# Scalable whole-genome single-cell library preparation without preamplification

Hans Zahn[1,2,7], Adi Steif[1,3,7], Emma Laks[1,3], Peter Eirew[3,4], Michael VanInsberghe[1,2], Sohrab P Shah[3–5], Samuel Aparicio[3–5] & Carl L Hansen[2,6]



**Single-cell genomics is critical for understanding cellular heterogeneity in cancer, but existing library preparation methods are expensive, require sample preamplification and introduce coverage bias. Here we describe direct library preparation (DLP), a robust, scalable, and high-fidelity method that uses nanoliter-volume transposition reactions for single-cell whole-genome library preparation without preamplification. We examined 782 cells from cell lines and triple-negative breast xenograft tumors. Low-depth sequencing, compared with existing methods, revealed greater coverage uniformity and more reliable detection of copy-number alterations. Using phylogenetic analysis, we found minor xenograft subpopulations that were undetectable by bulk sequencing, as well as dynamic clonal expansion and diversification between passages. Merging single-cell genomes *in silico*, we generated 'bulk-equivalent' genomes with high depth and uniform coverage. Thus, low-depth sequencing of DLP libraries may provide an attractive replacement for conventional bulk sequencing methods, permitting analysis of copy number at the cell level and of other genomic variants at the population level.**

Tumors consist of heterogeneous cell populations that are subject to Darwinian evolution[1,2]. As tumor cells divide, branched evolutionary processes lead to the acquisition of a range of genetic variants, including copy-number alterations (CNAs), single-nucleotide variants (SNVs), and breakpoints[3]. Somatic genome variation can be used to define clones (cells related by descent from a unitary origin), which may feature different phenotypes[2]. Selective pressures such as drug treatment and metastatic progression differentially affect the survival of different clonal lineages[2]. Thus, enumerating the constituent populations of tumors and studying their dynamics may reveal genome variants associated with disease progression[2].

The majority of genomic studies examining the clonal architecture of tumors have used two approaches: (i) sequencing of DNA extracted in bulk from many cells[4–11] and (ii) sequencing of highly amplified single-cell DNA[11–20].

Bulk approaches derive DNA from a mixture of cells, sequence it at high coverage depth, and report variant allele prevalence (VAP) values[2]. Computational methods then attempt to disentangle the effects of contaminating normal cells, CNAs, and loss of heterozygosity (LOH) to predict clusters of co-occurring mutations[21]. Although such methods can identify major tumor subclones, their capacity to resolve minor populations is limited by sequencing error rates[22]. In addition, they are of limited utility when tumor cellularity is low, and they have difficulties in adequately addressing subclonal CNAs at low prevalence, which are more difficult to detect in bulk than low-prevalence subclonal SNVs[11,23].

Most existing single-cell approaches attempt to capture complete genomes through whole-genome amplification (WGA) before library construction[11–19]. Preamplification permits sequencing to higher coverage depth and breadth[18,24]; however, amplification biases decrease coverage uniformity, thereby obscuring the detection of CNAs, whereas polymerase errors lead to false SNV calls[13,25–27]. Among the WGA methods, degenerate oligonucleotide-primed PCR (DOP-PCR) achieves higher coverage uniformity than do multiple displacement amplification (MDA) and multiple annealing- and looping-based amplification cycles (MALBAC), thus making it most amenable to single-cell CNA inference[28]. However, DOP-PCR libraries are less suitable for SNV analysis because their coverage breadth saturates with deeper sequencing[18]. Moreover, sequencing a biologically meaningful number of cells to high depth and breadth[19] can be prohibitively expensive. Previous studies have examined anywhere from approximately ten to several hundred single-cell genomes[12,13,18–20], and some studies have relied on exome capture or targeted sequencing to reduce costs[11,14–17,29]. Thus, WGA methods are unlikely to provide an unbiased representation of a large heterogeneous population[30].

Here we present an alternative approach whereby indexed libraries are constructed directly from single-cell template DNA without any preamplification or sorting steps (**Fig. 1a**). Our DLP protocol involves direct tagmentation[31] of single-cell DNA in nanoliter

[1]Genome Science and Technology Graduate Program, University of British Columbia, Vancouver, British Columbia, Canada. [2]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada. [3]Department of Molecular Oncology, BC Cancer Agency, Vancouver, British Columbia, Canada. [4]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [5]Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [6]Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, Canada. [7]These authors contributed equally to this work. Correspondence should be addressed to C.L.H. (carl.lars.hansen@gmail.com), S.P.S. (sshah@bccrc.ca) or S.A. (saparicio@bccrc.ca).

RECEIVED 21 APRIL 2016; ACCEPTED 2 DECEMBER 2016; PUBLISHED ONLINE 9 JANUARY 2017; DOI:10.1038/NMETH.4140

volumes, followed by several cycles of PCR to add sequencing adaptors and index barcodes. Indexed libraries are pooled for multiplex sequencing at low depth, thus enabling inference of single-cell copy-number profiles and identification of clonal subpopulations (**Fig. 1b**). Sequencing reads from all cells may then be merged *in silico* to produce a high-depth 'bulk-equivalent' genome amenable to SNV, LOH, and breakpoint inference; alternatively, all cells within each copy-number clone can be merged to produce a set of high-depth 'clonal genomes' (**Fig. 1b**).

It should be emphasized that PCR after tagmentation has very different consequences from those of WGA (**Supplementary Fig. 1**). WGA methods generate many copies of each template strand in the form of long molecules that are only later fragmented into sequencing inserts. Hence, any given region in the original template is represented by multiple inserts with nonoverlapping start and end coordinates, which cannot be computationally filtered as duplicates. In contrast, DLP involves fragmentation of the original DNA template as the first step. Although several PCR cycles are used to incorporate index barcodes and sequencing adaptors, all copies are exact duplicates that can be identified and removed computationally, thus generating single-cell genomes in which any correctly mapped read is a unique representation of the original template. Here we demonstrate, in 268 cell-line and 514 breast xenograft cells, that the lack of preamplification results in more uniform single-cell genome coverage than that provided by DOP-PCR, and that merged single-cell libraries achieve a uniformity equivalent to that of bulk genomes at the same sequencing depth. The DLP approach thus overcomes some of the principal challenges of both bulk and WGA-based single-cell methods.

## RESULTS

### Direct single-cell library construction in nanoliter reaction volumes

Nanoliter reaction volumes allowed us to adapt conventional transposase-based library construction[31] to unamplified single-cell DNA (Online Methods). We designed and fabricated a 192-chamber microfluidic device that integrates the entire library preparation workflow (**Supplementary Figs. 2** and **3**). The device features inflatable reaction chambers and incorporated index primers, thereby permitting protocol customization in a compact architecture. Isolated cells are imaged on-chip, and single cells can thus be distinguished from doublets and contaminating debris. No-template control (NTC) chambers are included to verify the lack of contamination (**Supplementary Fig. 4**). Whereas only true single-cell libraries are included in subsequent single-cell CNA analyses, reads derived from other libraries are not wasted, because they are included in the merged bulk-equivalent genome. Libraries are indexed and pooled on-chip, and low-coverage multiplex sequencing is then performed (192 libraries per HiSeq lane; **Supplementary Table 1**).

