Cell, Volume *179*

Supplemental Information

Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing

Emma Laks, Andrew McPherson, Hans Zahn, Daniel Lai, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Jerome Ting, Diljot Grewal, Cydney Nielsen, Samantha Leung, Viktoria Bojilova, Maia Smith, Oleg Golovko, Steven Poon, Peter Eirew, Farhia Kabeer, Teresa Ruiz de Algara, So Ra Lee, M. Jafar Taghiyar, Curtis Huebner, Jessica Ngo, Tim Chan, Spencer Vatrt-Watts, Pascale Walters, Nafis Abrar, Sophia Chan, Matt Wiens, Lauren Martin, R. Wilder Scott, T. Michael Underhill, Elizabeth Chavez, Christian Steidl, Daniel Da Costa, Yussanne Ma, Robin J.N. Coope, Richard Corbett, Stephen Pleasance, Richard Moore, Andrew J. Mungall, Colin Mar, Fergus Cafferty, Karen Gelmon, Stephen Chia, The CRUK IMAXT Grand Challenge Team, Marco A. Marra, Carl Hansen, Sohrab P. Shah, and Samuel Aparicio

Data S1

This text is a supplement to:

Clonal decomposition and DNA replication states defined by scaled single cell genome sequencing

Emma Laks[1,2,3,*], Andrew McPherson[1,2,8,*], Hans Zahn[1,3,4*], Daniel Lai[1,2,*], Adi Steif[1,2,3,*], Jazmine Brimhall[1,2], Justina Biele[1,2], Beixi Wang[1,2], Tehmina Masud[1,2], Jerome Ting[1,2], Diljot Grewal[1,2,8], Cydney Nielsen[1,2], Samantha Leung[1,2,8], Viktoria Bojilova[1,2,8], Maia Smith[1,2], Oleg Golovko[1,2], Steven Poon[1], Peter Eirew[1,2], Farhia Kabeer[1,2], Teresa Ruiz de Algara[1,2], So Ra Lee[1,2], M. Jafar Taghiyar[1,2], Curtis Huebner[1,2], Jessica Ngo[1,2], Tim Chan[1,2], Spencer Vatrt-Watts[1,2,8], Pascale Walters[1,2], Nafis Abrar[1,2], Sophia Chan[1,2], Matt Wiens[1,2], Lauren Martin[1,2], R. Wilder Scott[1,2], T. Michael Underhill[4], Elizabeth Chavez[7], Christian Steidl[7], Daniel Da Costa[1,4], Yussanne Ma[5], Robin J. N. Coope[5], Richard Corbett[5], Stephen Pleasance[5], Richard Moore[5], Andrew J. Mungall[5], Colin Mar[9], Fergus Cafferty[9], Karen Gelmon[10], Stephen Chia[10], The CRUK IMAXT Grand Challenge Team[11], Marco A. Marra[6], Carl Hansen[4], Sohrab P. Shah[1,2,8,+], and Samuel Aparicio[1,2,+]

[1]Department of Molecular Oncology, BC Cancer Research Centre, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada

[2]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, V6T 2B5, Canada

[3]Genome Science and Technology Graduate Program, University of British Columbia, Vancouver, British Columbia, Canada.

[4]Centre for High Throughput Biology, Michael Smith Laboratories, University of British Columbia,Vancouver, BC,

V6T 2B5, Canada

[5]Michael Smith Genome Sciences Centre, BC Cancer, Vancouver V5Z 1L3, Canada

[6]Department of Medical Genetics, University of British Columbia,Vancouver, BC, V6T 2B5, Canada

[7]Centre for Lymphoid Cancer, BC Cancer Research Centre, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada

[8]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 417 East 68th St., New York, NY 10065, USA

[9]Department of Radiology, BC Cancer, 600 West 10th Avenue,Vancouver V5Z 4E6, Canada

[10]Department of Medical Oncology, BC Cancer, 600 West 10th Avenue,Vancouver V5Z 4E6, Canada

[11]The CRUK IMAXT Grand Challenge Team

[*]equal contribution

[+]+ to whom correspondence should be addressed. shahs3@mskcc.org, saparicio@bccrc.ca

Cost of DLP+ consumables

DLP+ is a flexible platform, a researcher can choose to target hundreds or thousands of cells in a single experiment. For the purposes of this cost analysis we will assume 1000 cells are included per library, 3000 cells per open array chip, and 6000 cells (two chips) per library construction and library quality control check.

