Clonal genotype and population structure inference from single-cell tumor sequencing

Andrew Roth[1,2], Andrew McPherson[1,3], Emma Laks[1], Justina Biele[1], Damian Yap[1,4], Adrian Wan[1], Maia A Smith[1], Cydney B Nielsen[1,4], Jessica N McAlpine[5], Samuel Aparicio[1,4], Alexandre Bouchard-Côté[6] & Sohrab P Shah[1,4,7]

Single-cell DNA sequencing has great potential to reveal the clonal genotypes and population structure of human cancers. However, single-cell data suffer from missing values and biased allelic counts as well as false genotype measurements owing to the sequencing of multiple cells. We describe the Single Cell Genotyper (https://bitbucket.org/aroth85/scg), an open-source software based on a statistical model coupled with a mean-field variational inference method, which can be used to address these problems and robustly infer clonal genotypes.

© 2016 Nature America, Inc. All rights reserved.

Single-cell DNA sequencing has emerged as a promising approach for studying tumor evolution and dissecting clonal populations in human cancers[1–6]. A number of methods use bulk sequencing data and statistical deconvolution to study cancer cell populations[7–11], but single-cell sequencing offers the possibility of measuring clonal genotypes and prevalences directly. However, current single-cell sequencing protocols have been limited by high levels of experimental noise that result from missing data, allelic dropout (failure to measure both alleles at heterozygous loci, **Fig. 1a**,**b**), and doublets (unintended measurements of pairs of cells)[2,12,13], affecting the quality of genotype measurements.

Previous genotyping studies based on targeted sequencing have represented observed values of point-like events such as single-nucleotide variants (SNVs), small insertions or deletions (indels), and rearrangement breakpoints as binary variables that indicate the presence or absence of the variant allele. The Bernoulli mixture model (BMM) has been used to cluster mutations and predict genotypes, with the number of clusters selected using the Bayesian information criterion (BIC) score[2]. The BitPhylogeny model[14] makes use of a nonparametric Bayesian clustering approach that enforces phylogenetic constraints using an emission density similar to that of the BMM. These approaches ignore the zygosity of loci and have limited interpretability, as they model the hidden genotype states as continuous random variables. Thus, *ad hoc* postprocessing is required to derive discrete genotypes.

In this contribution, we systematically develop an interpretable probabilistic model, simultaneously addressing technical sources of noise in single-cell sequencing data and inferring a discrete set of genotypes present in a cell population and the genotype 'membership' of each cell (**Fig. 1c**,**d**). Using a cell-target matrix as input, our Single Cell Genotyper (SCG) model infers genotypes defined by point-like events (possibly from multiple data types such as SNVs and rearrangement breakpoints) with a discrete number of observable states. We introduce a novel mixture model or, more generally (to take doublets into account), a feature allocation model[15] that identifies groups of cells with shared genotypes and determines the genotype of each group. Our approach borrows statistical strength across measurements (cells and targets) to more accurately predict genotypes in the presence of missing data and allelic dropout. We use mean-field variational inference to efficiently estimate model parameters and the number of clonal populations (see **Supplementary Note 1** for full details of the model and see the inference algorithm and **Supplementary Discussion** for limitations of the model). Software to perform clonal genotype analysis using the SCG and Categorical mixture models is open source and freely available (**Supplementary Software** and https://bitbucket.org/aroth85/scg).

To assess performance, we generated 170 synthetic data sets using a range of parameters and used the data sets to compare the effectiveness of current cell clustering and clonal genotype prediction methods (**Supplementary Figs. 1–3** and **Supplementary Tables 1–10**). We compared the performance of doublet-naive variants of the SCG model with two (SCG2) and three states (SCG3), a doublet-aware variant of the SCG model (D-SCG3), the Categorical mixture model with two (CMM2) and three states (CMM3), BitPhylogeny[14], and two approaches based on hierarchical clustering (HC-VAF and HC-Discrete) (Online Methods). Note that the CMM2 model is the same as the BMM[2], but we use the current nomenclature to clarify the comparison. Our results suggest that the SCG family of models are more accurate than other approaches and can correct for doublets when present (see **Supplementary Results** for detailed discussion).

We applied the SCG model to previously published data from a patient with childhood acute lymphoblastic leukemia[2].