### Coverage uniformity in near-diploid cells and tumor-cell sequencing metrics

To evaluate the coverage uniformity obtained by DLP, we first sequenced 192 indexed libraries from each of two immortalized near-diploid cell lines (184-hTERT-L2 (ref. 32) breast epithelial cell line ($n$ = 152 single cells, mean 0.07 ± 0.01× depth per cell, $n$ = 8 NTCs) and GM18507 (ref. 33) lymphoblastoid cell line ($n$ = 123 single cells, mean 0.12 ± 0.03× depth per cell, $n$ = 44 NTCs); **Supplementary Table 2**). We used a hidden Markov model (HMM) to infer copy-number profiles[34] and found that whereas most cells were diploid, both cell lines featured several subpopulations with shared integer CNAs, as well as a minority of cells with unique integer alterations (**Supplementary Figs. 5** and **6**). In addition, both data sets included subpopulations with a multitude of noninteger alterations across the genome; such alterations have been reported in previous studies and may be related to biological processes, such as DNA replication and apoptosis[35].

To examine how coverage breadth increases with the number of single-cell genomes merged, we carried out bootstrap sampling ($n$ = 30 draws per condition) and merging of single-cell libraries

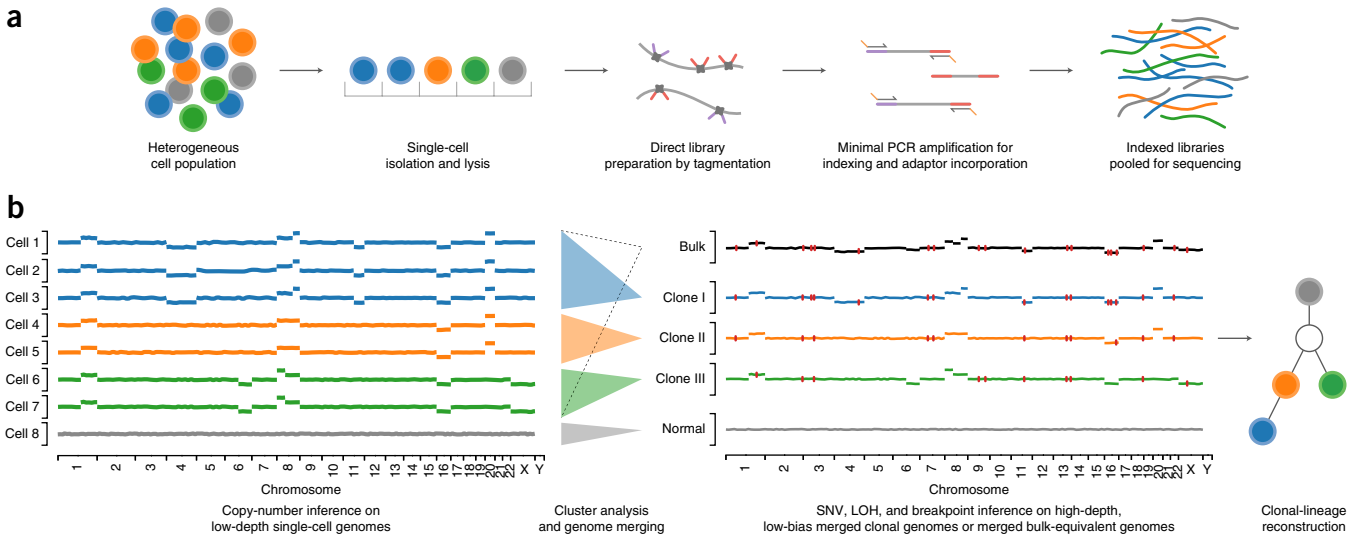


**Figure 1** | Single-cell genome analysis with DLP. (**a**) In the experimental workflow, single cells are isolated and lysed. Unamplified DNA is tagmented to produce unique sequencing inserts; index barcodes and sequencing adaptors are added by PCR; and indexed libraries are pooled for multiplex sequencing. (**b**) In the analytical workflow, a copy-number profile is inferred for each low-depth single-cell genome. Cells with shared profiles are merged *in silico* to produce clonal genomes, or all cells are merged to produce a bulk-equivalent genome. SNVs, LOH and breakpoints are inferred from the high-depth merged genomes, thus enabling reconstruction of the clonal lineage.