Component	CAD for a unit	Quantity/unit	Amount/use	Cells/use	CAD/cell
Primers	1353.60	504 Smartchips filled	1 chip	3000	0.00090
SmartChip	270.00	1 Smartchip	1 chip	3000	0.09000
Microseal A	189.00	200 seals	6 seals	3000	0.00189
CFSE	232.50	180 ul	0.4 ul	50000	0.00001
Live Dead Red	415.23	250 ul	0.4 ul	50000	0.00001
Lysis Buffer	134.93	50 ml	42.5 ul	3775	0.00003
Protease	95.00	7 ml	25 ul	3775	0.00009
Nextera Kit	10455.80	1.55 ml NPM	75 ul	3375	0.14990
Ampure	1555.72	60 ml	135 ul	1000	0.00350
Agilent HS	849.00	10 Aglilent HS chips	1 chip	6000	0.01415
TOTAL					0.26048

Identification of physical parameter ranges for implementation of DLP+

First, for assessment of per cell and library construction performance, we developed a series of 46 pre and post alignment sequencing metrics as features of each single-cell genome. In addition to standard sequencing and alignment metrics, we developed several new quality metrics based on "integer-ness" scores of copy number states that were shown to substantially improve discrimination (Methods). We explored the weights of these features using a random forest applied to manually classified libraries from the diploid lymphoblastoid cell line GM18507 (**Figure S2c**). The total high quality aligned sequence reads (total mapped reads), the median of bin residuals from segment integer copy number states (MBRSI dispersion non-integerness), and the median of segment residuals from segment integer copy number states (MSRSI non-integerness), a read depth independent metric of non-integer copy number assignment, were the metrics with the highest weights in identifying poorly performing libraries. To simplify the evaluation of the library quality from hundreds of cells, we implemented a random forest classifier to jointly evaluate 18 of these sequencing and post-alignment metrics and provide a single quality score for each cell (QS; **Figure S3a**, sTable 1).

For initial reaction parameter exploration we built sequencing libraries from the GM18507 lymphoblastoid cell line (The International HapMap Consortium, 2005), used an HMM (Ha et al., 2012) to infer copy-number profiles, and classified cells in each library using our trained random forest classifer. We included all wells containing single cells (as identified by microscopy) in the following analysis. Cells that did not produce any reads were assigned a quality score of zero (quality score = 0) during classification.

We examined limiting dilution cell dispensing against real-time selected cells (cellenONE, Scienion) dispensed in a block or limiting dilution-like scatter pattern , and found no significant improvement in library quality of the actively selected cells over the passively dispensed cells for the high Tn5 concentration (KW test, p value = 0.678 (2.2 nL tn5 condition)). The overall library quality built from single cells remained poor, while libraries built from gDNA or crude lysate produced high-quality copy number profiles (data not shown), motivating further optimization. The cellenONE block spotted cells did show a significant increase in total mapped reads compared to the scattered pattern cells that were cellenONE or Poisson spotted (mean total mapped reads cellenONE block = 592240, cellenONE scatter = 207090, limiting dilution = 247197 (2.2 nL tn5 condition), KW test scatter vs block p value = 3.94e-08) (**Figure S2f**). This motivated a switch to cellenONE well specific spotting for all future cell spotting.

Lysis volume, buffer type and crucially, time, proved to be one of the most important parameter sets determining overall performance. We tested new lysis buffer conditions (**Figure S2e**), which are a trade-off between reducing off-reactions in low volumes, vs. evaporation losses and the need to dilute buffer components in subsequent one-pot reactions. The low volumes of lysis buffer (1 nL Buffer G2, Qiagen) used in the microfluidic device (Zahn et al.,

2017) proved not to be robust in the open-array format (**Figure S2a, e ii** 1 nL G2) because the droplets do not fully cover the well. However we found that higher volumes of the same buffer poisoned the downstream reactions due to insufficient dilution (**Figure S2e ii** 10 nL G2; mean total number of reads 1 nL G2 buffer = 1443217, 10 nL G2 buffer = 4136). To address this, we evaluated a PCR compatible lysis buffer (DirectPCR Cell Lysis Reagent, Viagen). The new lysis buffer in combination with the increased lysis volume (10 nL Viagen; **Figure S2e ii**) significantly improved the mean library quality compared to the initial microfluidic lysis condition (KW test, p = 3.00 e-03), but the overall quality remained poor compared to the MF-DLP dataset (**Figure S2e i**). We next tested the PCR compatible lysis buffer (Protease, Qiagen; **Figure S2e iii**) against a low (2.2 nL), medium (3.5 nL), and high (6.5 nL) Tn5 concentration. The overall quality score remained poor suggesting that the genomic DNA was still not fully accessible due to incomplete lysis. However, the library quality significantly improved for the medium (KW test, p = 0.00705) and high (KW test, p = 0.0763) tagmentation concentrations in comparison to the low one, suggesting that the higher Tn5 concentration is able to recover more genomic fragments when the single cell is insufficiently lysed. In contrast, libraries built from gDNA or crude cell lysate produced high-quality copy number profiles (data not shown). Based on this observation, we speculated that the lysis step using the new lysis buffer was not sufficient to expose the DNA for efficient tagmentation and we sought to investigate extended lysis times (2 hours and overnight at 4 °C) over a range of protease concentrations. For all experimental conditions the extended lysis improved copy-number quality significantly compared to the shorter lysis (**Figure S2e iii-v**; KW tests, p = 5.88753e-30). In addition, we found that increasing the protease concentration beyond the amount used in the microfluidic device had little impact on library quality (**Figure S2g**; KW test, p = 0.008).

