Interactive analysis and quality assessment of single-cell copy-number variations

Tyler Garvin*, Robert Aboukhalil*, Jude Kendall, Timour Baslan, Gurinder S. Atwal, Jim Hicks, Michael Wigler, Michael C. Schatz**

Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor NY, 11724

Email:

Tyler Garvin tgarvin@cshl.edu,

Robert Aboukhalil raboukha@cshl.edu

Jude Kendall: kendall@cshl.edu

Timour Baslan: tbaslan@cshl.edu

Gurinder S. Atwal: atwal@cshl.edu

James Hicks: hicks@cshl.edu

Michael Wigler: wigler@cshl.edu

Michael Schatz mschatz@cshl.edu

*Equal Contribution

**Corresponding Author: mschatz@cshl.edu

Abstract

Single-cell sequencing is emerging as a critical technology for understanding the biology of cancer, neurons, and other complex systems. Here we introduce Ginkgo, a web platform for the interactive analysis and quality assessment of single-cell copy-number alterations. Ginkgo fully automates the process of binning, normalizing, and segmenting mapped reads to infer copy number profiles of individual cells, as well as constructing phylogenetic trees of how those cells are related. We validate Ginkgo by reproducing the results of five major single-cell studies, and discuss how it addresses the wide array of biases that affect single-cell analysis. We also examine the data characteristics of three commonly used single-cell amplification techniques: MDA, MALBAC, and DOP-PCR/WGA4 through comparative analysis of 9 different single-cell datasets. We conclude that DOP-PCR provides the most uniform amplification, while MDA introduces substantial biases into the analysis. Furthermore, given the same level of coverage, our results indicate that data

prepared using DOP-PCR can reliably call CNVs at higher resolution than data prepared using either MALBAC or MDA. Ginkgo is freely available at http://qb.cshl.edu/ginkgo.

Introduction

In recent years, single-cell sequencing [1] has become an important tool for probing cancer [2], neurobiology [3], and developmental biology [4-6]. Studying genomic variation at the single-cell level allows investigators to unravel the genetic heterogeneity within a sample and enables the phylogenetic reconstruction of subpopulations beyond what is possible with standard bulk sequencing, which averages signals over millions of cells. To date, thousands of individual human cells have been profiled to map the subclonal populations within cancerous tumors [7], and circulating tumor cells [8,25], discover mosaic copy number variations in neurons [3], and identify recombination events within gametes [5, 9], among many other applications.

One of the most significant applications of single-cell sequencing is to identify large-scale (>10kb) copy-number variations (CNVs) [3, 7, 10]. These events can radically alter cell biology and are important in neurodegenerative and developmental disorders among many other conditions. Within heterogeneous samples, CNVs form a "genetic fingerprint" from which the phylogenetic history of a sample may be inferred [11]. For example, within cancer biology, this technique has been used to study the widespread heterogeneity within tumors and has led to greater understanding of tumor development and metastasis [7].

CNVs are also the most readily accessible variants to analyze by single-cell sequencing. Whole-genome amplification (WGA) techniques are needed to prepare the DNA within a single cell for high-throughput sequencing [12]. Despite dramatic improvements since their introduction, WGA protocols inevitably have variable amplification efficiency across the genome and even total loss of coverage in certain regions [13, 14]. This makes detecting single-nucleotide mutations and other small mutations unreliable, but large copy-number variations (>10kbp) are robustly detected through statistical analysis of changes in read depth across the genome [15].

Given the insight derived from single-cell sequencing, many researchers are now interested in applying the technology to diverse cell types and species. However, the downstream analysis is complex, requiring a multistage pipeline to bin reads into genomic regions, normalize for GC-content and other sequencing biases, segment bin counts into regions of consistent copy-number state, and cluster copy-number states across cells [16]. As such, a major challenge is the lack of validated, easy to use, open-source software that can execute this pipeline automatically and correctly.