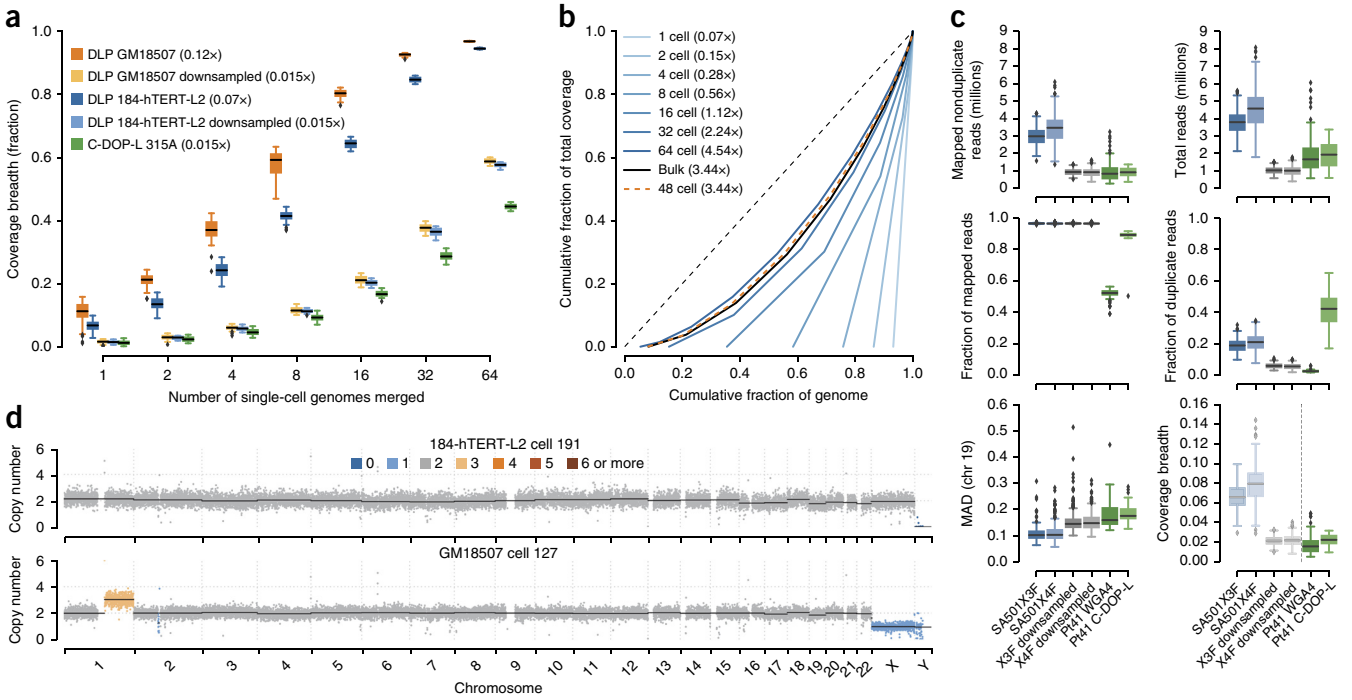


**Figure 2** | Coverage uniformity and sequencing metrics. (**a**) Effect of merging single-cell genomes (using bootstrap sampling, $n = 30$ draws per condition) on coverage breadth in near-diploid cell lines. DLP GM18507 ($n = 122$ cells); DLP 184-hTERT-L2 ($n = 146$ cells); C-DOP-L 315A[19] ($n = 95$ cells). Original and downsampled libraries are indicated. (**b**) Coverage uniformity for merged single-cell genomes. Blue curves, median merged 184-hTERT-L2 genomes from **a**. Bulk, 184-hTERT-L2 bulk genome. 48 cell, merged genome from 48 DLP 184-hTERT-L2 single cells. Dotted black line, perfectly uniform coverage. (**c**) Sequencing metrics for breast cancer cells sequenced with DLP (SA501X3F, $n = 295$ cells; SA501X4F, $n = 297$ cells), WGA4 ($n = 74$ cells), and C-DOP-L[19] ($n = 64$ cells). Xenograft cells (X3F and X4F) were downsampled to the same median number of mapped nonduplicate reads as in the Pt41 data sets. Dotted line (at bottom right) indicates that coverage breadth should not be compared across different tumors. For all box plots, the middle line represents the median, the box limits indicate the quartiles, the whiskers show the rest of the distribution, and the diamonds are outliers. (**d**) Copy-number profile for the 184-hTERT-L2 cell with median breadth from the bootstrap analysis (top), and for a GM18507 cell with a unique integer gain in chr 1 (bottom). Colors correspond to integer HMM copy-number states[34]; black lines indicate segment medians.

from our cell lines (**Fig. 2a**). Merging the genomes of 64 DLP single cells resulted in a median 94.5% (184-hTERT-L2) and 96.8% (GM18507) coverage breadth. We compared our results with those for cells from an immortalized lymphoblastoid line (315A) sequenced with the C-DOP-L protocol[19], a variant of DOP-PCR, which was the only published data set that we identified with a comparable number of near-diploid cells sequenced in multiplex[19] (**Fig. 2a**). Despite pooling of half as many libraries per HiSeq lane (96 versus 192), cells from the C-DOP-L data set had a mean depth of 0.015× (relative to our mean depth of approximately 0.07× per cell). To provide a fair comparison, we trimmed our reads by 44.8%, aligned our data as single-end reads, and downsampled each single-cell genome (184-hTERT-L2 by 48%; GM18507 by 66%) to achieve the same mean depth per cell as that in the 315A data set (**Fig. 2a**). In the trimmed and downsampled data sets, 64 DLP 184-hTERT-L2 and GM18507 cells achieved a median of 57.7% and 58.8% coverage breadth, respectively. In contrast, 64 C-DOP-L 315A cells had a median breadth of 44.5%. These results demonstrate that DLP yields higher-depth cells with a more uniform distribution of reads across the genome.

To determine how coverage uniformity for our merged genomes compared with that for a bulk genome, we plotted one Lorenz curve for each condition in our bootstrap analysis, using the merged 184-hTERT-L2 genome with the median breadth for that condition (**Fig. 2b**). We sequenced a bulk genome for the same 184-hTERT-L2 passage at 3.44×, using the standard Nextera library preparation protocol, and found that a merged genome corresponding to 48 DLP 184-hTERT-L2 cells with the same depth achieved equivalent breadth and uniformity (**Fig. 2b**). We also generated Lorenz curves for the GM18507 and 315A data sets (**Supplementary Fig. 7**).

We next examined two passages of a patient-derived primary triple-negative breast cancer (TNBC) xenograft (patient identifier SA501, 768 indexed libraries). Among these libraries, 296 single cells were from a third-passage xenograft (SA501X3F, mean 0.07 ± 0.01× depth per cell, 32 NTCs), and 299 single cells were from a fourth-passage xenograft derived from the third passage (SA501X4F, mean 0.087 ± 0.02× depth per cell, 22 NTCs). We compared sequencing metrics between these tumor cells prepared with DLP and those from a published WGA data set with similar ploidy (Pt41 (ref. 19), ER-positive breast cancer) sequenced with two different DOP-PCR protocols: WGA4 (refs. 12,19) and C-DOP-L[19] (**Fig. 2c**; metrics in **Supplementary Table 3**; results of Kruskal–Wallis (KW) tests in **Supplementary Table 4**). To yield the same median number of mapped nonduplicate reads per cell as that in the Pt41 libraries (KW $P = 0.11$), we downsampled the DLP libraries by 73% (SA501X3F) and 78% (SA501X4F).

WGA4 libraries suffered from low mappability resulting from WGA adaptor contamination, whereas C-DOP-L libraries had high duplicate rates (**Fig. 2c**), thereby requiring nearly twice the total reads to achieve the same number of mapped nonduplicate reads as the DLP libraries. The heavily downsampled DLP cells,