[1]Department of Molecular Oncology, BC Cancer Agency, Vancouver, British Columbia, Canada. [2]Graduate Bioinformatics Training Program, University of British Columbia, Vancouver, British Columbia, Canada. [3]School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada. [4]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [5]Department of Gynecology and Obstetrics, University of British Columbia, Vancouver, British Columbia, Canada. [6]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada. [7]Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. Correspondence should be addressed to S.P.S. (sshah@bccrc.ca) or A.R. (andrewjlroth@gmail.com).

RECEIVED 14 OCTOBER 2015; ACCEPTED 6 APRIL 2016; PUBLISHED ONLINE 16 MAY 2016; DOI:10.1038/NMETH.3867

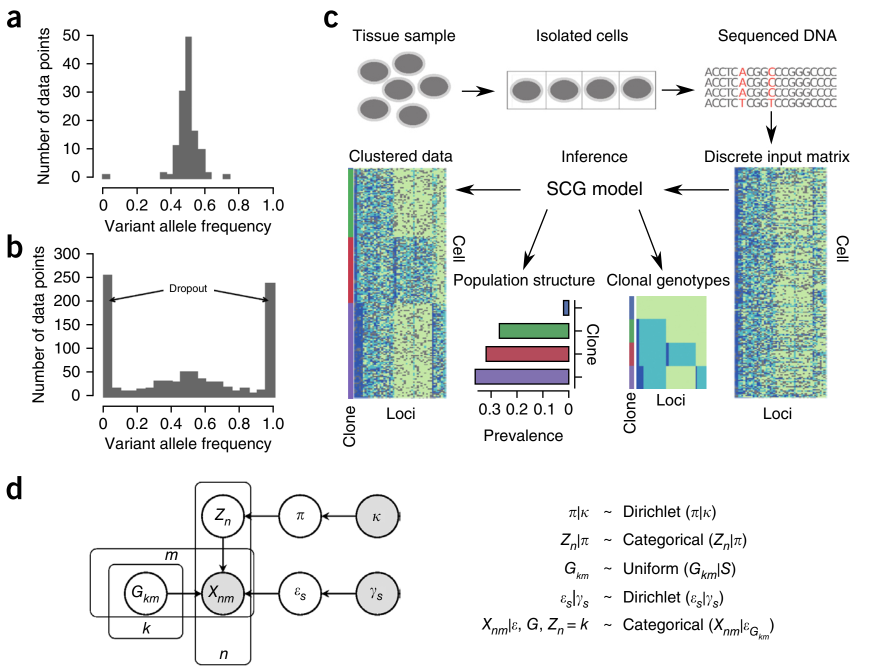


Figure 1 | Overview of the SCG model. (**a**,**b**) Histograms of variant allele frequencies of diploid heterozygous loci from (**a**) bulk sequencing and (**b**) single-cell sequencing of a 184-hTert cell line sample. The same set of loci are used for single-cell and bulk sequencing. (**c**) Schematic workflow of single-cell sequencing experiment. The SCG model is applied to the discrete data input matrix to cluster the data, predict clonal genotypes, and infer the prevalence of clones. (**d**) Probabilistic graphical model representing the basic SCG model. Shaded nodes represent observed values or fixed values; a posterior distribution over the values of the unshaded nodes is approximated using a variational Bayesian method. π, clone prevalence; Z_n, variable indicating clone of origin for cell n; G_{km}, variable indicating genotype of locus m for clone k; ε_s, error profile for genotype state s; X_{nm}, observed data from cell n and locus m; γ_s, parameter of Dirichlet distribution prior for the error profile for genotype state s; and k, parameter of Dirichlet distribution prior for the clone prevalence.

BMM (CMM2) was used in the original study to cluster single-cell data into clonal populations, generating clusters that appeared to be made up of disjoint sets of mutations. In two patients, a subset of data points associated with these clusters were assumed to be doublets and were manually removed from downstream analysis. We applied the CMM3, SCG3, and D-SCG3 models (**Fig. 2a–c** and **Supplementary Tables 11–14**) to data from one of these patients (patient 1, **Supplementary Table 11**) and plotted predicted genotypes associated with each cluster (**Fig. 2d–f**). As the CMM3 and SCG3 models are both doublet naive, we expected them to infer the spurious doublet clusters, whereas we expected the genotype-aware D-SCG3 model to identify them as doublets and assign them to their true pair of clonal clusters without the need for preprocessing.