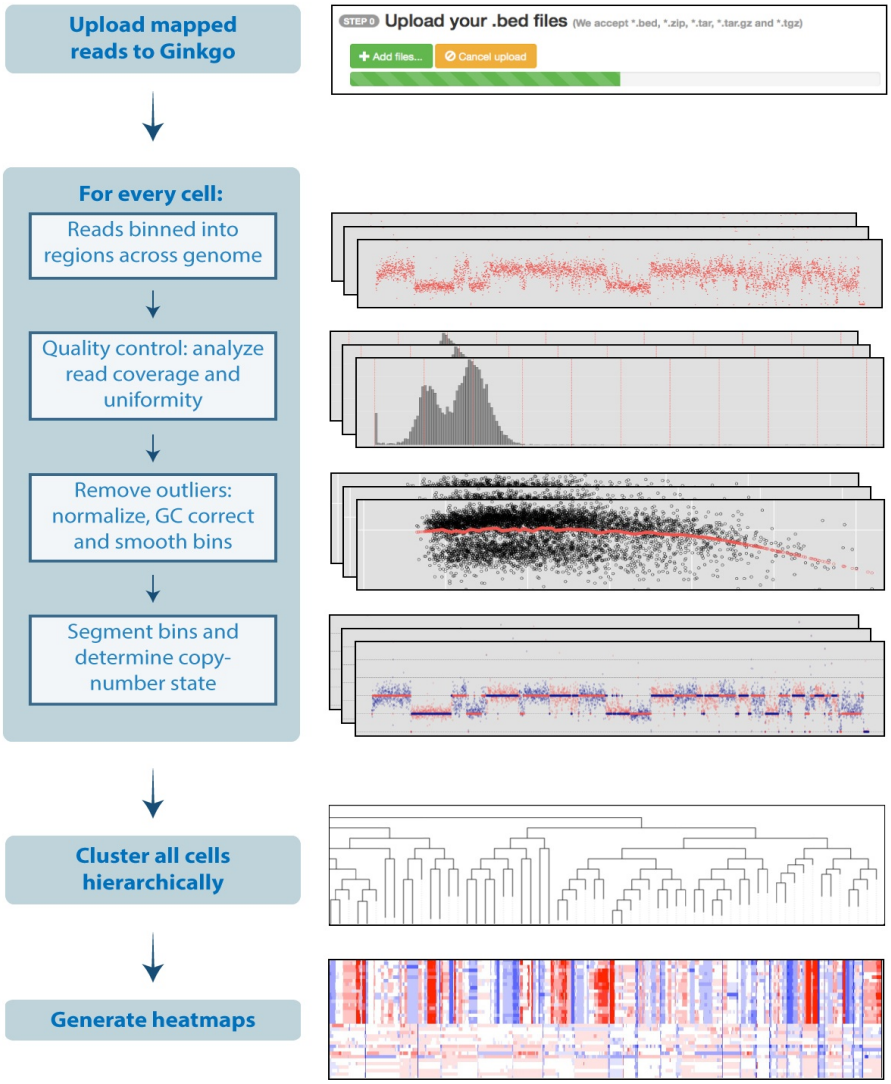


Figure 1: The Ginkgo flowchart for performing single-cell copy-number analysis. Usage and parameters are described in the supplementary information and on the website.

Here we present our new open-source web analytics platform, Ginkgo, for the automated and interactive analysis of single-cell copy-number variations (Figure 1). Ginkgo enables researchers to upload samples, select processing parameters, and after processing, explore the population structure and cell-specific variants revealed by the CNVs within a visual analytics framework in their web browser. Detailed reports, a wide array of quality assessment metrics, and high-resolution figures are also generated.

We validate Ginkgo by reproducing major findings of several single-cell sequencing studies that employ three different WGA techniques: MALBAC [8, 9, 17], DOP-PCR/WGA4 [3, 7], and MDA [5, 18, 19]. Taken together, we analyze the data characteristics of nine datasets across five tissue types (Table 1). Next, we explore

the biases present in single-cell data and the techniques employed by Ginkgo to overcome these biases. Finally, we compare biases and coverage uniformity between the three most widely published WGA techniques.

We find that both DOP-PCR outperform both MALBAC and MDA in terms of data quality. As previously reported [13, 14], MDA displays poor coverage uniformity and low signal-to-noise ratios. Coupled with overwhelming GC biases, MDA is unreliable for accurately determining CNVs compared to the other two techniques. Furthermore, we show that while both DOP-PCR and MALBAC data can be used to generate CNV profiles and identify large variants, DOP-PCR data has substantially lower coverage dispersion and smaller GC biases when compared to MALBAC data. Given the same level of coverage, our results indicate that data prepared using DOP-PCR can reliably call CNVs at higher resolution with better signal-to-noise ratios, and is more reliable for accurate copy-number calls.

Ginkgo is available open-source and can be run directly from the website at: http://qb.cshl.edu/ginkgo.

Study	WGA Method	Tissue Type	# of cells	Accession
Kirkness *et al.* (2013)	MDA	Sperm	11	SRP017516
Wang *et al.* (2012)	MDA	Sperm	31	SRA053375
Evrony *et al.* (2012)	MDA	Neuron	8	SRA056303
Lu *et al.* (2012)	MALBAC	Sperm	99	SRA060945
Ni *et al.* (2013)	MALBAC	Lung	29	SRP029757
Hou *et al.* (2013)	MALBAC	Oocyte	181	SRA091188
Navin *et al.* (2011)	DOP-PCR	Breast (T10)	100	SRX021401
Navin *et al.* (2011)	DOP-PCR	Breast (T16)	100	SRX037035/SRX037132
McConnnell *et al.* (2013)	DOP-PCR	Neuron	109	SRP030642

Table 1: List of the 9 datasets analyzed across 8 different studies.
Note that there are two distinct datasets from the same Navin *et al.* study.

Results

Ginkgo pipeline validation

Ginkgo provides a user-friendly interface that guides users through every aspect of the analysis, from binning reads into regions across the genome, to quality assessment, GC bias correction, segmentation, copy-number calling, visualization and exploration of results (**Figure 1**, and details below). The processing pipeline is based on previous methods [16], although it is a completely new implementation that packages together all of the steps into a flexible but integrated platform. It also

includes several novel features not previously described to advance the state of the art, including a new algorithm for determining absolute copy number state from the segmented raw read depth, and a new method for controlling for quality issues in the reference assembly (see "Bad bins" below). Once a user uploads data files and selects analysis parameters, Ginkgo displays the progress at each step. Since samples may take several minutes to hours to process depending on the number of cells and sequencing depth of coverage, the user can also be notified by e-mail once the analysis is done.

Once the analysis completes, Ginkgo displays an overview of the data in a set of heatmaps as well as an interactive phylogenetic tree [20] of all cells used in the analysis. Clicking on a cell in the interactive phylogenetic tree allows the user to view an interactive plot of the genome-wide copy number profile of that cell, search for genes of interest, and link out to the UCSC genome browser [26]. Ginkgo also outputs quality assessment graphs for each cell: a plot of read distribution across the genome, a histogram of read count frequency per bin, and a Lorenz curve to assess coverage uniformity [13]. All plots, statistical measurements and clustering results can be downloaded in publication-quality figures or as simple text files. All results are saved on our servers for several months, allowing the user to return to their results at a future date and run different analyses with the same data. A unique cryptographically secure URL is generated for each project, allowing researchers to easily share the displays with collaborators of their choosing while maintaining security of their data. Alternatively, for ease of use and potential privacy concerns, we provide and document all the software necessary for hosting the web tool on private servers for extended analysis or customization.