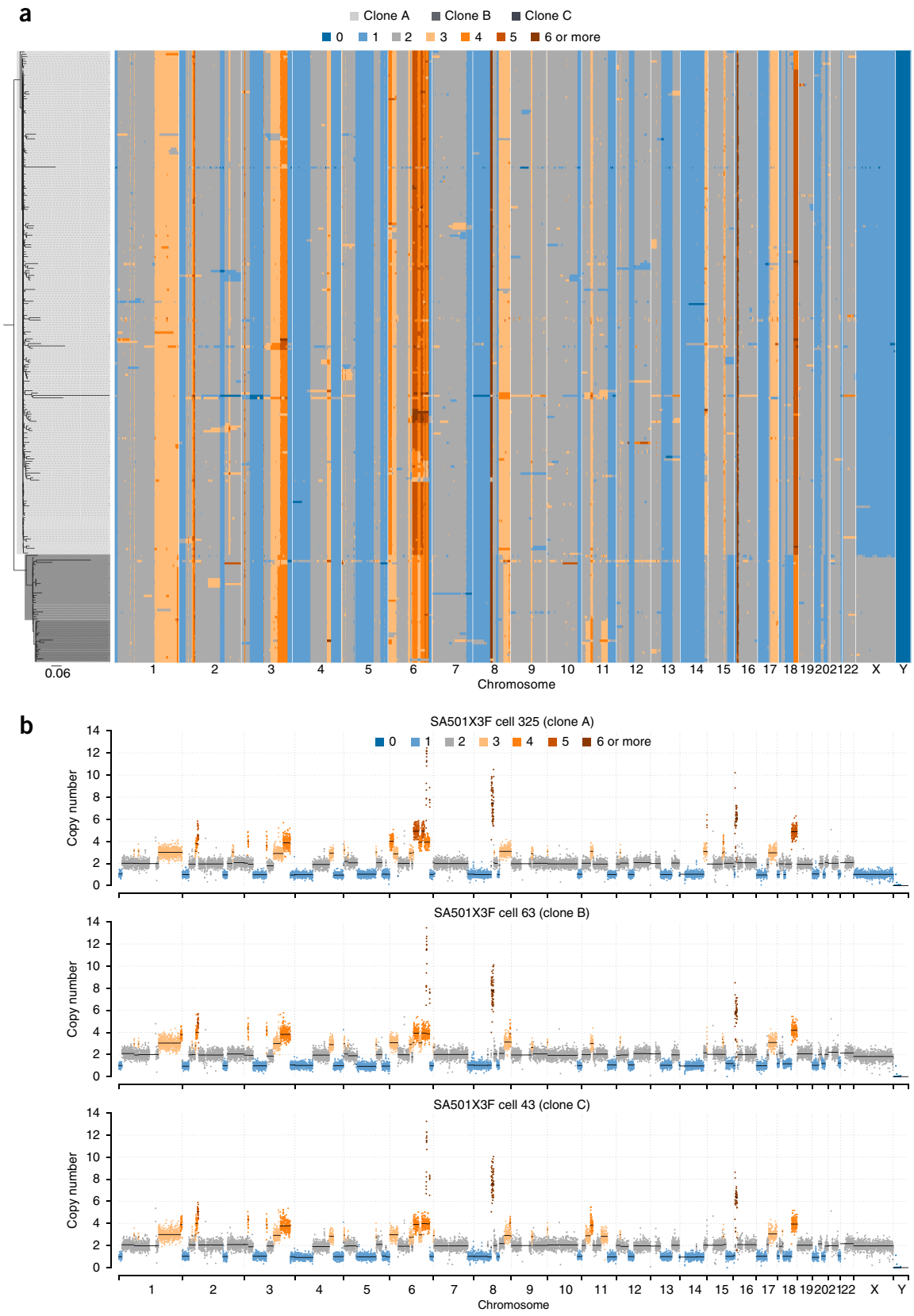


**Figure 3** | Single-cell copy-number profiles from xenograft SA501X3F. (**a**) Integer copy-number heat map. Rows correspond to cells ($n$ = 260), and columns correspond to 150-kb genomic bins. Colors correspond to integer HMM copy-number states[34], and a Bayesian phylogeny[36] with superimposed clones is shown at left. (**b**) Representative single-cell copy-number profiles from clones A ($n$ = 214) cells, B ($n$ = 28 cells), and C ($n$ = 18 cells). Colors correspond to integer HMM copy-number states[34]; black lines indicate segment medians.

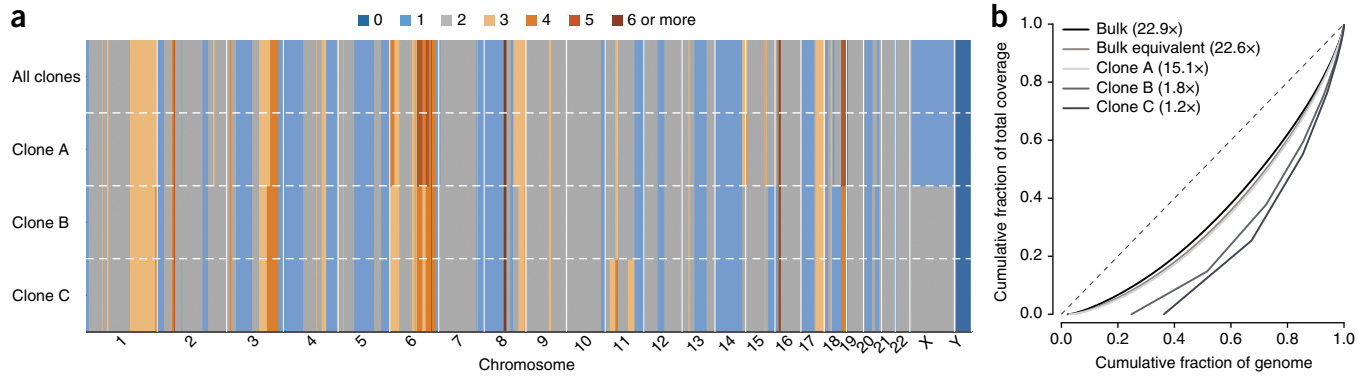


**Figure 4** | Analysis of merged clonal genomes for xenograft SA501X3F. (**a**) Inferred copy-number profiles for the merged genome from all single cells (top) and from all cells in clone A (82.3% of cells), clone B (10.8% of cells), or clone C (6.9% of cells). (**b**) Lorenz curves for the three clonal genomes, the merged bulk-equivalent genome, and a standard bulk genome.

compared with DOP-PCR cells, had significantly lower median absolute deviation (MAD) values in binned read counts for chromosome (chr) 19, the only chromosome that was diploid in all samples, (KW $P < 2.2 \times 10^{-16}$). Although coverage breadth cannot be fairly compared across different tumors, the C-DOP-L protocol achieved higher median breadth than did the WGA4 protocol when applied to the same tumor (**Fig. 2c**, KW $P = 1.89 \times 10^{-5}$), suggesting that WGA4 would not outperform DLP in a bootstrap merging analysis. A DLP single-cell copy-number-profile example from each of the immortalized normal human cell lines is shown in **Figure 2d**.

### Copy-number heterogeneity and clonal evolution in serial breast cancer xenograft passages

We next sought to examine copy-number heterogeneity and to identify subpopulations with shared profiles in the xenograft cells. We inferred integer copy-number states[34] for 260 cells from SA501X3F (**Fig. 3a**). Clonal lineage reconstruction with a Bayesian phylogenetic model[36] revealed three major subpopulations: a dominant clone with one copy of chr X (clone A, $n = 214$ cells), a minor population with two copies of chr X and numerous smaller alterations relative to the dominant population (clone B, $n = 28$ cells; chr 1, 2, 3, 5, 6, 8, 14, 15, 18, and 20), and a third subpopulation of clone B featuring additional alterations in chr 11 (clone C, $n = 18$ cells) (representative single-cell profiles from each clone in **Fig. 3b**). The high fidelity of the DLP approach also revealed many unique alterations, demonstrating rich diversity within each subpopulation. The lack of contamination in our NTCs and the clear placement of segment medians along integer values suggested that many of these unique events represent genuine diversity in this highly rearranged tumor (**Supplementary Fig. 8**).

