PyClone: statistical inference of clonal population structure in cancer

Andrew Roth[1,2], Jaswinder Khattra[2], Damian Yap[2], Adrian Wan[2], Emma Laks[2], Justina Biele[2], Gavin Ha[1,2], Samuel Aparicio[2,3], Alexandre Bouchard-Côté[4] & Sohrab P Shah[2,3]

We introduce PyClone, a statistical model for inference of clonal population structures in cancers. PyClone is a Bayesian clustering method for grouping sets of deeply sequenced somatic mutations into putative clonal clusters while estimating their cellular prevalences and accounting for allelic imbalances introduced by segmental copy-number changes and normal-cell contamination. Single-cell sequencing validation demonstrates PyClone's accuracy.

Human cancer progresses under Darwinian evolution, in which genetic or epigenetic variation alters molecular phenotypes in individual cells[1]. Consequently, tumors at diagnosis often consist of multiple, genotypically distinct cell populations[2] (**Supplementary Fig. 1**). These populations, referred to as clones, are related through a phylogeny and act as substrates for selection in tumor microenvironments or with therapeutic intervention[2,3]. The prevalence of a particular clone measured over time or in anatomic space is a reflection of its growth and proliferative fitness. Thus, ascertaining the dynamic prevalence of clones can help identify precise genetic determinants of phenotypes such as acquisition of metastatic potential or chemotherapeutic resistance.

We provide a hierarchical Bayes statistical model, PyClone (**Supplementary Figs. 2** and **3**), for analysis of deeply sequenced (coverage > 100×) mutations to identify and quantify clonal populations in tumors, which extends to modeling mutations measured in multiple samples from the same patient. Our approach uses the measurement of allelic prevalence to estimate the proportion of tumor cells harboring a mutation (referred to herein as the 'cellular prevalence'). Owing to the cell lysis involved in the preparation of bulk samples for sequencing, we cannot determine the complete set of genomic aberrations defining a clonal population. However, assuming that clonal populations follow a perfect (that is, no site mutates more than once in its evolutionary history, and each harbors at most one somatic mutant genotype) and persistent (mutations do not disappear or revert) phylogeny, we can identify and quantify the prevalence of clonal populations. These assumptions imply that clusters of mutations occurring at the same point in the clonal phylogeny are present at shared cellular prevalences. Thus, clusters of mutations can be used as markers of clonal populations. (Limitations for when these assumptions do not hold are presented in the **Supplementary Discussion**.)

Despite progress in measuring allele prevalence with deep sequencing[4–8], statistical approaches to cluster deep digital sequencing of mutations into biologically relevant groupings remain underdeveloped, with poorly understood analytical assumptions. The allelic prevalence of a mutation is a compound measure of several factors: the proportion of 'contaminating' normal cells, the proportion of tumor cells harboring the mutation and the number of allelic copies of the mutation in each cell, plus uncharacterized sources of technical noise. Consequently, allelic prevalence does not straightforwardly relate to cellular prevalence. This is particularly exacerbated when only single sample allelic prevalence estimates are taken. Multiple sample measurements (taken across time[5,9] or space[10]) have two distinct advantages: reduction of noise owing to repeated measures and the potential to identify sets of mutations whose cellular prevalences shift together. The latter is a route to precisely identifying clones whose prevalences are changing under selective pressures.

PyClone systematically addresses these deficiencies (**Supplementary Figs. 2** and **3**). The inputs to the model are a set of deeply sequenced mutations from one or more samples derived from a single patient and a measure of allele-specific copy number at each mutation locus in each sample (**Supplementary Fig. 4**). Allele-specific copy number can be obtained from high-density genotyping arrays or from whole-genome sequencing data (Online Methods). PyClone outputs posterior densities for model parameters such as the cellular prevalence for each mutation in the input (**Supplementary Fig. 5**) and the clustering structure over the mutations (**Supplementary Figs. 6–11**). The posterior densities of these quantities are then postprocessed to give interpretable point estimates of the cellular prevalences and mutational clustering.

The PyClone framework (full mathematical and implementation details are in the **Supplementary Note**) overcomes the challenges outlined earlier through four novel modeling advances. First, it uses beta-binomial emission densities, which models data sets with more variance in allelic prevalence measurements more effectively than a binomial model. Second, flexible prior probability estimates ('priors') of possible mutational genotypes are used, reflecting how allelic prevalence measurements are deterministically linked to zygosity and coincident copy-number

[1]Bioinformatics Graduate Program, University of British Columbia, Vancouver, British Columbia, Canada. [2]Department of Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada. [3]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [4]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada. Correspondence should be addressed to S.P.S. (sshah@bccrc.ca).