To validate Ginkgo, we set out to reproduce the major findings of six human datasets across five single-cell studies that used MALBAC and DOP-PCR amplification. These datasets address vastly different scientific questions, were collected from a variety of tissue types, and make use of different experimental and computational approaches at different institutions. The three MDA datasets were not used to validate our pipeline due to lack of published CNV analysis and poor coverage uniformity (see below). Using Ginkgo, we replicated the published CNVs in all aneuploid cells across the remaining six datasets (Supplementary Note 1: Replication of the single-cell analysis results), except one cell in Hou *et al*., which we believe was due to mislabeling in the NCBI SRA.

Moreover, the Navin *et al*. and Ni *et al*. datasets used the identified CNVs to generate phylogenetic trees across all samples, which Ginkgo can also reproduce. In the first Navin *et al*. dataset (T10), Ginkgo clusters all 100 samples into the same four distinct subpopulations of the original study (**Supplementary Figure 1A**). In the second Navin *et al*. dataset (T16P/M), Ginkgo clusters all 100 samples into the

same three distinct subpopulations as the original publication, linking the primary tumor to its metastasis (**Supplementary Figure 2B**).

Finally, through CNV analysis and hierarchical clustering of 29 CTCs across 7 patients with lung adenocarcinoma (ADC) or small-cell lung cancer (SCLC), Ni *et al*. discovered that CNVs appear specific to cancer types and distinct from patient to patient. Using default settings in Ginkgo, we can also reproduce the published clustering results (**Supplementary Figure 2A**).

However, careful consideration of gender must be given when analyzing patients from mixed populations, as the combined set of the X and Y-chromosomes make up a large fraction of the human genome that can distort the clustering results. When we examined the Ni *et al*. dataset with Ginkgo with sex chromosomes masked, we could still discriminate between individual patient's tumors, but we could no longer discriminate between ADC and SCLC (**Supplementary Figure 2B**); the SCLC patients were exclusively female and ADC patients were almost entirely male. Ginkgo comes prepackaged with the ability to selective mask sex chromosomes to prevent gender biases from dominating the clustering.

Initial comparisons of three single-cell sequencing techniques

We used Ginkgo's quality assessment tools to evaluate the characteristics and merits of three commonly used whole-genome amplification (WGA) techniques: MDA, MALBAC**,** and DOP-PCR. Raw sequencing reads from each of nine datasets were downloaded from NCBI using the accession numbers in Table 1. Reads were mapped to the human genome and downsampled to match the lowest coverage sample (see **Methods**). Finally, aligned reads were binned into 500kb variable-length intervals across the genome, such that the intervals average 500kb in length but contain the same number of uniquely mappable positions (see **Methods**). We use these binned read counts to measure two key data quality metrics: GC bias and coverage dispersion. Importantly, raw bin counts provide a robust view of the data quality impartial to the different approaches to segmentation, copy-number calling, or clustering.

GC content bias

GC content bias refers to preferential amplification and sequencing because of the percentage of G+C nucleotides in a given region of the genome [21]. This introduces cell-specific and library-specific correlations between GC content and bin counts. In particular, when GC content in a genomic region falls outside a certain range (typically <0.4 or >0.6), read counts rapidly decrease (**Methods**). In order to compute the GC biases across all nine datasets we calculate the lowess fit of the log base 2 normalized read counts with respect to the bin GC content for each sample. A

sample with no GC bias would have a flat normalized read count of zero across all bins and all GC values. After the lowess fit, we monitor the bias of each cell by calculating the proportion of bins that show a two fold change from the expected coverage in either direction (by +/- 1, log base 2). We find that MDA has very high GC bias compared to MALBAC and DOP-PCR (**Figure 2**). Only 45.9% of MDA bin counts fall within the bounded range compared to 94.0% of MALBAC bin counts and 99.6% of DOP-PCR bin counts. It is important to note that, regardless of WGA approach, each cell has unique GC biases that must be individually corrected as explained later in the manuscript.