Key to the robust utilization of the approach, we determined that an overnight lysis at 4 °C in combination with the lowest protease concentration provided the best overall performance across all Tn5 concentrations (**Figure S2g**; mean quality score for 2.2 nL Tn5 = 0.814 $\pm$ 0.367; 3.5 nL Tn5 = 0.842 $\pm$ 0.318, 6.5 nL Tn5 = 0.676 $\pm$ 0.451).

For the lysis expansion comparisons (2 hour and 19 hour cold lysis) with the original MF-DLP data (**Figure S2h**) we evaluated how the genome-wide coverage uniformity of our merged DLP+ libraries compared to the MF-DLP dataset. For each condition in the bootstrap analysis, we plotted one Lorenz curve for the merged genome with median coverage breadth and found that merged DLP+ genomes achieved comparable coverage uniformity to the MF-DLP libraries (**Figure S2i**). It has previously been shown that the MF-DLP single-cell libraries achieved equivalent coverage breadth and uniformity to a standard Nextera bulk genome of equivalent depth (Zahn et al., 2017). It can, therefore, be reasoned that the merged DLP+ genomes also have a comparable quality with that of a bulk library at the same depth. Combined, these results demonstrate that either DLP+ lysis condition sufficiently disrupts cell membranes and proteins and provides adequate access to the genomic DNA during library preparation, generating single-cell libraries

with uniformity equivalent to microfluidic DLP.

While cell lysis is critical, we also explored whether adjustments in Tn5 concentration from the MF-DLP range were required. We observed an improvement in library quality with increased Tn5 concentration for insufficiently lysed cells (**Figure S2e iii**); however, we also detected a significant increase in GC-bias with the increase of Tn5 concentration (**Figure S2l**). GC-bias is a library characteristic that reflects the correlation between coverage depth of a specific genomic location and its GC-content. Strong GC-bias introduced during library preparation can lead to the under-representation of some genomic regions and over-representation of others. This not only complicates CN inference, it can also result in the dropout of AT- and GC-rich regions due to low coverage, thereby undermining SNV and breakpoint inference in merged clonal or bulk-equivalent genomes (Benjamini and Speed, 2012). The other reaction parameter affecting genome representation was the number of post-tagmentation indexing PCR cycles. We prepared DLP+ libraries using a range of Tn5 concentrations and PCR cycles (**Figure S2l, m**) and observed an increase of the average GC-content in our single-cell libraries with increasing Tn5 concentration (8 PCR cycles with 2.2 nL Tn5 (2.2 nL/8PCR) = 47.1%; 3.5 nL Tn5 (3.5 nL/8PCR) = 52.7%; 6.5 nL Tn5 (6.5nL/8PCR) = 61.5%). In contrast, the increased number of PCR cycles had a very small effect on the GC-content (11 PCR cycles with 2.2 nL Tn5 (2.2 nL/11PCR) = 49.4%, **Figure S2l**).

To determine if the increased GC bias compromised our ability to generate a high-coverage merged genome, we carried out bootstrap sampling and merging of single-cell libraries from the different experimental conditions and compared the results to our microfluidic GM18507 DLP dataset (Zahn et al., 2017). It should be noted that the MF-DLP dataset was previously shown to produce equivalent breadth and uniformity to that of a bulk genome at the same coverage depth (Zahn et al., 2017). Before the bootstrap analysis, all single-cell libraries were downsampled to achieve the same mean coverage depth of 0.05X (KW test, p= 0.16). Merging 64 downsampled single-cell libraries resulted in 3.16X median coverage depth and 90% median coverage breadth across all libraries (**Figure S2m**). However, the 6.5nL/8PCR condition and the 2.2 nL/11PCR condition had a significant lower genome coverage breadth (86.2% and 89.9%, respectively) compared to the MF-DLP dataset and both the 2.2 nL/8PCR and 3.5 nL/8PCR conditions (90.6%, 91.3%, 90.7%, respectively; KW test, p = 1.542e-13; a post-hoc Dunn's test with Benjamini-Hochberg correction showed all comparisons between MF-DLP, 2.2 nL/8PCR, nL/8PCR and the 6.5 nL/8PCR, 2.2 nL/11PCR condition were significant, but there was no significant difference between MF-DLP, 2.23.5 nL/8PCR, 3.5 nL/8PCR conditions) at the same coverage depth (KW test, p = 0.6974). Finally, pooling 128 cells at a mean coverage depth of 0.05X per cell resulted in 96.9% coverage breadth at an aggregated depth of 6.35X for the 2.2 nL/8PCR protocol. In comparison, the 6.5nL/8PCR condition achieved significantly less genome coverage (93.2%) at the same depth (KW test, p = 0.3302).

This suggests that the higher GC-bias associated with the increased Tn5 concentration indeed reduced genomic

breadth. To evaluate genome-wide uniformity, we again plotted Lorenz curves for each condition in the bootstrap analysis (**Figure S2m**, **Figure S2n**). We found that the 6.5 nL Tn5 condition is considerably biased, while there was no major difference between the MF-DLP dataset and the 2.2 nL and 3.5 nL Tn5 conditions. For the 2.2 nL and 3.5 nL Tn5 conditions, 64 merged single cells achieve a comparable coverage breadth and uniformity as 64 merged single-cell genomes from the microfluidic dataset (**Figure S2m**).