The CMM-3 model infers five clusters, with cluster 3 appearing to be the result of doublet cells from clusters 0 and 1 (**Fig. 2d** and **Supplementary Table 15**). The SCG3 model infers six clusters, with cluster 2 appearing to be the result of doublets from clusters 1 (or 3) and 0 (or 5) (**Fig. 2f** and **Supplementary Table 16**). In contrast, the D-SCG3 model identifies only three clusters, none of which can be explained as a doublet mixture of the other clusters (**Fig. 2e** and **Supplementary Table 17**). Maximum-parsimony phylogenetic trees of predicted clonal genotypes show that, for both the CMM3 and SCG3 models, doublet clusters cause the inference of spurious clades relative to the D-SCG3 model (**Fig. 2g–i**).

To demonstrate the model's ability to facilitate the analysis of multiple data types and multiple samples, we applied D-SCG3 to data from a patient with high-grade serous ovarian cancer (HGSOC)[16]. We performed whole-genome sequencing of five synchronously obtained samples taken from three tumor masses: one from the patient's left ovary, one from the right ovary, and one from the omentum (**Supplementary Fig. 4**). We identified a

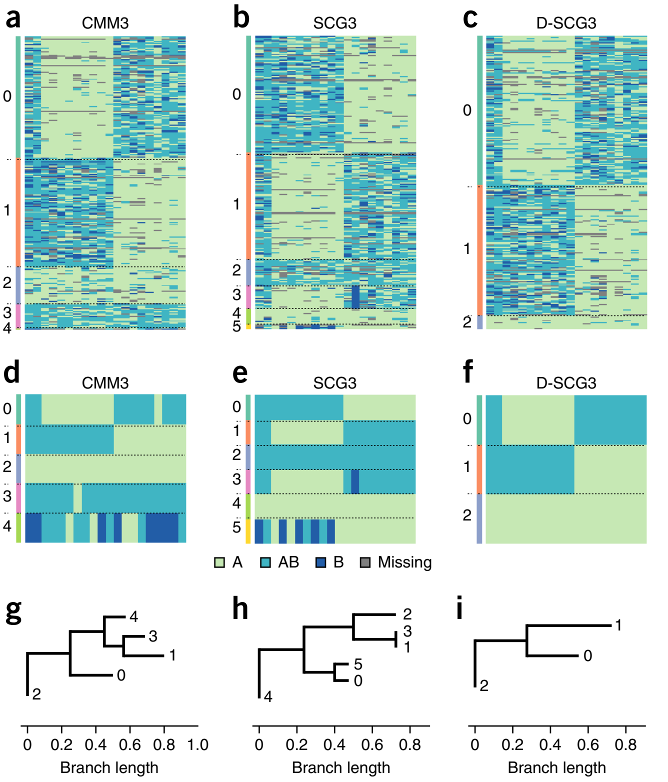
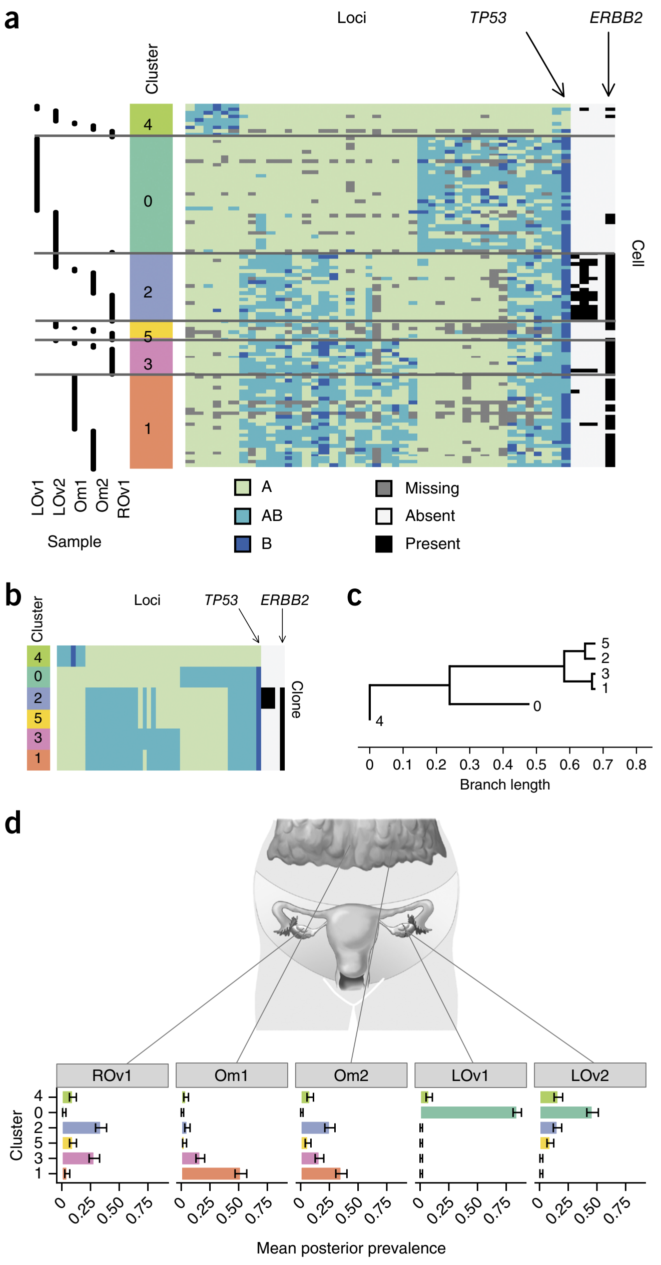