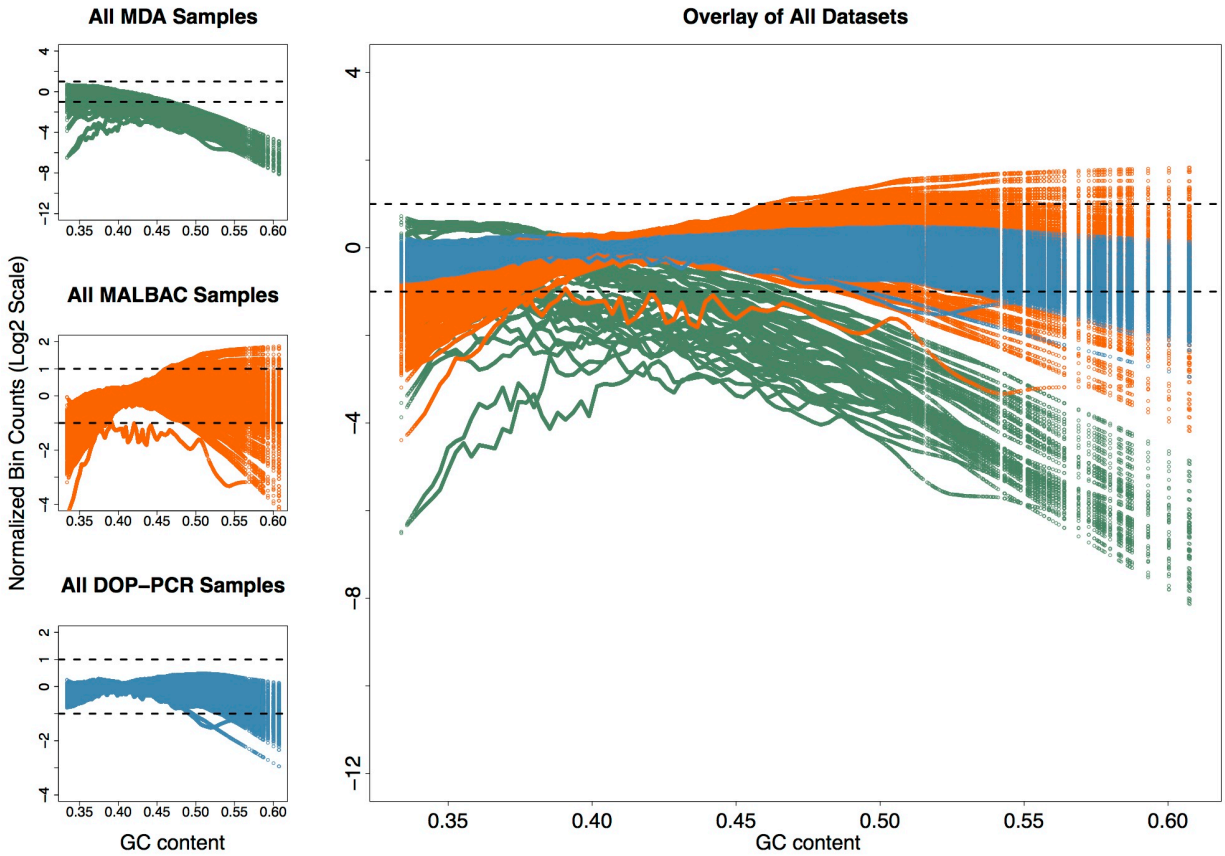


Figure 2 | Lowess fit of GC content with respect to log normalized bin counts for all samples in each of the 9 datasets analyzed: 3 for MDA (top left – green), 3 for MALBAC (center left – orange), and 3 for DOP-PCR (bottom left – blue). Each colored line within a plot corresponds to the lowess fit of a single sample. Note that the three MDA datasets (top left) have a different y-axis scale due to the large GC biases present.

Coverage dispersion

Once reads are mapped into bins and each cell GC-corrected, we normalize within cells by dividing the count in each bin by the mean read count across all bins. As a measure of data dispersion, we calculate the median absolute deviation (MAD) of all pair-wise differences in read counts between neighboring bins for each sample.

MAD is resilient to outliers caused by copy-number breakpoints, as transitions from one copy-number state to another will be infrequent. Instead, pair-wise MAD reflects the bin count dispersion due to technical noise. For each of the nine datasets, the MAD was calculated for each cell and displayed in a box-and-whisker plot (**Figure 3**). As expected from previous comparisons of MDA to other WGA techniques [13, 14], MDA data displays high levels of coverage dispersion on average with a mean MAD 2-4 times that of the DOP-PCR datasets. In addition the MALBAC and MDA datasets show large differences in data quality between studies while the DOP-PCR datasets show consistent flat MAD across all three studies (**Supplementary Figure 3**). While the MALBAC datasets outperform the MDA datasets on average, note that the samples from the Evrony *et al.* MDA dataset actually show a lower mean mad than the Lu *et al.* MALBAC dataset. We also calculate the pair-wise MAD for neighboring bins that are offset by 2, 3, or 4 bins (**Figure 4**, x-axis). An increase in the MAD as the bin offset increases indicates the presence of biases due to local trends in the data.

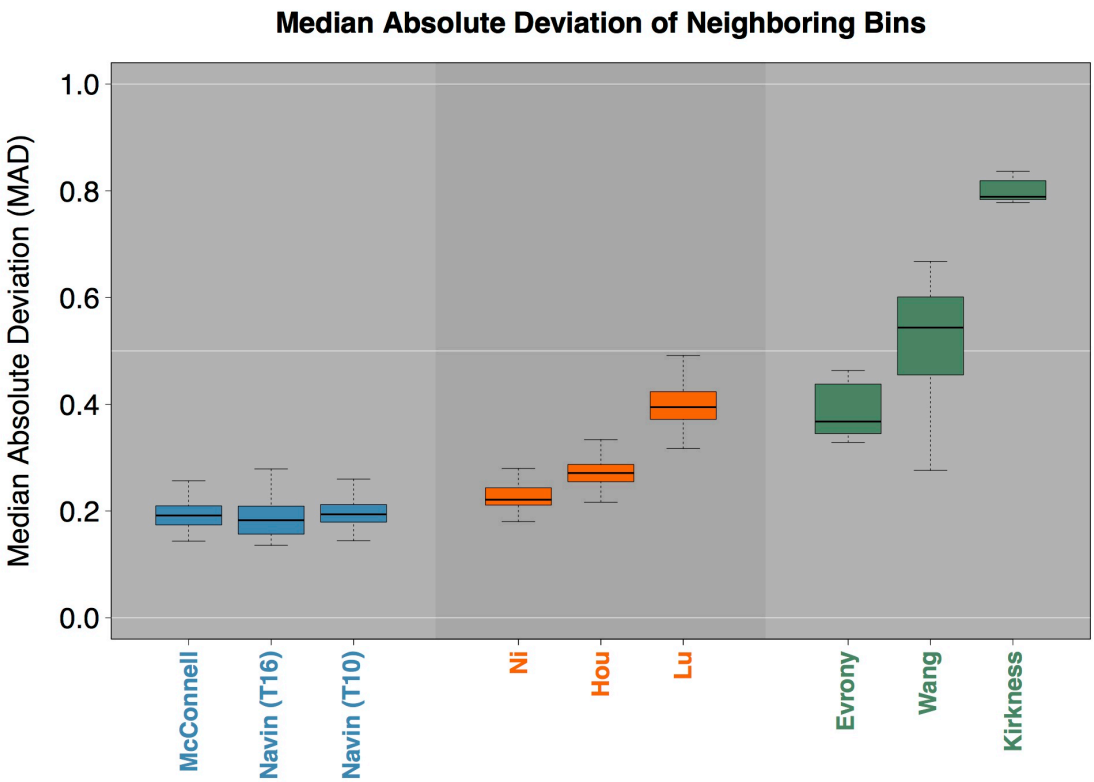


Figure 3 | The median absolute deviation (MAD) of neighboring bins: A single pair-wise MAD value is generated for each sample in a given dataset and represented by a box and whisker plot. The high biases present in the MDA datasets make comparing DOP-PCR and MABLAC samples difficult. Figure 4 shows this comparison more clearly.