RECEIVED 18 NOVEMBER 2013; ACCEPTED 31 JANUARY 2014; PUBLISHED ONLINE 16 MARCH 2014; DOI:10.1038/NMETH.2883

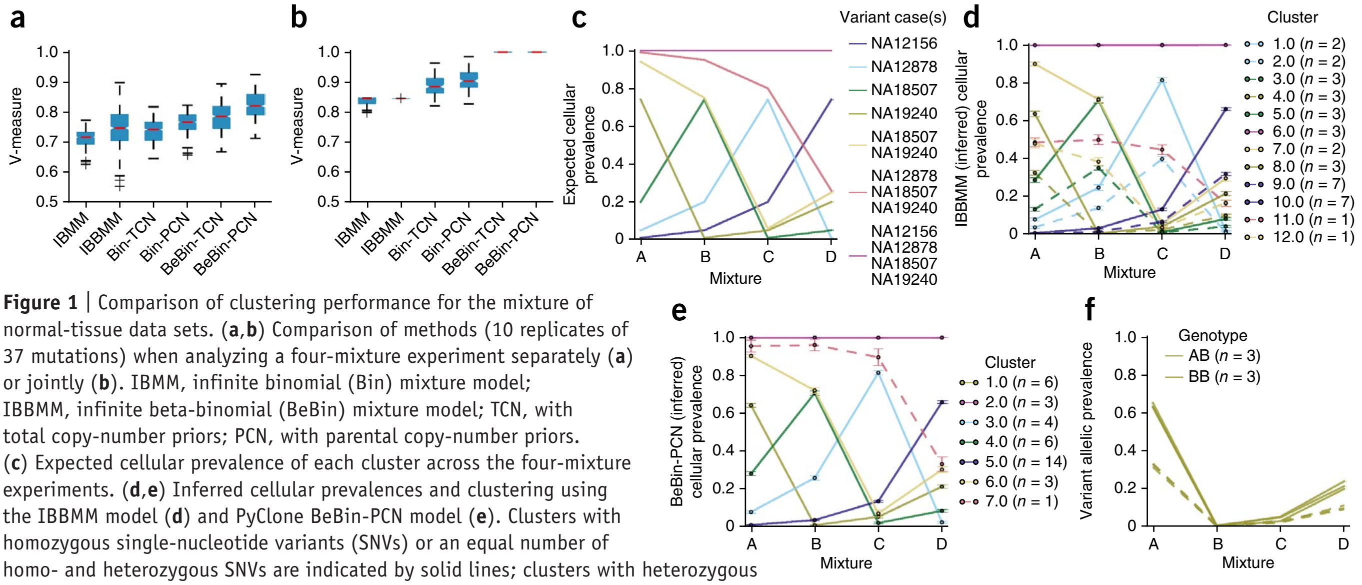

Figure 1 | Comparison of clustering performance for the mixture of normal-tissue data sets. (**a**,**b**) Comparison of methods (10 replicates of 37 mutations) when analyzing a four-mixture experiment separately (**a**) or jointly (**b**). IBMM, infinite binomial (Bin) mixture model; IBBMM, infinite beta-binomial (BeBin) mixture model; TCN, with total copy-number priors; PCN, with parental copy-number priors. (**c**) Expected cellular prevalence of each cluster across the four-mixture experiments. (**d**,**e**) Inferred cellular prevalences and clustering using the IBBMM model (**d**) and PyClone BeBin-PCN model (**e**). Clusters with homozygous single-nucleotide variants (SNVs) or an equal number of homo- and heterozygous SNVs are indicated by solid lines; clusters with heterozygous SNVs, by dashed lines. (**f**) Variant allelic prevalence for mutations assigned to cluster 1 by the PyClone BeBin-PCN model. Dashed lines represent heterozygous SNVs; solid lines represent homozygous SNVs. Whiskers (**a**,**b**) indicate 1.5× the interquartile range, red bars indicate the median and boxes represent the interquartile range. Error bars (**d**,**e**) indicate the mean s.d. (using 9,000 post-'burn-in' samples (Online Methods)) of Markov chain Monte Carlo–derived cellular prevalence estimates over all mutations in a cluster. (**d**–**f**) The number of mutations *n* in each cluster is shown in the legends in parentheses.

variation events. Third, Bayesian nonparametric clustering is used to discover groupings of mutations and the number of groups simultaneously. This obviates fixing the number of groups a priori and allows for cellular prevalence estimates to reflect uncertainty in this parameter. Fourth, multiple samples from the same cancer may be analyzed jointly to leverage the scenario in which clonal populations are shared across samples. Simulated data sets systematically illustrate improvements in performance of each of the novel modeling components (**Supplementary Figs. 12** and **13**).