In order to investigate possible breakpoints in the protocol, we next investigated the effects of storing isolated single cells and nuclei on the open-well device. After dispensing and imaging, chips were sealed and stored at -20 °C for 2 months or overnight ("fresh" controls) (**Figure S2e iv**). The median library quality was similar for nuclei regardless of storage time, but cells stored for longer had a poorer median quality score (**Figure S2e iv**). 55% (n = 15/27) of cells stored for 1 day had quality score <0.75%, 48% (n = 14/29) of cells stored for 63 days had quality score <0.75%, 59% (n = 27/46) of nuclei stored for 1 day had quality score <0.75%, 65%(n = 69/106) nuclei stored for 63 days had quality score <0.75%, which indicates a similar rate of successful cells/nuclei regardless of storage time.

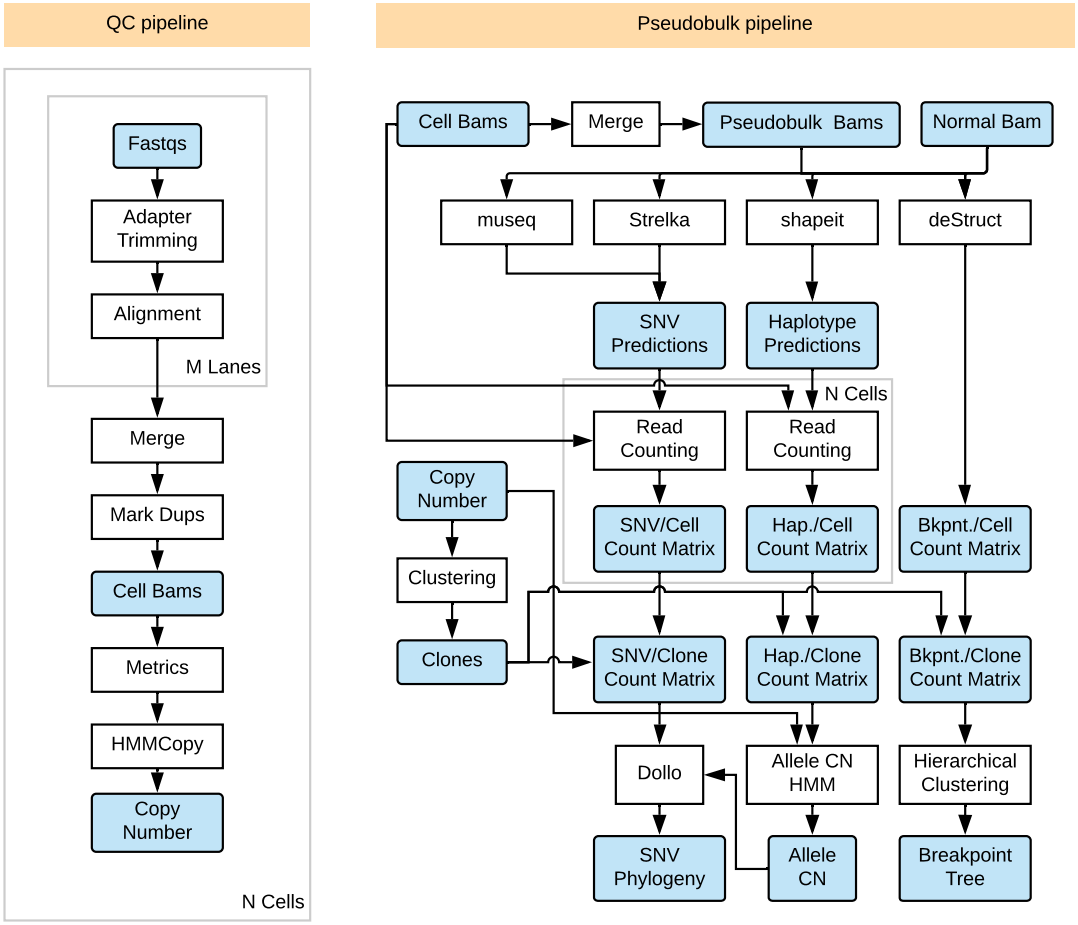
Finally, we examined the impact of dead cells on library quality. We used imaging information to split the lysis time optimized DLP+ libraries (libraries with overnight lysis, 2.2 nL/8PCR and 3.5 nL/8PCR conditions) by cell state (live vs. dead; **Figure S2** e v) and found a significant improvement in the quality of live cells over dead cells (KW test, p = 1.043E-12). Dead cells had a high dropout rate (n= 69% of dead wells failed to produce a library, i.e., less that 250,000 total mapped reads), whereas live cells had a low dropout rate (n= 6% of live wells failed to produced libraries). It is, however, noteworthy that we are able to detect some high-quality CN-profiles in the dead population. This may be related to our inability to distinguish between early and late apoptotic cells with the vital dye staining approaches.