Detailed comparison of MALBAC and DOP-PCR protocols

Whole-genome amplification using MDA introduces a large degree of biases compared to MALBAC or DOP-PCR, limiting its applicability to CNV analysis. As such, we focus the scope of the remaining comparisons on the latter two WGA techniques. For a fine-grained comparison of MALBAC and DOP-PCR, we compare the T10 dataset from Navin *et al*. and the CTC dataset from Ni *et al*. due to their similar biological and technical conditions and similar published analysis. Both datasets contain aneuploid cancer cells, were sequenced to similar depth (CTC mean read count: 4,133,466; T10 mean read count: 6,706,119), and were used to generate phylogenetic clusters of samples based on CNVs. We begin by comparing the coverage dispersion and investigate the minimum coverage and bin size needed to reproduce the published results.

Coverage dispersion

Using the MAD criteria described above, the DOP-PCR-based T10 dataset shows markedly better bin-to-bin correlation than the MALBAC-based CTC dataset as judged by a lower MAD of adjacent and offset bin counts (**Figure 4**). For adjacent bins, the first quartile of the CTC MAD comparison (black) is higher than the third quartile of the T10 MAD comparison (red). As we increase the bin offset, greater variation is seen in the CTC data as show by the separation of the mean MAD between the T10 and CTC datasets. We interpret this to mean that there is more local trending in amplification efficiency in MALBAC than in DOP-PCR data.

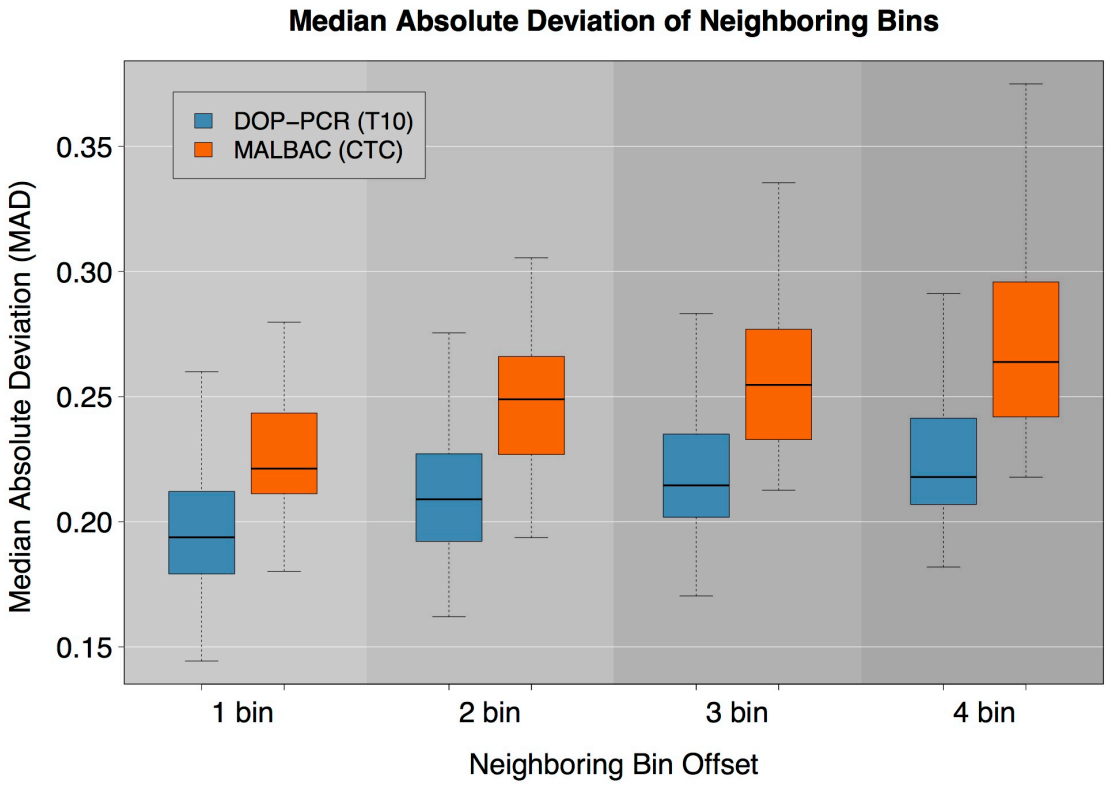


Figure 4 | A comparison of MAD between the Navin *et al.* (T10) shown in blue and Ni *et al.* (CTC) shown in orange. As the bin offset increases the separation between the mean T10 MAD and mean CTC MAD grows.

Minimum coverage requirement

We next explore whether WGA protocols differ with respect to the minimum coverage required to observe the same population/clonal substructure identified at full coverage. To this end, we down-sample all datasets and analyze each in Ginkgo to determine: (1) how well segment breakpoints are conserved and (2) how well the phylogenetic relationships are maintained. With all degrees of downsampling (from 25% to 99%), the T10 data shows better breakpoint conservation than the CTC data, but as expected, increased degrees of downsampling lead to substantial erosion of breakpoint boundaries in both datasets (see **Supplementary Figure 4**).