We first evaluated our approach on idealized data sets (**Fig. 1**) produced from physical mixtures of DNA extracted from four 1000 Genomes Project samples[11,12] (Online Methods). Each mixture contained DNA in approximate proportions of 0.01, 0.05, 0.20 and 0.74. Specific single-nucleotide variants (SNVs) were amplified using PCR and then sequenced deeply on the Illumina MiSeq platform. This data set simulates a hierarchically related population with ground truth for quantitative benchmarking (**Fig. 1c**). We selected positions with variants present in exactly one case,

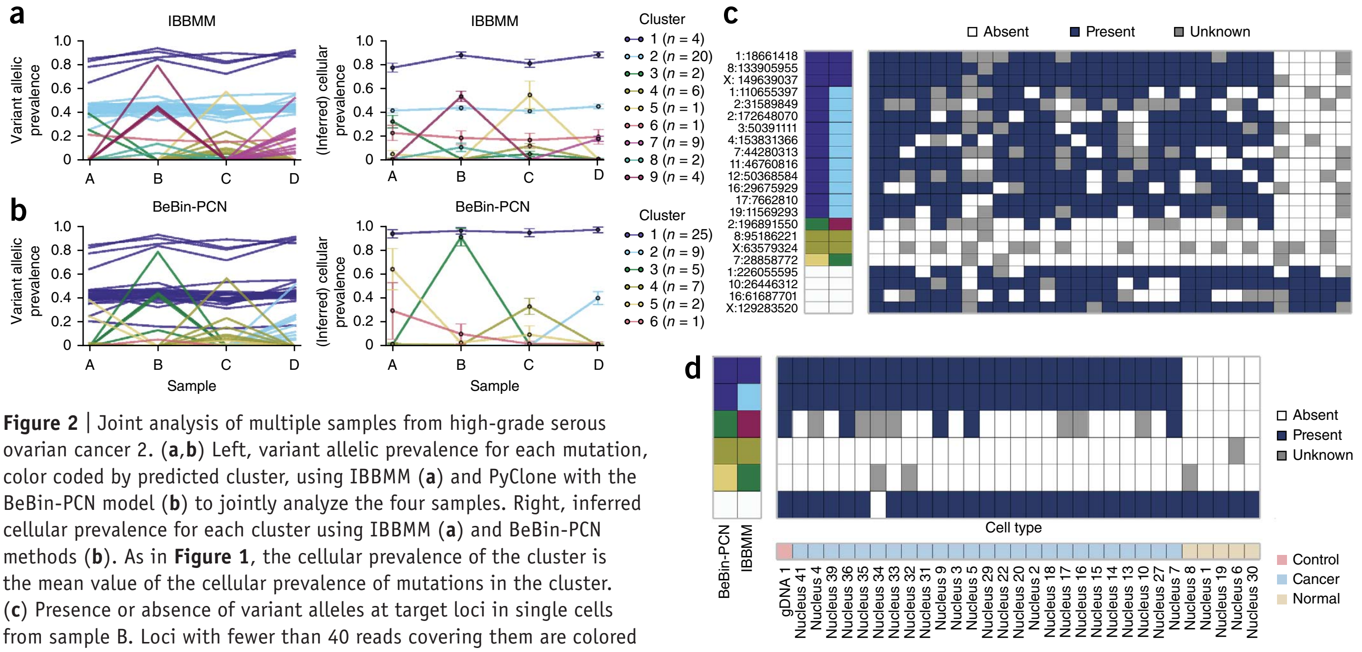

Figure 2 | Joint analysis of multiple samples from high-grade serous ovarian cancer 2. (**a**,**b**) Left, variant allelic prevalence for each mutation, color coded by predicted cluster, using IBBMM (**a**) and PyClone with the BeBin-PCN model (**b**) to jointly analyze the four samples. Right, inferred cellular prevalence for each cluster using IBBMM (**a**) and BeBin-PCN methods (**b**). As in **Figure 1**, the cellular prevalence of the cluster is the mean value of the cellular prevalence of mutations in the cluster. (**c**) Presence or absence of variant alleles at target loci in single cells from sample B. Loci with fewer than 40 reads covering them are colored gray. Predicted clusters for each method are shown on the left, with white cells indicating nonsomatic control positions. Row labels indicate hg19 chromosome and chromosome coordinate separated by a colon. (**d**) Presence or absence of IBBMM clusters in single cells from sample B. Clusters were deemed present if any mutation in the cluster was present. gDNA, bulk genomic DNA control. Error bars (**a**,**b**) indicate the mean s.d. (using 50,000 post-'burn-in' samples) of Markov chain Monte Carlo–derived cellular prevalence estimates over all mutations in a cluster. The number of mutations *n* in each cluster is shown in the legends in parentheses.

shared by specific subsets of cases, and shared by all cases (Online Methods). Using combinations of emission densities and genotype priors, we compared PyClone to two genotype-naive methods: an infinite binomial mixture model (IBMM) and an infinite beta-binomial mixture model (IBBMM) (Online Methods). Empirical comparisons to ground truth showed that PyClone with beta-binomial emission densities with parental (BeBin-PCN) and total copy-number (BeBin-TCN) priors outperformed all other methods on the basis of clustering accuracy, as determined by the 'validity' V-measure[13] (Online Methods). Performance gains were consistent when we analyzed each data set separately (**Fig. 1a**) or all four samples jointly (**Fig. 1b**). Accounting for mutational genotype and joint inference each conferred independent performance gains, with the best results obtained when we used PyClone BeBin-PCN. To illustrate the effect of accounting for mutational genotype, we randomly selected one of the ten joint analysis runs contrasting PyClone BeBin-PCN with IBBMM, a baseline analogous to clustering the raw allelic data without considering mutational genotype. IBBMM output 12 clusters, consistently assigning heterozygous and homozygous SNVs from each ground-truth cluster to separate clusters (**Fig. 1d**). By contrast, PyClone identified the seven correct clusters (**Fig. 1e**), placing heterozygous and homozygous SNVs from the same clusters together (**Fig. 1f**).