Extension to isolated single nuclei

To extend the range of biospecimens amenable to scWGS, we investigated the ability of DLP+ to process nuclei from flash frozen tissues as well as live cells. We observed that 69.9% of nuclei had quality score above 0.75 (n = 933/1335), showing that DLP+ is equally successful producing high quality libraries from live cells or nuclei (**Figure S3**). Nuclei and cell copy number profiles from the same sample cluster together rather than forming two clusters (**Figure S3b**) and had similar quality score, total mapped reads, duplicate reads, and integerness within libraries where cells and nuclei were prepared in parallel (**Figure S3c**), producing copy number profiles of equal quality (**Figure S3d**).

Single cell pipeline

The single cell pipeline performs a majority of the processing of DLP+ data described in this paper. A high-level depiction of the 2 main pipelines appears in the figure below. Shaded boxes denote datasets and unshaded boxes represent tasks. Light grey boxes show parellelization of parts of each pipeline.



Quality control (QC) pipeline

The QC pipeline performs the necessary steps for quality control of DLP+ data including alignment and copy number calling using HMMcopy. Per cell, per sequencing lane fastqs are processed with TrimGalore to remove adapters from both read ends. Processed fastq files are subsequently aligned using BWA (Li and Durbin, 2009) to produce per lane, per cell BAMs. The resulting BAMs are merged using Picard MergeSamFiles and postprocessed with Picard MarkDuplicates to produce per cell BAMs. Finally, the per cell BAMs are provided as input to HMMcopy to produce per cell copy number calls. The QC pipeline produces results ready to be loaded into Montage.

Pseudobulk pipeline

The Pseudobulk pipeline uses bulk whole genome variant calling tools on merged single cell sequencing data, followed by per cell genotying of discovered variants. Cell BAMs are merged using samtools to produce "pseudobulk" BAMs as input to bulk whole genome analysis tools. MutationSeq and Strelka is used to predict Single Nucleotide variants (SNVs), deStruct is used to predict breakpoints, and shapeit is used to infer haplotype blocks from heterozyous Single Nucleotide Polymorphisms (SNPs). MutationSeq and Strelka results are merged and per cell read counts are obtained from Cell BAMs for each predicted SNV. Per cell SNP read counts are generated similarly and deStruct produces a per cell breakpoint read count matrix directly.

Clonal analysis begins with clustering of per cell copy number data to produce putative genomic clones. SNV, SNP and breakpoint read counts are merged by clone to produce per clone variant read count matrices. A custom HMM is used to infer allele specific copy number from HMMcopy predicted total copy number, per clone SNP read counts, and haplotype block information. An SNV phyogeny is inferred from per clone SNV read counts and allele specific copy number. Finally, a tree relating clones by their breakpoint composition is inferred using hierarchical clustering.

Computational parameters for DLP+ analyses

Genomic analyses including HMMcopy and pseudobulk variant calling were run with consistent parameters across all DLP+ libraries, whereas copy number clustering used two different approaches. Future analyses involving non-standard datasets or scientific questions not considered in this paper may require parameter tuning. Below we describe the most relevant parameters of interest used in the computational analyses for this paper.

HMMcopy

Bin size

The most critical parameter for copy number calling is the bin size used to generate a histogram of read counts across the genome. Selecting an appropriate bin size requires finding a balance between copy number calling sensitivity and ability to accurately correct for GC bias. A smaller bin size will allow HMMcopy to more accurately infer the location copy number changepoints, and will also allow for identification of smaller copy number changes and resolution of the precise details of complex copy number changes. Nevertheless, a smaller bin size will also result in increased noise. Individual bins will contain a smaller number of reads increasing sampling variance. Local variation in sequencing the sequencing efficiency of DNA within each bin will also result in increased read count variance. Finally, with increased

variance in binned read counts, calculation of the sequencing associated GC bias will be less accurate.

We have selected a bin size of 500k nucleotides for this study based on an estimated average 120 reads per bin for 1 lane of HiseqX sequencing (375 million reads) for a library of approximately 1000 cells. In practice, calling copy number changes smaller than 4 bins is not feasible limiting the resolution of the method to 2Mb features. Focul high level amplifications are an exception however, and can be identified with sensitivity related to the level of amplification. As an example of the kind of events we have focused on in this paper we list below the amplifications found in the FNA sample including the length of the encompassing segment and the copy number state of that segment.

Gene name	Clone	Segment Length	Copy number
CCNE1	B	3000000	8
CCNE1	C	3000000	8
CCNE1	D	1500000	7
MCL1	B	33500000	8
MCL1	C	31000000	8
MCL1	D	35500000	8
MYC	B	14000000	10
MYC	C	10000000	11
MYC	D	13500000	9
RAB18	B	7000000	9
RAB18	C	7000000	9
RAB18	D	28000000	5
RAD18	B	9000000	8
RAD18	C	500000	7
RAD18	D	19500000	4

In future work on copy number with DLP+ data will consider re-evaluating this parameter in the context of an improved copy number calling method.