Nevertheless, these downsampling experiments prove MALBAC and DOP-PCR remarkably robust with respect to preserving the overall clonal/population structure, even at extremely low coverage. The clonal structure of the T10 dataset remains fully intact across all downsampling experiments even as the mapped reads are downsampled by 99% (from ~608 reads/bin to ~6 reads/bin). The population structure of the CTC dataset is preserved when downsampled by 95% (from ~597 reads/bin to ~30 reads/bin); when downsampled to 99%, one cell from one patient is incorrectly clustered.

Although depth of coverage in both studies was originally very low (< 0.15x), our downsampling results indicate that Ginkgo can correctly determine the phylogenetic relationship between samples even when sequenced to a depth of coverage of only 0.01x. If generally applicable, which we have not proven here, this approach will allow sparser sequencing with higher throughput at equivalent cost. After low-coverage sequencing, a number of cells from the same phylogenic branch can be pooled for deeper sequencing if desired.

Optimizing bin sizes

Bin size directly impacts the resolution at which CNVs can be called [16]. Thus far we used 500kb-bins to reproduce the results of Navin *et al*. and Ni *et al*. following the procedure by Ni *et al*. However, such large bin sizes hinder the identification of smaller copy-number events. To identify the minimum bin size needed to reproduce the published results, we decreased bin size from 500kb to 10kb (**Supplementary Table 1)** for both datasets until the hierarchical clustering of the copy number profiles produced different results.

The T10 dataset retained its hierarchical structure until bin sizes dropped below 25kb (**Supplementary Figure 5**), while the CTC dataset lost it original hierarchical structure at a bin size of 100kb. In the T10 dataset, when bin sizes drop to 10 kb, a few hypodiploid cells incorrectly cluster. In the CTC dataset, as bin sizes approach 100kb, cells from two patients (4 and 7) begin to overlap. Using 50kb bins, there is widespread overlap between nearly all patients' cells, and only the cells from two patient cluster correctly (**Supplementary Figure 6**). This indicates that at the same level of coverage, DOP-PCR can resolve smaller CNVs than MALBAC, but more comparably structured studies are needed.

Detecting integer copy-number states

Preliminary analysis of bin counts indicate that at the same level of coverage, MALBAC data had a higher level of coverage dispersion and therefore a worse signal to noise ratio than DOP-PCR data. Our downsampling experiments support this claim as the ability to properly discriminate between CTC patients based on the CN states is lost at a bin resolution that is easily resolved with the T10 dataset. To understand the effects of noise further, we evaluated each dataset to discriminate distinct copy number states.

Because the copy-number states of individual cells are integer, we expect the data to be centered at integer values. If the data is highly uniform, read coverage per bin will tightly surround integer copy-number states. As bin count dispersion around copy-number states increases, or is influenced by local chromosomal trends, the distinction between copy-number states will blur.

To examine this, we generated a histogram of the normalized read count distribution for the CTC and T10 datasets (**Supplementary Figure 7**). **Figure 5** shows distributions of bin counts for representative cells: excellent, typical, and lower quality cells as well as the highest quality population average. All T10 profiles have distinct peaks representative of integer copy-number values. While there are a few cells in the CTC dataset that have distinct peaks, many of the CTC profiles have considerably worse resolution with substantial blurring between CN states. Furthermore, the scaled read count distributions illustrate the substantial difference in signal-to-noise between the T10 and CTC datasets (**Supplementary Figure 8**).

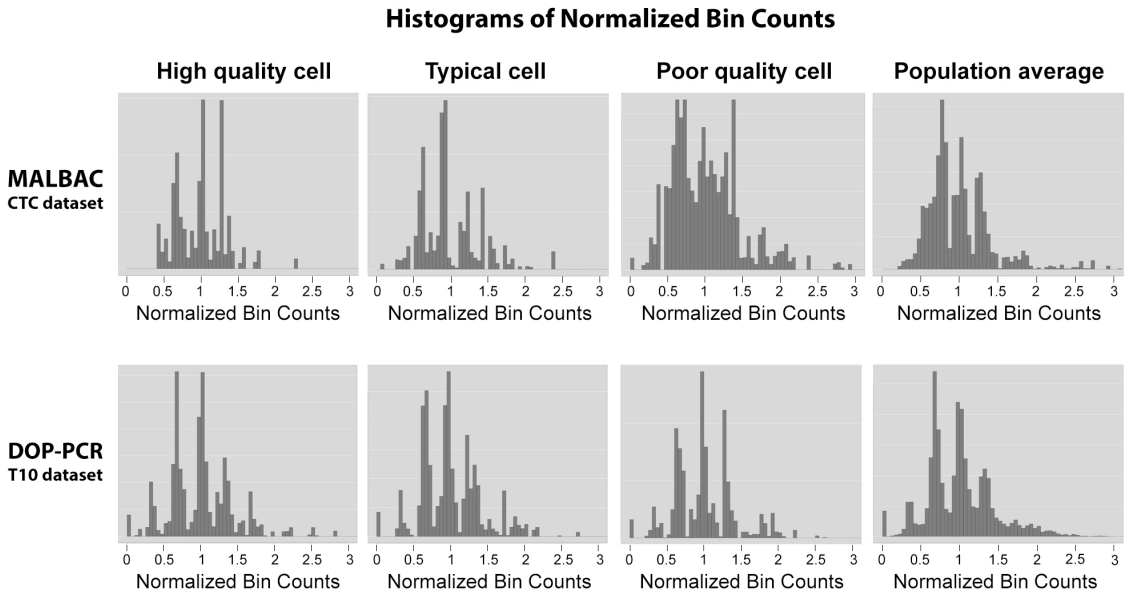


Figure 5 | Histograms of normalized bin counts across the CTC and T10 datasets, for a high-, typical-, and poor-quality cell. The rightmost column contains histograms of high quality cell population averages. Distinct peaks are representative of clean data from which accurate copy number calls can be made.