We next compared results from multisample BeBin-PCN and IBBMM in the cancer setting, using recently published mutational profiles of multiple samples from a high-grade serous ovarian cancer (HGSOC)[14]. Four spatially separated samples were taken from a primary, untreated ovarian tumor (study case 2). Mutations called in exomes were deeply sequenced (~5,000×), yielding 49 validated mutations. Copy-number priors were obtained from high-density genotyping arrays (Online Methods).

IBBMM inferred nine clusters, whereas PyClone identified six clusters. Clusters 1, 2 and 6 identified by IBBMM showed a similar pattern of variation across the four samples. PyClone placed these 25 mutations together in cluster 1 (**Fig. 2a**,**b**). We propose that these mutations strictly co-occurred, with the observed variation in allelic prevalence due to differing mutational genotypes. Supporting this hypothesis, 75% (3 of 4) mutations placed in cluster 1 by IBBMM were predicted to be in regions of loss of heterozygosity, whereas 90% (18 of 20) mutations placed in cluster 2 by IBBMM were in regions predicted to be diploid heterozygous. The single mutation placed in cluster 6 by IBBMM fell in a region predicted to have major copy 3 and minor copy 1 (**Supplementary Table 1**). Major and minor copy refer to the number of copies of the most or least prevalent allele in the genotype, respectively.

The sum of allelic prevalences for IBBMM clusters 1 and 2 exceeded 1.0, which implies that cells exist with mutations from both clusters, with another group of cells containing only mutations from cluster 1. However, PyClone predicted that mutations from IBBMM cluster 1 and cluster 2 strictly co-occurred. To test these competing hypotheses, we performed single-cell sequencing of tissue from sample B, obtaining reliable measurements for 25 nuclei (**Fig. 2c**). When interpreting the data, we advise the reader that failure to detect a mutation in single-cell data can occur even if the mutation is present, owing to biased PCR amplification of alleles. This problem does not affect homozygous mutations (IBBMM cluster 1) but does affect heterozygous mutations (IBBMM clusters 2 and 9 and germline heterozygous positions). Thus, we may observe only partial representations of these clusters when they are present in a cell, motivating a collapsed representation of the data. To evaluate the clusters generated by IBBMM and PyClone, we collapsed the raw data to presence or absence of clusters predicted by IBBMM, wherein a cluster was determined to be present in a cell if any mutation from the cluster was present (**Fig. 2d**). IBBMM cluster 1 mutations were always detected with mutations from IBBMM cluster 2 (**Fig. 2c**,**d**). This is parsimonious with IBBMM clusters 1 and 2 comprising a single cluster, as predicted by PyClone. Results from a second HGSOC case from the same study led to similar conclusions (**Supplementary Results**, **Supplementary Fig. 14** and **Supplementary Table 2**).