Additional HMMcopy parameters

HMMcopy as used to call copy number, originally developed for calling CNV events in WGS bulk analysis. Two major novelties were introduced to allow the tool to work optimally for single-cell copy number analysis. Firstly, the original HMMcopy only have 6 states to capture the most commonly observed range of events from homozygous deletion (0 copies) to high level amplifications ($\geq$ 5 copies). In single cell, due to the more digital nature of the observed data (i.e. data rounding to integer copy number counts), we were able to resolve the higher copy number events with much higher confidence and clarity, and we were able to expand the state space to 12 unique states ranging from 0 to 11 copies.

Secondary, due to the digital nature of the copy number, instead of working in logarithmic space, we were able to operate in linear space, which mean an adjustment to multiple variables, such as the expected mean of states and their

priors. Together, this means that for HMMcopy, we now run an 11 state parameter input with linear space values as follows:

	strength	e	mu	lambda	nu	kappa	m	eta	gamma	S
1	1e+30	0.9	0	20	2.1	25	0	5e+04	3	0.01858295
2	1e+30	0.9	1	20	2.1	100	1	5e+04	3	0.01858295
3	1e+30	0.9	2	20	2.1	670	2	5e+05	3	0.01858295
4	1e+30	0.9	3	20	2.1	100	3	5e+04	3	0.01858295
5	1e+30	0.9	4	20	2.1	25	4	5e+04	3	0.01858295
6	1e+30	0.9	5	20	2.1	25	5	5e+04	3	0.01858295
7	1e+30	0.9	6	20	2.1	25	6	5e+04	3	0.01858295
8	1e+30	0.9	7	20	2.1	10	7	5e+04	3	0.01858295
9	1e+30	0.9	8	20	2.1	5	8	5e+04	3	0.01858295
10	1e+30	0.9	9	20	2.1	5	9	5e+04	3	0.01858295
11	1e+30	0.9	10	20	2.1	5	10	5e+04	3	0.01858295
12	1e+30	0.9	11	20	2.1	5	11	5e+04	3	0.01858295

Seen here are the parameters for the 12 states, corresponding to copy numbers 0 to 11, with expected means set as mu, and expected distribution set with $kappa$. The remaining parameters are as described in HMMcopy.

Copy number clustering

To cluster cells into clones with similar copy number we used a combination of dimensionality reduction using UMAP followed by a clustering step. The relevent UMAP parameters are `n_neighbours`, `min_dist`. The `n_neighbours` parameter constrains the size of the local neighbourhood UMAP will look at when learning the manifold structure. Since we expect the cells to cluster into clones with very similar genotype, we used the default value of 15 allowing UMAP to consider the larger neighbourhood of each cell, and find a representation that preserves the global structure between clones. The `min_dist` parameter controls the compactness of the low dimensional embedding, and here also we use the default value as a balance between producing an embedding that will be over-clustered in the second step, while retaining the ability to identify chains of highly related cells with similar copy number.

We used two distinct clustering methods, Gaussian Mixture Models and HDBScan. From our observations of the libraries we have generated thus far, most populations are predominantly composed of a set of clusters with very similar copy number, in addition to outlier cells produced by biological noise including cells with mitotic error, cells in s-phase and failed cells. For both clustering methods we perform a clustering step followed by a filtering of outliers and smaller clones, allowing for identification of the major populations and removal of cells with biological features that make them difficult to cluster. Note that in some analyses we have reassigned both s-phase and mitotic error cells

post-hoc to major populations to understand the proportion of these cell states in specific clones.

For the GMM based method, the major parameter are the number of mixture components, the parameters for outlier removal, and the threshold on minimum clone size. To identify the major populations, we use the strategy of overspecifying the number of clones in the mixture model, and pruning outliers and small clones. The number of mixture components was selected as at least double the number of components visible in the low dimensionality representation (n=20). Outliers were classified as having an RMS deviation from the mean copy number of 0.8 or more and clusters were removed if they were composed of fewer than 50 cells. Note that the minimum cluster size parameter is also highly relevant for the accuracy of the pseudobulk analysis (see below).

HDBScan was used for the analysis of larger collections of cells including the clonal analysis of the 184-hTert cell line passages. The HDBScan algorithm automatically selects the number of clusters and adds outlier data points to an outlier cluster. Two parameters of the HDBScan algorithm are relevant, `min_samples` and `min_cluster_size`. For all analyses we use a larger than default value of `min_cluster_size`$= 30$ to reflect the fact that we are only considering clusters of size 50 or more cells. We set `min_samples`$= 15$ as recommended in the HDBScan guide, to mitigate the over calling of outliers resulting for larger values of `min_cluster_size`. As an added filtering step, we removed all cells that showed higher pearson correlation with a cluster other than that to which they were assigned. As for the GMM method, we remove clones smaller than 50 cells in size.