Analysis of the subsequent xenograft passage SA501X4F ($n = 254$ cells) revealed that minor clones B and C were no longer detectable, and the descendants of clone A dominated the population. Furthermore, the population had diversified and featured numerous small subclones with distinct integer CNAs shared by at least three cells (**Supplementary Figs. 9** and **10**). These included a subclone with a set of amplifications and deletions in chr 12 that was evident in the previous passage ($n = 2$ cells in SA501X3F, $n = 8$ cells in SA501X4F; population v), as well as several subpopulations not evident in the third passage, such as a group of cells that had lost the ancestral high-level amplification in chr 16 ($n = 20$ cells; population ix). We also computationally identified a single mouse cell and five xenograft libraries with mouse debris (**Supplementary Fig. 11**), all of which were excluded from the merged genome analysis.

### Merging of single-cell genomes yields high-depth, low-bias clonal and bulk-equivalent genomes

A distinguishing advantage of our method is its ability to generate high-depth clonal or bulk-equivalent genomes with the same uniformity as that of a standard bulk genome. We first generated a merged genome for each clone identified in SA501X3F, as well as a genome for all populations. Next, we inferred CNA profiles (**Fig. 4a**). Despite numerous differences in copy number, we found little evidence of minor clones B and C in the combined profile (**Supplementary Fig. 12a**). Next, we added filtered single cells, multicell libraries, and libraries with contaminating debris (excluding those with mouse contamination) to generate a bulk-equivalent genome. Once again, the bulk-equivalent genome featured coverage uniformity comparable to that of a bulk genome at the same sequencing depth (**Fig. 4b**). In contrast to results from a recent study of CNA inference in WGA single-cell data sets, in which known germline variants <5 Mb could not be reliably detected[37], we compared deletion and amplification calls in DLP single cells with their clonal genome profiles and were consistently able to infer segments in the range of 1–5 Mb (**Supplementary Fig. 12b**). Comparing CNAs in our highest-depth cells with known bulk CNAs, we decreased our bin size and were able to detect even smaller segments (100–500 kb; **Supplementary Fig. 13**), thus producing what is, to our knowledge, the best reported sensitivity for low-depth single-cell CNA inference.

Finally, we applied conventional SNV, LOH, and breakpoint inference to the SA501X3F bulk-equivalent and bulk genomes (**Supplementary Table 5**). VAP correlations were high between the two samples, both for high-probability SNV calls[38] (**Fig. 5a**) and for a set of SNVs previously validated through targeted sequencing in the SA501 xenograft series[11] (**Fig. 5b**). The output of an HMM for simultaneous inference of copy number and LOH[23] demonstrated that, in both genomes, evidence of the minor subpopulations with two copies of chr X was apparent as an inward shift in allele ratios (**Supplementary Fig. 14**). However, other alterations distinguishing these minor subpopulations were not evident, thus underscoring the need for a single-cell approach.

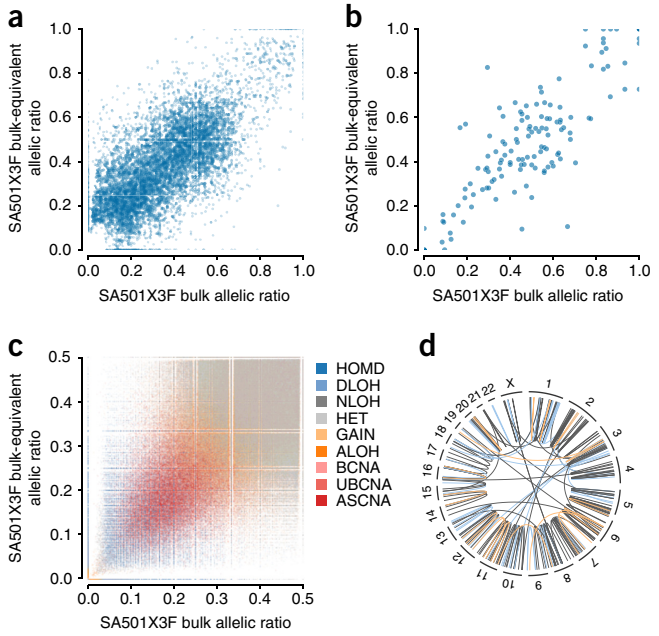


**Figure 5** | Analysis of SNVs, LOH, and breakpoints for xenograft SA501X3F. (**a**) Correlation of allelic prevalence values for high-confidence SNVs[38] between the DLP bulk-equivalent and bulk genomes (Pearson's correlation $\rho = 0.84$, $n = 8{,}745$ SNVs). (**b**) Correlation of allelic prevalence values between the bulk-equivalent and bulk genomes for previously validated SNVs[11] ($\rho = 0.93$, $n = 184$ SNVs). (**c**) Correlation of LOH-state calls for heterozygous germline variants between the bulk-equivalent and bulk genomes[23] ($\rho = 0.92$, $n = 1{,}516{,}294$ SNVs). HOMD, homozygous deletion; DLOH, deletion LOH; NLOH, copy-neutral LOH; HET, heterozygous diploid; GAIN, one-copy gain; ALOH, amplified LOH; BCNA, balanced CNA; UBCNA, unbalanced CNA; ASCNA, allele-specific CNA. (**d**) Rearrangement breakpoint calls[39] in the bulk-equivalent and bulk genomes, showing overlapping calls (gray, $n = 133$ breakpoints), calls with high probability only in the bulk-equivalent genome (orange, $n = 18$ breakpoints), and calls with high probability only in the bulk genome (blue, $n = 44$ breakpoints).

The overlap in LOH-state calls and correlation of VAPs for heterozygous germline variants in these two samples indicated that LOH can be reliably inferred from the bulk-equivalent genome (**Fig. 5c**). Finally, individual breakpoints inferred in the bulk-equivalent and bulk genomes show high overlap[39] (**Fig. 5d**). These analyses suggest that single-cell library construction without preamplification and subsequent *in silico* genome merging generate a bulk-equivalent genome to which standard variant inference methods can be applied.

## DISCUSSION

We developed a single-cell library preparation method without preamplification for the simultaneous acquisition of high-resolution single-cell copy-number profiles and bulk tumor genomes. In comparison with existing methods, DLP permits more cells to be multiplexed per lane and generates genomes with more uniform coverage. Merging cells *in silico* produces a bulk-equivalent genome with uniformity equal to that of a standard bulk genome. Analysis of 514 single cells from two TNBC xenograft tumors revealed many copy-number subpopulations featuring integer alterations shared by multiple cells, in addition to numerous unique events. Finally, we demonstrated that SNVs, LOH, and breakpoints can be reliably inferred in merged bulk-equivalent genomes.