As the practice of measuring allelic prevalences during the treatment cycle[15,16] or through retrospective analysis of multiple samplings increases[10,14,17], PyClone can contribute a robust statistical inference approach for studying selection patterns underpinning disease progression in cancer. PyClone is freely available for academic use at http://compbio.bccrc.ca/software/pyclone/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. European Genome-phenome Archive (EBI): HGSOC raw sequencing data are available under accession EGAS00001000547.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work is funded by Canadian Institutes for Health Research (CIHR), Genome Canada, Genome British Columbia, Canadian Cancer Society Research Institute and Canadian Breast Cancer Foundation grants to S.P.S. and S.A. S.P.S. is supported by the Michael Smith Foundation for Health Research and is the Canada Research Chair (CRC) for Computational Cancer Genomics. S.A. is the CRC for Molecular Oncology. A.R. is supported by a CIHR Banting scholarship.

AUTHOR CONTRIBUTIONS

Project conception and oversight: S.P.S., S.A., A.R.; method development: A.R., A.B.-C., S.P.S.; implementation and benchmarking: A.R.; manuscript writing and editing, study design and execution: A.R., A.B.C., S.P.S., S.A.; single-cell sequencing: J.K., D.Y., A.W., E.L., J.B.; data analysis and interpretation: G.H.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Nowell, P.C. *Science* **194**, 23–28 (1976).
2. Aparicio, S. & Caldas, C. *N. Engl. J. Med.* **368**, 842–851 (2013).
3. Greaves, M. & Maley, C.C. *Nature* **481**, 306–313 (2012).
4. Shah, S.P. *et al. Nature* **486**, 395–399 (2012).
5. Ding, L. *et al. Nature* **481**, 506–510 (2012).
6. Nik-Zainal, S. *et al. Cell* **149**, 994–1007 (2012).
7. Carter, S.L. *et al. Nat. Biotechnol.* **30**, 413–421 (2012).
8. Govindan, R. *et al. Cell* **150**, 1121–1134 (2012).
9. Shah, S.P. *et al. Nature* **461**, 809–813 (2009).
10. Gerlinger, M. *et al. N. Engl. J. Med.* **366**, 883–892 (2012).
11. The 1000 Genomes Project Consortium. *Nature* **467**, 1061–1073 (2010).
12. Harismendy, O. *et al. Genome Biol.* **12**, R124 (2011).
13. Rosenberg, A. & Hirschberg, J. in *Proc. 2007 Joint Conf. Empir. Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)* Vol. **410**, 420 (2007).
14. Bashashati, A. *et al. J. Pathol.* **231**, 21–34 (2013).
15. Forshew, T. *et al. Sci. Transl. Med.* **4**, 136ra68 (2012).
16. Dawson, S.J. *et al. N. Engl. J. Med.* **368**, 1199–1209 (2013).
17. Sottoriva, A. *et al. Proc. Natl. Acad. Sci. USA* **110**, 4009–4014 (2013).

ONLINE METHODS

PyClone model and implementation. The full description of the PyClone model and its derivatives used in the benchmarking experiments, methods used for synthetic data generation, methods for copy number–prior elicitation and implementation details are provided in the **Supplementary Note**.

Running PyClone and MCMC analysis. For the synthetic model comparison and normal mixing experiments, all analyses involving the PyClone genotype-aware models and IGMM, IBMM and IBBMM models were run for 10,000 MCMC iterations, with the first 1,000 samples discarded as 'burn-in', and no thinning was done. Burn-in refers to the initial samples from the MCMC chain that are discarded because the chain has not started sampling from the posterior distribution. For the synthetic investigation varying the number of mutations and HGSOC analyses, we used 100,000 iterations of the sampler, discarding the first 50,000 as burn-in. To generate cellular frequency plots, we fit Gaussian kernel density estimators to the post-burn-in MCMC trace using the SciPy 0.12.0 Python library. To assess convergence for HGSOCs, we ran three MCMC chains from random starting positions. The posterior densities of the cellular prevalence estimates were then inspected to ensure they were visually similar (**Supplementary Figs. 8** and **11**). For the synthetic data set and normal mixing experiment, this was not done because of the large number of runs. In general we have found that 10,000–100,000 iterations is sufficient for convergence in data sets with hundreds of mutations.