Finally, we note that clustering copy number profiles of single cells is a problem that requires further method development in addition to the development of methods for objectively assessing performance. Future method development will likely not combine dimensionality reduction with clustering as this approach is known to suffer from reproducability issues in some contexts. We leave further improvements to the clustering analysis for future work.

Pseudobulk analysis

Suitability of input data

The DLP+ platform's primary use is identification of copy number changes in single cells, and as shown by our results, excels when applied to aneuploid cell types or in the context of whole chromosome loss and gains due to mitotic error. The ovarian cell lines to which we applied our phylogenetic analysis are highly genomically instable, resulting in clearly defined clones divergent in copy number. At the other extreme, DLP+ may not be the ideal platform for phylogentic analysis of a sample divergent in SNVs but not copy number change. Between these two extremes there may be samples with some clonal divergence at the copy number level for which DLP+ will perform well. This may especially be the case with further improvements to algorithms or models for clustering single-cell copy number,

including advances such as the ability to borrow strength across cells or between adjacent genomic bins.

As implemented in this paper, copy number clustering treats the read count observations for each bin as independent. Thus each dataset amounts to N observations in a K dimensional space for N cells and K bins, and conclusions regarding this very generic problem should apply. Existing work on clusterability (Ackerman and Ben-David, 2008) may help understand the amount of copy divergence that is required for a DLP+ dataset to be 'clusterable'. For instance, in (Zhang and Li, 2013), the authors use as a measure of clustered-ness the ratio of inter-cluster variance over within-cluster variance. For DLP+ data, within cluster variance resulting from sequencing and alignment noise is impacted by coverage, whereas inter cluster variance results from copy number divergence. The number of cells and how balanced they are between clusters will also affect clusterability, higher within cluster variance may cause a small number of cells to be subsumed into a larger cluster especially if the copy number divergence is low. Future simulation studies will allow for more quantitative statements surrounding the amount of copy number divergence that may be detected and the impact of imbalanced populations and additional sequencing coverage.

Minimum clone size

One of the most critical features for both the clustering and pseudobulk analyses is minimum clone size. A smaller minimum clone size may be necessary when tracking fitness in initally small populations of cells in time series datasets, or for studying negative selection due to mitotic error or other genomic features. However, analyses of the proportions of cells with specific genomic features, such as proportions of s-phase and mitotic error cells, will be less accurate for smaller clones. Minimum clone size will also impact the depth of coverage per clone, and with it the ability to call additional clone specific single position based genomic features including SNVs, rearrangement breakpoints, and allele specific copy number change. DLP+ sequencing data is largely uncorrelated between cells, thus average coverage depth is roughly proportional to the number of cells. For the OV2295 sample with 4 clones for instance, clone size is highly correlated with total coverage at SNV positions (pearson-r=0.97, p-value=0.03). We discuss the coverage requirements for pseudobulk analysis in more detail below.

SNV and breakpoint detection

To detect SNVs and breakpoints, used standard bulk whole genome sequencing parameters for MutationSeq, Strelka and deStruct. For Strelka we used a somatic score filter of 20. For MustationSeq we used a probability score filter of 0.9. For deStruct, we retained breakpoints for which `num_split` ≥ 5 and `template_length_min` ≥ 250.

Additionally, the structure of DLP+ data provides the additional opportunity to filter for variants that are nominated by 2 or more independent molecules. DLP+ barcodes can be used in a way that is similar to Unique Molecular

Barcodes (UMIs); an SNV with support in two different cells must have originated from two indepednent molecules (not-withstanding artifacts in the barcodes themselves). We leveraged this fact to provide an additional level of filtering for SNVs, removing SNVs with evidence of the alternate allele in only 1 cell.

Coverage requirements for pseudobulk analysis

Coverage breadth and depth has a significant impact on the ability to call position specific genomic features such as SNVs, breakpoints and allele specific copy number. Specifically, insufficient coverage will result in insufficient sampling of both alternate and reference alleles.

As an example, below we calculate the proportion of SNVs falsely called as absent (false negative rate) for varying mean coverages and genotypes. We make the following assumptions: 1. coverage at each mappable genomic loci is poisson distributed according to mean genome wide coverage, 2. the number of alternate reads is binomial distributed 3. copy number does not vary significantly across the genome 4. presence is detection of at least 1 read supporting the SNV Note that in this specific instance, 10X coverage appears to provide a lower bound for detection of SNVs at a reasonable false negative rate.

genotype coverage	AB	AAB	ABB	AAAB	ABBB
1	0.606	0.717	0.515	0.779	0.472
2	0.367	0.513	0.264	0.606	0.223
5	0.083	0.189	0.036	0.287	0.023
10	0.007	0.036	0.001	0.082	0.001
20	0.000	0.001	0.000	0.007	0.000
30	0.000	0.000	0.000	0.001	0.000
40	0.000	0.000	0.000	0.000	0.000
50	0.000	0.000	0.000	0.000	0.000

We performed HMMcopy and pseudobulk analysis on the 3 OV2295 cell line libraries at two levels of coverage: 41-44X per library / 0.059-0.068X per cell, and approximately 88-93X per library / 0.12-0.16X per cell. We sought to estimate the relative benefit of sequencing to approximately 40X versus 90X with respect to accuracy of SNV detection. First we calculated the number of clones in which each SNV was detected as present, and compared these numbers between the 40X and 90X datasets (**Figure S5a**). In total 3634 SNVs differed in the number of clones in which they were identified, while 9639 SNVs were identified in the same set of clones between each dataset. Approximately 40% of the SNVs called as private were reassigned as shared between 2 or more clones in the higher coverage data. In general, SNVs were called as more ancestral in the higher coverage data.