Figure 2 | Comparison of clustering performance on real data with doublets. SCG3, D-SCG3, and CMM3 models were used to identify clones in single-cell sequence data from a patient with childhood acute lymphoblastic leukemia. The same data were included in all plots, and doublet cells predicted by D-SCG3 were arbitrarily assigned to clusters. (**a–c**) Raw data ordered by cluster for (**a**) CMM3, (**b**) SCG3, and (**c**) D-SCG3 models. (**d–f**) Predicted genotypes for each cluster for (**d**) CMM3, (**e**) SCG3, and (**f**) D-SCG3 models. (**g–i**) Maximum-parsimony trees relating clonal genotypes for predicted genotypes from (**g**) CMM3, (**h**) SCG3, and (**i**) D-SCG3 models. Clusters are annotated to the left of each heat map and at each cladogram branch terminus. Branch length expressed as normalized proportion of change (i.e., fraction of locus change).

© 2016 Nature America, Inc. All rights reserved.

© 2016 Nature America, Inc. All rights reserved.

Figure 3 | The D-SCG3 model identified clonal cell populations in multiple samples from an HGSOC patient. The data set contained both SNV and breakpoint events. (**a**) Input data ordered by cluster (inner left bar). Originating samples for each cell are annotated on the far left. (**b**) Predicted clonal genotypes. (**c**) Maximum-parsimony tree relating clonal genotypes. (**d**) Estimated prevalence of clones across samples. Error bars indicate ±1 s.d. from the posterior mean, based on posterior distribution estimates from the model. Samples consisted of 84 cells each. By cluster, $n = 123$ (cluster 0), 102 (cluster 1), 75 (cluster 2), 37 (cluster 3), 35 (cluster 4), and 20 (cluster 5). LOv, left ovary; Om, omentum; ROv, right ovary.



high level of amplification of the *ERBB2* locus in four of the five samples. *ERBB2* amplification is known to be an early driver event in *HER2*-positive breast cancer, but it has not been associated with HGSOC. The absence of the *ERBB2* amplification in the left ovary sample 1 (LOv1) suggested that the amplification occurred after tumor initiation.

We designed PCR primers for a panel of clonally informative SNVs and control heterozygous single-nucleotide polymorphisms in a region of clonal loss of heterozygosity (LOH). We also incorporated several rearrangement breakpoints, including one associated with the *ERBB2* amplification. Using the D-SCG3 model, we jointly clustered the SNV and breakpoint events. Our analysis identified eight tumor clones and a normal population (cluster 4) of cells (**Fig. 3** and **Supplementary Tables 18–21**). We excluded three clusters (6, 7, and 8) that contained a large proportion of low-quality cells (mean ≥ 20% SNV events missing per cell, **Supplementary Fig. 5**). Only a subset of the predicted clones had a genotype that contained the *ERBB2* breakpoint (**Fig. 3a**,**b**; clusters 1, 2, 3, and 5). This suggested that the amplification was acquired after the tumor-initiating events of *TP53* mutation and LOH of chromosome 17. Furthermore, owing to the use of three zygosity genotype states, the D-SCG3 model determined that the *TP53* mutation was homozygous. Homozygous mutation of *TP53* is a hallmark of HGSOC[17] and therefore acts as an important diagnostic ground-truth marker for HGSOC.