To cluster the data, we generated the pairwise posterior similarity matrix (the matrix indicating how frequently any two mutations appeared in the same cluster in the post-burn-in trace). We then hierarchically clustered the data using average linkage, and the resulting dendrogram was used as a guide to find a clustering that optimized the MPEAR criterion described in Fritsch and Ickstadt[18] (the code for doing this is built into PyClone). We note that there are other approaches, such as using the sample with the highest posterior probability to infer flat clusters using a Dirichlet process, but Fritsch and Ickstadt[18] have shown that these tend to perform poorly in comparison to the method we use. Because this step requires the formation of the posterior pairwise similarity matrix, the computational complexity scales as $O(N^2)$, meaning the computational complexity scales according to the square of the number (N) of mutations.

Evaluation and benchmarks. To assess the clustering performance of the methods, we computed the V-measure[13], calculated using the scikits-learn Python package 0.14.1. The V-measure is a measure of clustering accuracy between 0 and 1; a V-measure score of 1 represents perfect clustering. Cellular prevalence estimates were evaluated using the mean error over MCMC samples, where a mean error of 0 represents perfect cellular prevalence estimates. Explicitly, for each mutation the mean of the post-burn-in trace of the cellular prevalence parameter was used as a point estimate. The absolute difference between this value and the true value for each mutation in each data set was computed. For each of the data sets, the mean value of the absolute error across mutations was taken and used to generate the box plots. Statistical analysis was performed with the ANOVA (aov) and TukeyHSD functions in the R statistical computing package using RStudio v.0.96.331.

Alternative methods. To compare genotype-aware clustering to other clustering methods that ignore genotype, we implemented three standard clustering models in the PyClone software package: the infinite Gaussian mixture model (IGMM), the infinite binomial mixture model (IBMM) and the infinite beta-binomial mixture model (IBBMM). We interpreted the probability of success for the IBMM, and the mean parameter m for the IBBMM as the cellular prevalence of mutations. For the IGMM analysis we first computed the variant allelic prevalence for each mutation and then clustered these values, interpreting the mean of the clusters as the cellular prevalence. All MCMC analysis, clustering and cellular frequency inference was done as described earlier and in the **Supplementary Note**. All methods implemented in the PyClone software package, including the IGMM, IBMM and IBBMM, use the PyDP package to perform inference. Thus, any variation in performance should be due to differing distributional assumptions, not inference methods or implementation.

Normal-tissue mixture experiments. For the idealized mixture data presented in **Figure 1**, data from mixture experiments A (SRR385938), B (SRR385939), C (SRR385940) and D (SRR385941) from Harismendy *et al.*[12] were downloaded from the NCBI short-read archive. Each data set was generated by physically mixing DNA from four tissue samples from the 1000 Genomes Project in different proportions. Thus, mixtures were generated from the source DNA material and not *in silico*. The resulting mixtures were then subjected to targeted amplification using the UDT-Seq protocol and sequenced on the Illumina MiSeq platform.

FASTQ files were extracted from the downloaded .sra files using the "fastq-dump -split-files -clip" command from the NCBI SRA-SDK version 2.3.33. Sequences were aligned to the targeted genome using the "mem" command from the bwa 0.7.5a package. Count data were extracted from the BAM files using a custom Python script that filtered out positions with base or mapping qualities below 10. Because the primers used in the UDT-Seq protocol were designed to target mutational 'hot spots' in cancer and not the SNV positions we were analyzing, some candidate positions lay near the start or end of the primers. These positions tended to show significant strand bias, which could translate to biased allelic abundance estimates. To address this, we postprocessed the count data to remove positions showing significant strand bias ($P < 0.05$) as determined using a Fisher exact test. No multiple test correction was done.

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

Variants positions in the four cases used were previously identified in Ng *et al.*[19]. We compiled a list of positions with variants (i) in exactly one of the four cases used in the mixture, (ii) shared by NA18507 and NA19240, (iii) shared by NA18507, NA19240 and NA12878 and (iv) shared by all four cases. For SNVs present in multiple cases, we only considered positions that had the same genotype in all the variant cases. We manually removed positions that appeared to have variant allelic prevalence deviating significantly from the expected values. These outliers were either due to incorrect annotation of the genotype in one of the four cases

doi:10.1038/nmeth.2883

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 selected simulated an idealized cancer consisting of seven clonal cell populations that evolved according to a bifurcating tree. SNVs in all four cases represent the root of the tree. SNVs present in two or three cases represent interior nodes of the tree. SNVs unique to each sample represent leaf nodes in the tree.