Options and limits in processing pipeline

In addition to diverse experimental methods, single-cell sequencing is subject to varying computational methods such as the choice of mapping parameters, binning methods, normalization, and segmentation (**Supplementary Table 1**). Ginkgo offers several tunable parameters and we provide some guidance below by comparing modalities. Our pipeline is expanding, and we expect to add additional tools in the future.

Mappability: fixed vs. variable length bins

As previously demonstrated [10], fixed-length bins introduce mappability issues in highly repetitive regions such as centromeres. These regions are often "dead zones"

that remain empty even if the overall coverage is high, and will bias segmentation algorithms into identifying nonexistent breakpoints. While Ginkgo supports normalizing copy-number profiles by a diploid cell, this can introduce noise in bins with low counts. We recommend using a variable-binning strategy in which variable-length bins with equal mappability are used, allowing uniform mapping of reads. When neglected, like in Ni *et al.*, dead zones are present at chromosome boundaries and centromeres. In comparison, the same profiles generated by Ginkgo are less noisy and free of dead zones (see Ginkgo website). We note, however, that in using the variable-length bin strategy, certain regions of the genome, specifically near the centromeres of several chromosomes, have consistently higher read depth than expected whether profiling bulk DNA or single cells. This problem also occurs in fixed bins but is less severe due to the substantial read drop out caused by using fixed intervals. These peaks may be the result of occasional mismapping of highly repetitive sequences from elsewhere in the genome to unique but similar sequences located near the peri-centromeric DNA. This is also likely influenced by imperfect reference assemblies that do not fully represent the correct genomic sequence. We have termed these "bad bins" and have provided an option in Ginkgo to mask them for human (hg19) for simplicity of presentation (see **Online Methods**).

GC correction

As presented earlier, when GC content in a genomic region falls outside a certain range, read counts rapidly decrease. Ginkgo automatically corrects for GC biases in all cells and references to ensure uniform average read counts at all GC values.

Segmentation and copy-number calls

Two common approaches for segmenting bin data into copy-number profiles are Circular Binary Segmentation (CBS) and Hidden Markov Models (HMMs). The HMM approach requires a specified number of states, each parameterized, and a transition matrix, for copy-number calls. For this reason, we have chosen not to implement this in our initial version of Ginkgo, but offer CBS with its typical choice of parameters (see **Methods**).

Other options

Ginkgo offers additional options. Ploidy estimated from fluorescence intensity during cell sorting can be used to infer copy number instead of our quantal copy number algorithm. There are several options for building phylogenies, with the option to mask sex chromosomes for mixed-gender samples. Moreover, there is an interactive viewer for inspecting copy-number profiles at varying levels of resolution. Finally, quality control parameters are accessible across each step of the pipeline. Ginkgo supports data from human, chimpanzee, mouse, rat and fruit fly.

Discussion

Single-cell sequencing has tremendous potential to shed light on genetically complex environments. Early applications have already discovered surprisingly high levels of heterogeneity and copy-number mosaicism in tumors, neurons, and sperm. The implications are profound and provide a new computational lens to observe, for example, the founding cancer cell population and trace its development through a tumor and metastasis.

Many projects are now underway to apply the technology to diverse tissue and cell types. The experimental protocols are maturing, and with Ginkgo, a validated, open-source end-to-end pipeline is now available for researchers as well. The interactive visual analytics environment provides researchers with an intuitive platform to explore and understand their population of cells. It begins with a high-level overview of the population represented by dendrograms and heatmaps. It then gives researchers the ability to zoom in on the copy-number profile of individual cells, filter the analysis for subsets of cells, and inspect the details of copy-number state or read depth on demand. Throughout each stage of the analysis, statistical summaries and quality metrics guide researchers to the most significant and most accurate data. Figures and data tables are available for download to be embedded into presentations or publications.

Using Ginkgo's automation and quality-assessment tools, we examined 9 single-cell datasets to compare the data quality obtained by 3 experimental WGA techniques. Based on our analysis it appears that DOP-PCR provides more reliable copy number profiles as MALBAC and MDA's poor coverage uniformity tends to obscure the underlying copy number landscape. Furthermore, for copy-number analysis we find that DOP-PCR shows higher coverage uniformity, lower coverage dispersion, less GC bias, and higher resolution than the other techniques.

Finally, to guide researchers in their analysis, we highlight common computational pitfalls in single-cell analysis and discuss how Ginkgo corrects for them. Correctly calling CNVs from single-cell sequencing data is still an open problem and Ginkgo's wide array of parameters leaves it flexible to users' needs while remaining robust. As single-cell sequencing methods further develop, we anticipate it will become practical to analyze SNPs and other smaller mutations reliably. As this occurs, and as other algorithmic improvements are made, we will incorporate those new ideas into our toolbox. We are also exploring ideas for the analysis and visualization of single-cell RNA sequencing as those protocols become more widely available. Users are encouraged to customize and contribute back to Ginkgo's open-source code base hosted on GitHub.

Acknowledgements

We would like to thank Nicholas Navin and Peter Andrews for their helpful discussions and assisting getting access to the data. The project was supported in part by National Institutes of Health award (R01-HG006677) to MCS, the National Science Foundation (DBI-1350041) to MCS, the CSHL Cancer Center Support Grant (5P30CA045508), and by the Watson School of Biological Sciences through a Training Grant (5T32GM065094) from the National Institutes of Health.