The putative primary site for this tumor was in the left ovary. We identified multiple populations of cells with the amplification in one of the two samples (LOv2) taken from this site, whereas the other sample (LOv1) contained no cells with the amplification. In contrast, all the putative metastatic sites contained clones with the *ERBB2* amplification (**Fig. 3b–d**). This suggested that the late acquisition of the *ERBB2* amplification may have provided the clones with the ability to spread to other sites.

The SCG model advances the field of single-cell sequencing data analysis in several ways. It introduces a novel probabilistic approach to denoise properties of the input data such as allelic dropout, missing data, and doublet cells in a unified framework. Our approach simultaneously clusters cells into groups with shared genotypes and infers the genotype for each group. In all scenarios using simulated and real data, our model showed increased accuracy in clustering cells and inferring genotypes relative to other approaches. The SCG model also improves the biological interpretation of single-cell data. The ability to determine heterozygous and homozygous mutation state by genotype inference has implications for downstream haploinsufficiency analysis. Furthermore, we showed that the SCG model is capable of executing joint inference over SNVs and structural breakpoints. This allows for analysis of epistatic interactions between point mutations and genomic rearrangements—an area of analysis that remains unexplored in single-cell cancer genomics. Finally, we showed that SCG output allows for accurate phylogenetic tree inference using routine methods (whereas input from other methods leads to spurious results), thereby advancing a major goal of single-cell sequencing of cancer populations: the accurate reconstruction of population evolutionary histories.

METHODS

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

Accession codes. Sequence data for the 184-hTert cell line used in the synthetic benchmarks were deposited in the NCBI Short Read Archive (accession number SRP072568). Sequence data for the childhood leukemia data set were deposited in the Sequence Read Archive database (accession number SRP044380). Sequence data for the HGSOC data set were deposited in the

European Genome-phenome Archive (accession number EGAS00001000547).

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

ACKNOWLEDGMENTS

We acknowledge generous long-term funding support from the BC Cancer Foundation. In addition, the S.P.S. and S.A. groups receive operating funds from the Canadian Breast Cancer Foundation, the Canadian Cancer Society Research Institute (impact grant 701584 to S.A. and S.P.S.), the Terry Fox Research Institute (PPG program on forme fruste tumors), Canadian Institutes for Health Research (CIHR) (grant MOP-115170 to S.A. and S.P.S.), CIHR Foundation (grant FDN-143246 to S.P.S.), and a CIHR new investigator grant (MSH-261515 to J.N.M.). A.R. is supported by a Frederick Banting and Charles Best CIHR doctoral scholarship. S.P.S. and S.A. are supported by Canada Research Chairs. S.P.S. is a Michael Smith Foundation for Health Research scholar. We thank V. Earle for artwork depicting anatomic sites sampled in the study.

AUTHOR CONTRIBUTIONS

A.R., project conception, algorithm development, software implementation, and data analysis; S.A., A.M., E.L., J.B., D.Y., and A.W., single-nucleus sequencing; M.A.S. and C.B.N., data visualization; J.N.M., surgery, sample acquisition, and tumor banking; A.R., S.A., A.B.-C., and S.P.S., manuscript writing; S.P.S., project oversight and senior responsible author.

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. Navin, N. *et al. Nature* **472**, 90–94 (2011).
2. Gawad, C., Koh, W. & Quake, S.R. *Proc. Natl. Acad. Sci. USA* **111**, 17947–17952 (2014).
3. Wang, Y. *et al. Nature* **512**, 155–160 (2014).
4. Baslan, T. *et al. Genome Res.* **25**, 714–724 (2015).
5. Eirew, P. *et al. Nature* **518**, 422–426 (2015).
6. Navin, N.E. *Sci. Transl. Med.* **7**, 296fs29 (2015).
7. Roth, A. *et al. Nat. Methods* **11**, 396–398 (2014).
8. Jiao, W., Vembu, S., Deshwar, A.G., Stein, L. & Morris, Q. *BMC Bioinformatics* **15**, 35 (2014).
9. Zare, H. *et al. PLoS Comput. Biol.* **10**, e1003703 (2014).
10. Malikic, S., McPherson, A.W., Donmez, N. & Sahinalp, C.S. *Bioinformatics* **31**, 1349–1356 (2015).
11. Popic, V. *et al. Genome Biol.* **16**, 91 (2015).
12. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
13. Ning, L. *et al. Front. Oncol.* **4**, 7 (2014).
14. Yuan, K., Sakoparnig, T., Markowetz, F. & Beerenwinkel, N. *Genome Biol.* **16**, 36 (2015).
15. Broderick, T., Pitman, J. & Jordan, M.I. *Bayesian Anal.* **8**, 801–836 (2013).
16. McPherson, A. *et al. Nat. Genet.* http://dx.doi.org/10.1038/ng.3573 (2016).
17. Ahmed, A.A. *et al. J. Pathol.* **221**, 49–56 (2010).