Because the data set was highly imbalanced for variants with the BB genotype, we randomly downsampled the BB positions to obtain ten smaller data sets of 37 mutations with a 50:50 representation of positions with AB and BB genotypes for the nodes with SNVs in exactly one of the samples. There were not enough mutations in the interior nodes to do this, so we used all mutations for these nodes.

The predicted genotypes from Ng *et al.*[19] were used to determine the homologous copy number for the PCN prior 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 reference clustering (**Fig. 1c**). Reference clustering was defined by the case(s) harboring the variant genotype. To be explicit, we expected seven clusters to be identified: one for SNVs present in all cases; one for SNVs present in NA18507, NA19240 and NA12878; one for SNVs present in NA18507 and NA19240; and four clusters for SNVs unique to each case. Because we knew which SNVs were present in each case, we could compute a ground-truth clustering based on the above expectation. The challenge for the methods we considered was to correctly predict the co-occurrence of SNVs from the same ground-truth clusters with different genotypes. We would expect methods that ignore genotype to fail at this task.

We did not attempt to benchmark cellular prevalence estimates for this data set because the cellular prevalence values were only approximately correct owing to experimental variability in the mixing of the tissues.

Copy-number analysis. In the following section, "array" refers to Affymetrix SNP6.0 arrays. No other array platform was used for copy-number analysis.

Normalization and feature extraction. ASCAT and OncoSNP both require that the input data be suitably normalized and that the B allele fraction (BAF) and log R ratio (LRR) be extracted from the raw CEL files. To do this we used a modified version of the workflow described on the PennCNV-Affy website (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html). To extract features in the hg19 coordinate system, we mapped the supplied PennCNV .pfb and .gcmodel files using the annotations in the GenomeWideSNP_6,Full,na31,hg19,HB20110328.ugp file downloaded from the AROMA project (http://aroma-project.org/). We used only steps 1.2 and 1.4 of the PennCNV workflow to perform normalization and BAF/LRR extraction. The output of step 1.4 was passed to OncoSNP. For ASCAT we applied the PennCNV GC correction to the output of step 1.4 as ASCAT has no built-in GC normalization strategy.

ASCAT. We used ASCAT[20] version 2.1 for all analyses. We used the paired analysis mode to analyze the tumor and matched normal arrays jointly. The standard ASCAT workflow and default parameters described in the software manual were used for all analyses.

PICNIC. The latest version of PICNIC[21] as of 10 June 2013 was used for all analyses. For all analyses, PICNIC was run using only the tumor array. The parameter files supplied with PICNIC were mapped to hg19 coordinates using the GenomeWideSNP_6,Full,na31,hg19,HB20110328.ugp file downloaded from the AROMA project. Quantitative pathology estimates were used to inform the prior for normal contamination in the PICNIC inference procedure. The standard PICNIC workflow as described in the software manual with default parameters was performed for all analyses. As PICNIC performs normalization and feature extraction as part of its analysis, raw CEL files were passed as inputs.

OncoSNP. We used OncoSNP[22] version 1.3 for all analyses. We used the paired analysis mode to analyze the matched normal and tumor arrays jointly. We used the stromal contamination and intra tumor heterogeneity modes. We also included the X chromosome in the analysis. With these modifications, the OncoSNP workflow and default parameters described in the software manual were used for all analyses.

High-grade serous ovarian cancer. Ethical approval for molecular study of high-grade serous ovarian cancers was obtained from the University of British Columbia (UBC) Ethics Board application #H08-01411. Women undergoing debulking surgery (primary or recurrent) for carcinoma of ovarian/peritoneal/fallopian tube origin were approached for informed consent for the banking of tumor tissue. Cases of high-grade serous carcinoma in which more than one sample were collected in different anatomic locations (for example, different locations within the ovary, omentum) or in which material was available over different time periods (for example, at primary surgery and at recurrence) were chosen for this analysis. Multiple PyClone analyses were performed, one analysis per copy-number prediction method (**Supplementary Results**). We specified the priors for cellularity estimation in PICNIC on the basis of quantitative pathology estimates. Allelic count data were obtained from Bashashati *et al.*[14].