Adey *et al.* previously introduced Tn5 transposition chemistry for bulk library preparation and have described testing this chemistry on as little as 10 pg of bulk DNA[31]. Here, we applied this Tn5 chemistry in an optimized nanoliter-volume process to enable, what is, to our knowledge, the first direct construction of single-cell libraries through tagmentation without preamplification. We implemented this approach using a novel microfluidic chip design, which features prespotted index primers, fluorescence imaging of trapped cells and inflatable reaction chambers that permit protocol customization. Prespotting of index primers during device fabrication enables pooling of all reaction products on-chip, thereby decreasing the device complexity. Compared with existing commercial devices, which use a series of chambers with fixed volumes, the inflatable chamber design permits an arbitrary number of reagent additions and volumes, decreases the risk of losing intermediate products during transfer steps, and allows implementation of a range of genomic workflows that would benefit from the use of nanoliter volumes.

Current DOP-PCR WGA protocols for single-cell library preparation take approximately 3 d to complete[35], are typically performed on 96 or fewer samples per experiment[12,18,35,40–42], cost an estimated $15 per cell for amplification (ignoring subsequent library construction in large volumes), and suffer from low mappability or high duplicate rates. The DLP protocol offers substantial gains in throughput in terms of cost (approximately $0.50 per cell; **Supplementary Table 6**), labor (192 libraries in 2.5 h of hands-on time), and sequencing effort (192 libraries per HiSeq lane). We believe that the DLP protocol would also be adaptable to other small-volume formats, such as microdroplets[43–46].

We emphasize that DLP is not meant to capture complete single-cell genomes but is instead meant to provide high-resolution single-cell copy-number profiles while producing high-quality bulk genomes in a single sequencing experiment. DLP also permits identification and exclusion of contaminating normal cells from merged bulk-equivalent genomes to rescue low-cellularity samples. Low tumor cellularity hinders variant calling and presents a major obstacle to the analysis of valuable patient tissues[11].

We posit that in most cases, sequencing many DLP cells to shallow depth is superior to sequencing a few WGA cells to high depth, because this approach provides a better representation of copy-number diversity in the tumor population. For example, sequencing the full genomes of ten cells with WGA methods to 30× depth would be equivalent to sequencing 6,000 single cells (at 0.05×) with DLP. The latter approach provides a much more comprehensive view of copy-number diversity in the population, with a detection sensitivity for low-prevalence subclones of approximately 0.05% (3/6,000 cells). Another distinct advantage of DLP is that when single-cell genomes are merged, information about which cell each read originated from is preserved. Future computational methods may exploit this property to infer SNVs, LOH, and breakpoints in merged DLP genomes with improved power, and to comprehensively characterize differences in somatic genome variation between copy-number subpopulations at lower sequencing depth. We envision that direct single-cell library preparation may become a new standard approach to the sequencing of heterogeneous populations.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

**ACKNOWLEDGMENTS**

We gratefully acknowledge funding support from the BC Cancer Foundation, the Canadian Breast Cancer Foundation, Genome Canada/Genome BC, the Natural Sciences & Engineering Research Council of Canada (grant RGPIN 386152-10 to C.L.H.), the Terry Fox Research Institute (grant NFP 1021 to S.A. and S.P.S.), the Canadian Institutes of Health Research (grant MOP 126119 to S.A. and S.P.S.), and the Canadian Cancer Society Research Institute (grant 701584 to S.A. and S.P.S.). S.A. and S.P.S. are supported as Canada Research Chairs, and S.P.S. is supported as a Michael Smith Foundation for Health Research Scholar. H.Z. and A.S. are each supported by a Vanier Canada Graduate Scholarship.

**AUTHOR CONTRIBUTIONS**

H.Z., A.S., S.P.S., S.A., and C.L.H. designed the research. H.Z. performed experiments. A.S. analyzed the data. A.S., H.Z., C.L.H., S.A., and S.P.S. wrote the paper. E.L. prepared tissue samples and bulk libraries. P.E. performed xenograft transplants. M.V. contributed to technology development. C.L.H., S.A., and S.P.S. supervised the research.

**COMPETING FINANCIAL INTERESTS**

The authors declare competing financial interests: details are available in the online version of the paper.



1. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
2. Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368**, 842–851 (2013).
3. Burrell, R.A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
4. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
5. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
6. Campbell, P.J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
7. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
8. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
9. Landau, D.A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
10. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
11. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).
12. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
13. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
14. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
15. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. USA* **110**, 21083–21088 (2013).
16. Gawad, C., Koh, W. & Quake, S.R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci. USA* **111**, 17947–17952 (2014).
17. Lohr, J.G. *et al.* Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.* **32**, 479–484 (2014).
18. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
19. Baslan, T. *et al.* Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Res.* **25**, 714–724 (2015).
20. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
21. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
22. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
23. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
24. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
25. Wang, J., Fan, H.C., Behr, B. & Quake, S.R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**, 402–412 (2012).
26. de Bourcy, C.F. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* **9**, e105585 (2014).
27. Macaulay, I.C. & Voet, T. Single cell genomics: advances and future perspectives. *PLoS Genet.* **10**, e1004126 (2014).
28. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).
29. Leung, M.L. *et al.* Highly multiplexed targeted DNA sequencing from single nuclei. *Nat. Protoc.* **11**, 214–235 (2016).
30. van den Bos, H. *et al.* Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimers disease neurons. *Genome Biol.* **17**, 116 (2016).
31. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
32. Burleigh, A. *et al.* A co-culture genome-wide RNAi screen with mammary epithelial cells reveals transmembrane signals required for growth and differentiation. *Breast Cancer Res.* **17**, 4 (2015).
33. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
34. Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* **22**, 1995–2007 (2012).
35. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).
36. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
37. Knouse, K.A., Wu, J. & Amon, A. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Res.* **26**, 376–384 (2016).
38. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).
39. McPherson, A. *et al.* nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res.* **22**, 2250–2261 (2012).
40. McConnell, M.J. *et al.* Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
41. Cai, X. *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.* **8**, 1280–1289 (2014).
42. Knouse, K.A., Wu, J., Whittaker, C.A. & Amon, A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc. Natl. Acad. Sci. USA* **111**, 13409–13414 (2014).
43. Mazutis, L. *et al.* Multi-step microfluidic droplet processing: kinetic analysis of an in vitro translated enzyme. *Lab Chip* **9**, 2902–2908 (2009).
44. Mazutis, L. *et al.* Single-cell analysis and sorting using droplet-based microfluidics. *Nat. Protoc.* **8**, 870–891 (2013).
45. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
46. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).

