panels a) and d) show the observed variant allelic prevalence for mutations from triple negative breast cancer (TNBC) cases SA029 and SA223 respectively. Panels b) and e) show the inferred cellular prevalence for each mutation in SA029 and SA223 using the IBBMM method. Panels c) and f) show the inferred cellular prevalence of mutations in SA029 and SA223 using PyClone with the PCN method for specifying priors and a Beta Binomial emission density. Each mutation is annotated with the gene containing the mutation, hg18 genomic coordinates, and predicted major and minor copy number (cn=(major cn, minor cn)).

{fig:tn}

Supplementary Tables

Table S1: Allelic counts, IBBMM and PyClone PCN cellular prevalence estimates for mutations in triple negative breast cancer SA029. Copy number predictions where inferred using PICNIC as described in the text. Cellular prevalences where computed by taking the mean of the post burnin trace for the cellular prevalences for the respective methods. Cluster ids where predicted from the post burnin trace using the mpear clustering criteria as described in the methods. Mutation ids list gene name, chromosome and chromosome coordinate. All coordinates are in the hg18 coordinate system.

{stab:tnbcSA029}

Table S2: Allelic counts, IBBMM and PyClone PCN cellular prevalence estimates for mutations in triple negative breast cancer SA223. Copy number predictions where inferred using PICNIC as described in the text. Cellular prevalences where computed by taking the mean of the post burnin trace for the cellular prevalences for the respective methods. Cluster ids where predicted from the post burnin trace using the mpear clustering criteria as described in the methods. Mutation ids list gene name, chromosome and chromosome coordinate. All coordinates are in the hg18 coordinate system.

{stab:tnbcSA223}

Table S3: Allelic counts, IBBMM and PyClone PCN cellular prevalence estimates for mutations in high grade serous ovarian cancer Case1. Copy number predictions where inferred using PICNIC as described in the text. Cellular prevalences where computed by taking the mean of the post burnin trace for the cellular prevalences for the respective methods. Cluster ids where predicted from the post burnin trace using the mpear clustering criteria as described in the methods. Mutation ids list gene name, chromosome and chromosome coordinate. All coordinates are in the hg18 coordinate system.

{stab:hgsc1}

Table S4: Allelic counts, IBBMM and PyClone PCN cellular prevalence estimates for mutations in high grade serous ovarian cancer Case2. Copy number predictions where inferred using PICNIC as described in the text. Cellular prevalences where computed by taking the mean of the post burnin trace for the cellular prevalences for the respective methods. Cluster ids where predicted from the post burnin trace using the mpear clustering criteria as described in the methods. Mutation ids list gene name, chromosome and chromosome coordinate. All coordinates are in the hg18 coordinate system.

{stab:hgsc2}