We obtained count data from additional mutations not validated as somatic from the authors. For each copy-number method we used the homologous copy-number information to elicit priors for PyClone using the PCN strategy, and we set the tumor content for the PyClone analysis to the value predicted by the copy-number method. MCMC analysis and postprocessing of the trace was done as discussed in the **Supplementary Note**.

Single-cell genotyping of frozen high-grade serous ovarian cancers. *Nuclei preparation and sorting.* Single cell nuclei were prepared using a sodium citrate lysis buffer containing Triton X-100 detergent. Solid tissue samples were first subjected to mechanical homogenization using a laboratory paddle-blender. The resulting cell lysates were passed twice through a 70-μm filter to remove larger cell debris. Aliquots of freshly prepared nuclei were visually inspected and enumerated using a dual–counting chamber hemocytometer (Improved Neubauer, Hausser Scientific) with Trypan blue stain. Single nuclei were flow sorted into

doi:10.1038/nmeth.2883

individual wells of microtiter plates using propidium iodide staining and a FACSAria II sorter (BD Biosciences).

Multiplex and single-plex PCRs. Somatic coding SNVs cataloged and validated in bulk-tissue genome sequencing experiments were picked for mutation-spanning PCR primers design using Primer3 (ref. 23). Common sequences were appended to the 5′ ends of the gene-specific primers to enable downstream barcoded adaptor attachment using a PCR approach. Multiplex (24) PCRs were performed using an ABI7900HT machine and SYBR GreenER qPCR Supermix reagent (Life Technologies). The 24-plex reaction products from each nucleus were used as input template to perform 48 single-plex PCRs using 48.48 Access Array IFCs according to the manufacturer's protocol (Fluidigm). Flow sorting plate wells without nuclei and 10-ng gDNA aliquots were used for negative and positive control reactions, respectively.

Nuclei-specific amplicon barcoding and nucleotide sequencing. Pooled single-plex PCR products from each nucleus were assigned unique molecular barcodes and adapted for MiSeq flow-cell NGS sequencing chemistry using a PCR step. Barcoded amplicon libraries were pooled and purified by conventional preparative agarose gel electrophoresis. Library quality and quantitation was performed using a 2100 Bioanalyzer with DNA 1000 chips (Agilent Technologies) and a Qubit 2.0 Fluorometer (Life Technologies). High-throughput DNA sequencing was conducted using a MiSeq sequencer according to the manufacturer's protocols (Illumina).

Bioinformatic analysis. Paired-end FASTQ files from the MiSeq sequencer were aligned to human genome build 37 downloaded from the NCBI using the "mem" command from the bwa[24] 0.7.5a package. Allelic count data were extracted from the BAM files using a custom Python script that filtered out positions with base or mapping qualities below 10. Any loci with fewer than 40 reads of coverage were deemed unusable and assigned an unknown status in plots. We removed any cell in which more than 80% of the loci were unusable. We also removed any loci that were unusable in more than 80% of cells.

Stochastic biased amplification of alleles due to limiting quantities of DNA in single cells made it difficult to detect presence or absence of the variant allele using allelic prevalence. To address this we applied a statistical test to determine the presence or absence of the variant allele. The null hypothesis for the test was that the variant allele was absent; thus, we observe reads with the variant allele only owing to sequencing error. We computed the proportion of reads with a variant allele at all positions on the amplicon targeting a loci, excluding the target loci. For these positions we defined the variant allele as the nonreference allele with the most reads supporting it. We used the mean of these values as the estimated sequencing error rate. This value was used to perform a one-tailed binomial exact test. For each cell we multiple test–corrected the *P* values for all loci with coverage using the Benjamini-Hochberg procedure. We used a false discovery rate of 0.001 to determine whether a variant allele was present. These methods have been applied in previous work, and additional details are reported therein[4,9].

18. Fritsch, A. & Ickstadt, K. *Bayesian Anal.* **4**, 367–392 (2009).
19. Ng, S.B. *et al. Nature* **461**, 272–276 (2009).
20. Van Loo, P. *et al. Proc. Natl. Acad. Sci. USA* **107**, 16910–16915 (2010).
21. Greenman, C.D. *et al. Biostatistics* **11**, 164–175 (2010).
22. Yau, C. *et al. Genome Biol.* **11**, R92 (2010).
23. Untergasser, A. *et al. Nucleic Acids Res.* **40**, e115 (2012).
24. Li, H. & Durbin, R. *Bioinformatics* **26**, 589–595 (2010).

doi:10.1038/nmeth.2883