Next we sought to determine the extent to which SNV absences could be explained by their phylogenetic origin as opposed to a lack of coverage at the SNV loci. Taking the infinite sites hypothesis, each SNV results from a mutation that occurs once in the evolutionary history of the tumour and cannot revert to the reference allele in descendent clones. An SNV is deemed to be phylogentically incompatible if it cannot be assigned to a branch of the phylogeny such that it is observed in all descendant observed clones. An SNV is said to be phylogenetically incompatible for a given clone if that SNV is not detected in that clone despite its predicted presence in ancestral clones.

We measured the total coverage for each SNV within each clone, splitting the SNVs into present, absent, and incompatible within those clones (**Figure S5b,c**). We repeated the analysis for both 40X and 90X datasets. In the 40X dataset, incompatible SNVs have significantly lower coverage than present or absent SNVs, whereas present and absent SNVs are approximately consistent in their coverage. In the 90X dataset by contrast, absent and incompatible SNVs are consistent for the majority of copy numbers. For copy number 2-8, present SNVs are marginally higher in average coverage than both absent and incompatible SNVs, possibly indicating that copy number loss may explain the absence of some of these SNVs. Thus for copy number in the range 2-8, we conclude that the median clone coverage of 15X, minimum 10X for the 90X dataset is sufficient to accurately reconstruct the clonal SNV genotypes in the OV2295 data.

In-silico cell mixing experiments

To evaluate the relative performance of DLP+ and bulk sequencing for the purposes of clonal inference, we compared clone proportions predicted by ReMixT v0.5.7, TheTA2 v0.7 and CloneHD v1.17.8 with clustering applied to DLP+ copy number profiles. We chose these 3 tools as they directly predict clone proportions whereas tools such as Titan and Battenberg estimate only the related measurement, cellular prevalence. In brief, we performed in-silico mixing of cells from each HGS ovarian cell line sample in known proportions, downsampling each cell's read data to a fixed coverage. We then applied HMMcopy, dimensionality reduction and clustering to unmixed single-cell data, and ReMixT, TheTA2 and CloneHD to merged data. We created 14 mixtures in total, varying the number of clones between 2 and 3 and clone proportions with each clone having at least 5% prevalence. We compared the results as follows. We first calculated a correlation between each methods clone copy number, and the average normalized read count in each sample. As a similarity metric we use correlation between predicted copy number and sample level read depth for two reasons. First, we wished to avoid calling copy number in the samples prior to mixing, so as to avoid any bias that might add to the comparison. Second, we wished to evaluate clone proportions even when the tool being evaluated predicted the ploidy of the clone incorrectly. Next we assigned each predicted clone the the sample with

which it is maximally correlated. Finally, we compared the predicted proportions of each clone with the proportions of each sample in each mixture. As expected, HMMcopy would frequently identify more than 1 clone in each sample, and in these cases we would sum the predicted proportions of all clones attributed to a single sample, and compare those summed proportions to the proportion of each sample in each mixture.

Mixture Name	Num. Clones	Minor Clone Frac.	Is Mixed Ploidy
mixture1	2	0.05	False
mixture2	2	0.10	False
mixture3	2	0.20	False
mixture4	2	0.30	False
mixture5	2	0.40	False
mixture6	2	0.50	False
mixture7	3	0.05	True
mixture8	3	0.05	True
mixture9	3	0.10	True
mixture10	3	0.10	True
mixture11	3	0.20	True
mixture12	3	0.20	True
mixture13	3	0.30	True
mixture14	3	0.30	True

HMMcopy and clustering performed on the unmerged data outperformed ReMixT, TheTA2 and CloneHD with respect to the total clone fraction error (**Figure S5d**), the number of simulated clones captured in the results (**Figure S5e**), and the average correlation with the clone copy number (**Figure S5f**). Neither ReMixT not TheTA2 was able in decisively identify a clone with copy number matching each sample in the mixture, instead predicting 2 clones that both matched more strongly with a single sample. CloneHD was able to predict clones that matched the samples in each mixture for a majority of the 2 clone simulations, but performance on other metrics remained poor. TheTA2 predicted several of the mixtures to have significant normal contamination, with an associated negative impact on the tools performance metrics for those mixtures. We suspect that none of the deconvolution tools are well suited to mixtures of tetraploid populations. Additionally, we hypothesize that although these tools may be capable of accurately predicting regions of subclonal copy number, they are not able to phase the subclonal copy number between these regions resulting in predictions of clone copy number that are a combination of the true clone copy number across the

genome.