© 2016 Nature America, Inc. All rights reserved.

ONLINE METHODS

SCG model. The SCG model uses a hierarchical Bayes approach to model a cell locus discrete input matrix derived from targeted single-cell data. The model jointly infers the genotypes present in the cell population and the assignment of cells to individual genotypes. This permits clustering of cells, estimation of clonal prevalences, and the specific co-occurrence and mutual exclusivity of mutations across the set of inferred genotypes. The model assumes each cell locus observation is generated by a Categorical distribution, permitting the specific modeling of wild-type, heterozygous, and homozygous mutations as well as the seamless integration of rearrangement breakpoints. Doublet cells are optionally modeled in order to avoid spurious inference in the scenario where more than two cells are erroneously sampled by the experimental platform. Inference is performed using a variational mean-field algorithm, which obviates the need to prespecify the number of clonal genotypes *a priori*. The full description of the SCG model and implementation details are provided in **Supplementary Note 1**.

Simulations. We simulated data according to a phylogenetic generative process. In this process, we first sample a clone phylogeny and clonal genotypes. We then populate the cell-locus matrix by simulating allelic count data from empirical distributions obtained from known diploid heterozygous positions sequenced from single 184-hTert breast epithelial cell line nuclei. The data are then discretized and a proportion of the values are stochastically perturbed as missing. A full description of the simulation methodology is provided in **Supplementary Note 2**. Parameters for simulations are in **Supplementary Table 1**. Code to generate simulated data is provided in the **Supplementary Software**.

Alternative methods. We compared the SCG model to hierarchical clustering; a Categorical mixture model that extends the Bernoulli mixture model to generalize multiple discrete states; and the BitPhylogeny model, which jointly infers clonal genotypes and phylogenies using a tree-structured stick-breaking prior over partitions of the data. A full description of the alternative methods and parameter settings is provided in **Supplementary Note 3**.

Performance metrics. When comparing performance in the absence of doublets, we use the V-measure metric[18] to assess clustering performance, and we use the mean hamming distance between predicted clonal genotype and true genotype averaged across cells to assess genotype accuracy. When comparing performance for data sets with doublets, we use the BCubed family of metrics extended to handle feature allocations[19] to assess clustering performance, and we use the maximum of the minimum hamming distance between predicted clones and true clones to assess genotype accuracy.

Cell-line sequencing. *Cell culture.* 184-hTert-L2 human mammary epithelial cells were cultured at 37 °C, 5% CO_2 in 500 mL of MEBM Mammary Epithelial Cell Growth Medium (Lonza) with 5 µg/mL transferrin (Sigma) and 2.5 µg/mL isoproterenol (Sigma), supplemented with Lonza MEGM Mammary Epithelial Cell Growth Medium Singlequots, excluding gentamicin. Cells were grown to near confluence, trypsinized, spun down, resuspended in cryopreservation media (50% media, 40% FBS, and 10% DMSO) and frozen at −1 °C/min to −80 °C in a Mr. Frosty Freezing Container. The parental 184-hTert cell line clones were generated by C. Barratt (Laboratory of Molecular Carcinogenesis at the National Institute of Environmental Health Sciences). The cell line was tested for mycoplasma with IDEXX Bioresearch h-IMPACT II human pathogen testing.