## ONLINE METHODS

**Microfluidic-device design.** We designed and manufactured a microfluidic device that integrates the entire library preparation workflow for multiplex single-cell whole-genome sequencing, including cell isolation, imaging, lysis, DNA tagmentation, barcoding, and sequencing-adaptor incorporation. The device features 192 single-cell processing units arrayed in four columns of 48 and has four separate cell-loading inlets to enable case–control studies (**Supplementary Fig. 2a**). The last cell-processing unit of each column lacks a cell trap, thus ensuring that each column contains at least one NTC reaction, which is used to assess contamination in the cell-suspension fluid. The core of each cell-processing unit (**Supplementary Fig. 2b**) is an inflatable reaction chamber (III), which is connected to a cell trap (II) optimized for single-cell capture, a reagent inlet (IV), and a chamber containing prespotted index primers (V). We designed the inflatable chamber (**Supplementary Fig. 2c**) such that the volume can be freely adjusted, and a nearly arbitrary sequence of reagent additions can be implemented in a single device architecture. The design consists of a reaction chamber aligned to a displacement chamber, with a thin elastomeric membrane separating the two (**Supplementary Fig. 2c**). This configuration allows the thin membrane to be deflected, thus decreasing or increasing the volume of the reaction chamber. The deflection can be controlled by changing the pressure gradient across the membrane or by adjusting the volume in the reaction chamber. To minimize the number of required world-to-chip connections and to simplify channel routing, contactless spotting technology (sciFLEXARRAYER S3, Scienion) was used to dispense picoliter volumes of unique molecular-index sequences directly into a specialized open chamber in each cell-processing unit (**Supplementary Fig. 2d**,**e**). These index barcodes are sealed in the device as fabrication is completed. During an experiment, reactions are assembled in parallel by metering out precisely defined volumes through the integrated peristaltic pumps. After a reagent addition, the mutual bus channel (VI) (**Supplementary Fig. 2b**) can be flushed and replaced with a new reagent. All reaction chambers are furthermore connected in series to enable pooled recovery of reaction products for downstream analysis. Interlayer connections (vias) are used to facilitate three-dimensional fluid routing to further decrease device complexity and maintain scalability[47]. The AutoCAD design file is provided as **Supplementary Data**, and the device fabrication process and details of the device operation procedure are provided in the **Supplementary Note**.

**On-chip direct library preparation.** We modified the Nextera protocol (Illumina) for rapidly generating genomic libraries for next-generation sequencing in a 'one-pot' nanoliter reaction. After priming, cell suspensions were injected into each of the four cell-loading inlets and pushed through the cell-sorting channels. Cells were sequentially caught in cell traps and washed with PBS to remove extracellular DNA, cell debris, and untrapped cells. Integrated valves above and below each cell trap were closed to isolate trapped cells into individual cell-processing units, and high-magnification fluorescence imaging with a DNA stain was used to identify single cells and to flag chambers with multiple cells and those with contaminating debris (**Supplementary Fig. 2f**–**h**). Cell calls were recorded for downstream use, and only libraries identified as true single cells were included in subsequent single-cell copy-number analyses. In place of fluorescent DNA stains, it is possible to use a more powerful microscope and other stains to examine cell morphology and phenotype in greater detail. Next, the isolated cells were transferred into individual inflatable reaction chambers with G2 lysis buffer (Qiagen) through the on-chip peristaltic pump. Single-cell libraries were then prepared by using a one-pot tagmentation library preparation protocol directly on unamplified single-cell template DNA. Hyperactive Tn5 transposome complexes (Nextera DNA Library Preparation Kit, Illumina) were used to create staggered double-stranded breaks across the genome and to simultaneously ligate unique adaptor sequences to the 5′ end of the target fragment. After the completion of the tagmentation reaction, bound transposase was released with a heat-inactivatable protease (Qiagen). Unique dual-index barcodes sealed into the device during fabrication were resuspended and pushed into each reaction chamber, and 11 PCR cycles were applied on-chip to incorporate both the barcodes and Illumina flow-cell adaptors onto both ends of the tagmented DNA inserts. Subsequently, valves separating the reaction chambers were opened, thus permitting the pooled recovery of indexed single-cell libraries for multiplex sequencing while maintaining the identity of sequencing reads. Additional details are provided in the **Supplementary Note**.

**Sample preparation.** Cells from the immortalized normal human breast epithelial cell line 184-hTERT-L2 were cultured at 37 °C and 5% $CO_2$ in MEBM Mammary Epithelial Cell Growth Medium (Lonza) with transferrin (Sigma) and isoproterenol (Sigma) supplemented with Lonza MEGM(tm) Mammary Epithelial Cell Growth Medium Singlequots. The parental 184-hTERT cell line was generated by C. Barratt (Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences), and the monoclonal 184-hTERT-L2 cell line was derived by A. Burleigh (Aparicio laboratory, BC Cancer Agency). Passage-17 cells were grown to near confluence, trypsinized, resuspended in cryopreservation medium and frozen to −80 °C at a rate of −1 °C per minute. Cells were then stored at −150 °C. For the experiment, 184-hTERT-L2 cells were rapidly thawed in a 37 °C water bath. GM18507 immortalized normal human lymphoblastoid cells were obtained from the Michael Smith Genome Science Centre (GSC) in Vancouver, British Columbia. Cells were cultured at 37 °C and 5% $CO_2$ in RPMI-1640 medium with 2.05 mM l-glutamine (HyClone) supplemented with 10% FBS (Gibco/Invitrogen). Before the experiment, passage-13 cells were cultured at $3 \times 10^5$ cells/mL, spun down at 128*g* for 10 min to enrich for live-cell clusters, and resuspended in fresh medium. We tested both cell lines for mycoplasma with h-IMPACT II human pathogen testing (IDEXX Bioresearch). Patient-derived breast cancer xenograft samples were transplanted and passaged as described in Eirew *et al.*[11]. The anonymized human tumor tissue for xenografting was collected with informed patient consent according to procedures approved by the Ethics Committee at the University of British Columbia, under protocols H06-00289 BCCA-TTR-BREAST and H11-01887 Neoadjuvant Xenograft Study. Female NOD/SCID interleukin-2 receptor gamma–null (NSG) mice were bred and housed at the Animal Resource Centre at the British Columbia Cancer Research Centre. Surgery was carried out on mice between the ages of 5 and 12 weeks. All experimental procedures were approved by the University of British Columbia Animal Care Committee.





doi:10.1038/nmeth.4140

After being harvested, the tissue was finely minced with scalpels, then mechanically disaggregated for 1 min in cold DMEM–F12 medium (Stemcell Technologies). Aliquots from the resulting suspension of cells and organoids were cryopreserved in viable freezing medium and stored at −196 °C until further processing. During an experiment, tissue aliquots were rapidly thawed in a 37 °C water bath and enzymatically dissociated into single cells. First, thawed tissue was incubated in 37 °C warm collagenase (Stemcell Technologies) and hyaluronidase (Stemcell Technologies) for 2.5 h, then incubated for 4 min each with trypsin/EDTA (Corning) and DNase I (Stemcell Technologies) while samples were triturated with a pipette. Finally, dissociated cells were passed through a 50-µm filter in preparation for loading. Additional details are provided in the **Supplementary Note**.

**Cell and device preparation for loading.** Before cell loading, thawed 184-hTERT-L2, fresh GM18507, and dissociated thawed xenograft cells were resuspended in fresh PBS (Life Technologies), filtered, and stained with SYTO 9 Green Fluorescent Nucleic Acid Stain (Life Technologies). Stained cells were mixed with a loading buffer, which was optimized for neutral cell buoyancy and can be adjusted for different cell types. The cell-sorting channels and inlet ports of the microfluidic device were primed with a Pluronic solution to prevent cells from adhering to the PDMS walls and to remove trapped air from the channels.