Nuclei preparation and sorting. Single nuclei were prepared from freshly thawed 184-hTert cells using Nuclei EZ lysis buffer (Sigma-Aldrich) and passed through a 70-µm filter. Aliquots of prepared nuclei were visually inspected and enumerated using a dual-counting-chamber hemocytometer (Improved Neubauer, Hausser Scientific) with trypan blue stain. Single nuclei were stained with propidium iodide and flow sorted into individual wells of microtiter plates using a FACSAria III sorter (BD Biosciences).

Multiplex and singleplex PCRs. Multiplex (48) PCRs were performed using an ABI 7900HT Fast Real-Time PCR System and SYBR GreenER qPCR SuperMix reagent (Life Technologies). The 48-plex reaction products from each nucleus were treated with ExoSAP-IT (Affymetrix) and used as input template to perform 48 singleplex PCRs using 48.48 Access Array IFCs according to the manufacturer's protocol (Fluidigm). Empty plate wells were used as negative controls and 10-ng gDNA aliquots were used for positive control reactions.

Nuclei-specific amplicon barcoding and nucleotide sequencing. Pooled singleplex PCR products from each nucleus were assigned unique Nextera molecular barcodes (Illumina) and adapted for MiSeq flow-cell NGS sequencing chemistry using a PCR step. Barcoded amplicon libraries were pooled and purified by electrophoresis on a 2% agarose gel and QIAquick Gel Extraction Kit (Qiagen). Library quality assessment and quantitation were performed using a 2100 Bioanalyzer (Agilent Technologies) and a Qubit 2.0 Fluorometer (Life Technologies). DNA sequencing was conducted using a MiSeq sequencer according to the manufacturer's protocols (Illumina).

Data sets. *Preprocessing.* All data for real data sets were preprocessed in the same way. Paired-end FASTQ files were aligned to the GRCh37 human reference genome downloaded from http://www.bcgsc.ca/downloads/genomes/9606/hg19/1000genomes/bwa_ind/genome/GRCh37-lite.fa using BWA version 0.7.12 with the aln and sampe commands using default parameters. Aligned BAM files were sorted with the sort command from sambamba version 0.5.4. Allelic counts from previously identified variant positions were extracted using a custom Python script. Reads with mapping quality < 30 or bases with base quality < 30 were not counted. In addition, we computed the background sequencing error rate by taking the mean proportion of nonreference allele counts from loci within 30 base pairs of either side of the target locus. Allelic count data were converted to a binary representation for use by the clustering algorithms using the binomial exact test described in ref. 20. We performed the test to check for the presence of each (A and B) allele at a P value threshold of 10^{-6}. If the test was negative for both alleles, we treated the data point as missing.

For breakpoint analysis we removed all reads less than 100 bp in length before alignment. We constructed a custom genome by creating chromosomes from the predicted rearrangement sequences. Alignments were performed using BWA as described above. We then counted the number of reads aligned to each rearrangement

© 2016 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.3867

using a Python script. We deemed a breakpoint as present in a cell if five or more reads supporting the breakpoint were found.

Childhood leukemia. We used previously published data[2]. Raw data were downloaded from the NCBI Short Read Archive in SRA format and converted to FASTQ format using the fastq-dump command from the SRA Toolkit version 2.4.3. The rest of the preprocessing was done as described above.

Preprocessed input data are in **Supplementary Table 11**. Clustering and genotype inference were performed using the SCG version 0.3.0 software. We clustered the data using the CMM3, D-SCG3, and SCG3 models. We set the value of κ to 1, used 40 clusters, and performed 1,000 random restarts for each method.

High-grade serous ovarian cancer. We used previously published data[16] available from the European Genome-phenome Archive under accession number EGAS00001000547. Preprocessed input data are in **Supplementary Table 18**. Clustering and genotype inference were performed using the SCG version 0.3.0 software. We clustered the data using the D-SCG3 model with sample-specific clone prevalences. We set the value of κ to 1, used 40 clusters, and performed 1,000 random restarts.

Code availability. Software to perform clonal genotype analysis using the SCG model and the Categorical mixture model is available under an open-source license as **Supplementary Software** and from https://bitbucket.org/aroth85/scg.

18. Rosenberg, A. & Hirschberg, J. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 410–420 (Association for Computational Linguistics, 2007).
19. Amigó, E., Gonzalo, J., Artiles, J. & Verdejo, F. *Inf. Retrieval* **12**, 461–486 (2009).
20. Shah, S.P. *et al. Nature* **461**, 809–813 (2009).

© 2016 Nature America, Inc. All rights reserved.