**Bulk library preparation.** Flash-frozen xenograft tissue was thawed and immediately homogenized in lysis buffer with a rotor-stator homogenizer (Polytron PT1000). DNA was prepared from the lysate with an AllPrep DNA/RNA Mini Kit (Qiagen). Cells from the 184-hTERT-L2 line were thawed, and DNA was extracted with a QIAmp DNA Mini Kit (Qiagen) by following the protocol for cultured cells. DNA was quantified with a Qubit dsDNA HS Assay Kit (ThermoFisher Scientific), and bulk libraries were constructed by following the Nextera DNA Sample Preparation Guide (Illumina) with the following alteration: after tagmentation, the DNA was purified from the transposome with a NucleoSpin PCR Clean-up Kit (Clontech).

**Whole-genome sequencing.** Library insert size and quantity were determined with a Bioanalyzer High Sensitivity DNA Kit (Agilent) and a Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific), respectively. The 184-hTERT-L2 and SA501X3F/X4F xenograft libraries were sequenced in multiplex (192 libraries per lane) on an Illumina HiSeq 2500 instrument with paired-end 125-bp reads at the GSC. The GM18507 libraries were sequenced in multiplex (192 libraries per run) on an Illumina NextSeq instrument with paired-end 125-bp reads at the UBC Biomedical Research Centre (BRC) in Vancouver, British Columbia.

**Data alignment and sequencing metrics.** Demultiplexed FASTQ files for the 184-hTERT-L2 and xenograft libraries were obtained from the GSC. For the GM18507 libraries, BCL files were demultiplexed and converted to FASTQ format with Illumina's bcl2fastq2 program (http://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html/). Demultiplexed paired-end FASTQ files were trimmed to remove Nextera adaptor contamination and low-quality bases on the 3′ ends of reads with Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and were aligned with BWA[48] to the human reference genome GRCh37-lite or the mouse reference genome GRCm38. The resulting BAM files were sorted with Picard (http://broadinstitute.github.io/picard/) and indexed with Samtools[49]. Indels were realigned with GATK[50], and duplicates were removed with Picard. Downsampling and merging of BAM files were also carried out with Picard, as was collection of sequencing metrics. Full details are provided in the **Supplementary Note**.

**Assessment of mouse contamination in xenograft libraries.** All patient-derived human breast cancer xenograft libraries were aligned to the mouse reference genome GRCm38 to assess mouse contamination. Among the 768 indexed libraries sequenced, we identified a single mouse cell (SA501X4F cell 106; 0.07× depth relative to mouse genome, 0.0007× depth relative to the human genome; flagged as a true single cell through fluorescence imaging). Five additional xenograft libraries had minute quantities of mouse contamination (defined as having a depth ≥0.001× relative to the mouse genome), but all had previously been flagged as containing contaminating debris through fluorescence imaging before library construction. The remaining sequenced libraries had depths on the order of 0.0001× relative to the mouse genome (**Supplementary Fig. 11**). The ability to identify libraries with mouse contamination and to exclude those from *in silico* merged bulk- equivalent genomes is an added benefit of the DLP method, which reduces the risk of introducing false-positive variant calls from mouse reads aligning to the human reference genome[11].

**Single-cell quality filtering.** For the comparative analyses in **Figure 2**, cells with fewer than 0.25 million total reads were excluded (184-hTERT-L2, $n$ = 6 cells; GM18507, $n$ = 1 cell; SA501X3F, $n$ = 1 cell; SA501X4F, $n$ = 1 cell). This exclusion criterion was selected because the DOP-PCR data sets used for comparison were prefiltered in this manner[19]. All other libraries identified as true single cells through fluorescence imaging were included, except for the single mouse cell identified in xenograft passage SA501X4F. For all single-cell copy-number analyses (**Fig. 3**, **Supplementary Figs. 5** and **6**, and **Supplementary Fig. 9**), the median absolute deviation (MAD) of all bins assigned to the neutral (two-copy) state was computed, and samples identified as true single cells through fluorescence imaging that also had MAD values <0.15 were retained for downstream analysis (**Supplementary Table 1**). Once again, the single mouse cell in xenograft SA501X4F was excluded.

**Single-cell copy-number inference.** Inference of single-cell copy-number profiles was carried out with HMMcopy[34], with several modifications to the standard usage of this tool. Mappability and GC-content files were generated for the reference genome with the generateMap, mapCounter, and gcCounter tools in the HMMcopy package. Owing to the inclusion of downsampled data sets, single-cell BAM files were binned into 200-kb bins for all data sets analyzed in **Figure 2c**; analysis of all other data sets was carried out with 150-kb bins. Instead of running the model on the logged, GC- and mappability-corrected values, the HMM was run on the nonlogged GC-corrected values, after filtering of bins with low mappability. The HMM with Student's $t$ emissions was run with seven hidden states, rather than the default six states, and the default model parameters were modified as follows:

```
params$m <- c(0, 0.5, 1, 1.5, 2, 2.5, 3)
params$mu <- c(0, 0.5, 1, 1.5, 2, 2.5, 3)
params$kappa <- c(25, 50, 800, 50, 25, 25, 25)
params$e <- 0.995
params$S <- 35
```

After segmentation, binned read counts were converted to integer copy-number scale by dividing all bin counts by half the median value for bins assigned to the neutral (two-copy) state. Segment medians were recomputed, and the median of each segment was rounded to the nearest integer to derive an integer copy-number profile for the cell. Thus, although the last HMM state encompassed all CNVs with six or more copies, high-level amplifications were assigned to integer values. Further details can be found in the **Supplementary Note**.

**Clonal and bulk-equivalent genome analysis.** Clonal genomes were generated by merging the BAM files of all single cells belonging to a given clonal group with Picard. Bulk-equivalent genomes were generated by merging all indexed libraries for a given sample (including those flagged as a single cell, as multiple cells, or as contaminated or ambiguous on the basis of fluorescence imaging), excluding those xenograft libraries flagged as containing mouse contamination. Copy-number inference for the clonal genomes was carried out in HMMcopy, with the same parameters applied to the single-cell libraries, as described above. For bulk-equivalent and bulk genomes, inference of LOH, SNVs, and breakpoints was carried out with Titan[23], mutationSeq[38], and deStruct[39], respectively. Full details can be found in the **Supplementary Note**.

**Data availability.** Genome data have been deposited in the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under accession number EGAS00001002170.


47. Huft, J., Da Costa, D.J., Walker, D. & Hansen, C.L. Three-dimensional large-scale microfluidic integration by laser ablation of interlayer connections. *Lab Chip* **10**, 2358–2365 (2010).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
50. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).






doi:10.1038/nmeth.4140