Author Contributions

T.G. and R.A. developed the software and conducted the computational experiments. M.C.S, M.W., J.H., and G.S.A. designed the experiments. T.B. and J.K. assisted with the analysis and helped design the experiments. All of the authors wrote and edited the manuscript. All of the authors have read and approved the final manuscript.

References

1. Shapiro, E., T. Biezuner, and S. Linnarsson, *Single-cell sequencing-based technologies will revolutionize whole-organism science.* Nature reviews. Genetics, 2013. **14**(9): p. 618-30.
2. Wigler, M., *Broad applications of single-cell nucleic acid analysis in biomedical research.* Genome medicine, 2012. **4**(10): p. 79.
3. McConnell, M.J., et al., *Mosaic copy number variation in human neurons.* Science, 2013. **342**(6158): p. 632-7.
4. Blainey, P.C., *The future is now: single-cell genomics of bacteria and archaea.* FEMS microbiology reviews, 2013. **37**(3): p. 407-27.
5. Wang, J., et al., *Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm.* Cell, 2012. **150**(2): p. 402-12.
6. Gundry, M., et al., *Direct, genome-wide assessment of DNA mutations in single cells.* Nucleic Acids Research, 2012. **40**(5): p. 2032-40.
7. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.
8. Ni, X.H., et al., *Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(52): p. 21083-21088.
9. Lu, S., et al., *Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing.* Science, 2012. **338**(6114): p. 1627-30.

10. Navin, N., et al., *Inferring tumor progression from genomic heterogeneity.* Genome research, 2010. **20**(1): p. 68-80.

11. Henrichsen, C.N., E. Chaignat, and A. Reymond, *Copy number variants, diseases and gene expression.* Human molecular genetics, 2009. **18**(R1): p. R1-8.

12. Pinard, R., et al., *Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.* BMC Genomics, 2006. **7**: p. 216.

13. Zong, C., et al., *Genome-wide detection of single-nucleotide and copy-number variations of a single human cell.* Science, 2012. **338**(6114): p. 1622-6.

14. Navin, N.E., *Cancer genomics: one cell at a time.* Genome Biology, 2014. **15**(8): p. 452.

15. Magi, A., et al., *Read count approach for DNA copy number variants detection.* Bioinformatics, 2012. **28**(4): p. 470-8.

16. Baslan, T., et al., *Genome-wide copy number analysis of single cells.* Nature protocols, 2012. **7**(6): p. 1024-41.

17. Hou, Y., et al., *Genome analyses of single human oocytes.* Cell, 2013. **155**(7): p. 1492-506.

18. Kirkness, E.F., et al., *Sequencing of isolated sperm cells for direct haplotyping of a human genome.* Genome research, 2013. **23**(5): p. 826-32.

19. Evrony, G.D., et al., *Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain.* Cell, 2012. **151**(3): p. 483-96.

20. Smits, S.A. and C.C. Ouverney, *jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web.* PLoS One, 2010. **5**(8): p. e12267.

21. Ross, M.G., et al., *Characterizing and measuring bias in sequence data.* Genome Biology, 2013. **14**(5): p. R51.

22. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

23. Cleveland, W.S., *Lowess - a Program for Smoothing Scatterplots by Robust Locally Weighted Regression.* American Statistician, 1981. **35**(1): p. 54-54.

24. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. **5**(4): p. 557-72.

25. Dago, Angel E., et al. *Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells*. PloS One 9.8 (2014): e101777.

26. Meyer, Laurence R., et al. *The UCSC Genome Browser database: extensions and updates 2013.* Nucleic acids research 41.D1 (2013): D64-D69.

Online Methods

Mapping reads to genome

Reads were mapped to hg19 using bowtie [22] and only uniquely mapped reads (mapping quality score >= 25) were kept. Mapped read counts ranged from 1,538,234 (Ni et al.) to 30,638,853 (Lu et al.) with a mean of 15,827,886. To perform an unbiased comparison, all samples were randomly downsampled to 1,538,234 reads to match the lowest coverage.

Binning reads

Copy number analysis begins with binning uniquely mapping reads into fixed-length or variable-length intervals across the genome. This aggregates read depth information into larger regions that are more robust to variable amplification and other biases. As discussed in the main text, fixed-length bins are generally discouraged as they lead to read drop out in regions that span highly repetitive regions, centromeres, and other complex genomic regions.

To generate boundaries for variable-length bins, we use the method outlined in [7], where we sample 101bp stretches of the reference assembly at every position along the genome. These simulated reads are mapped back to the genome using Bowtie and only uniquely mapping reads are analyzed. For a given bin size, we assign reads into bins such that each bin has the same number of uniquely-mappable reads. Consequently, intervals with higher repeat content and low mappability will be larger than intervals with highly mappable sequences, although they will both have the same number of uniquely mappable positions.

Using variable-length bins with sufficient depth of coverage and consistent ploidy, sequence reads are expected to map evenly across the entire genome with uniform variance. Users are provided with a variety of bin sizes from which to choose, depending on the overall coverage available; if the mean coverage per bin is too low, we encourage users to use larger bins.

GC bias correction

Once reads are placed into bins, Ginkgo normalizes each sample and corrects for GC biases prior to segmentation. The normalization process begins by dividing the count in each bin by the mean read count across all bins. This centers the bin counts of all samples at 1.0. To identify and correct GC biases, Ginkgo computes a locally-weighted linear regression using the R function *lowess* [23] (smoother span = .5, iterations = 3, delta=0.1*range(x)) to model the relationship between GC content and log-normalized bin counts. This lowess fit is then used to scale each bin such that the expected average log-normalized bin count across all GC values is zero.

After the lowess fit, we monitor the bias of each cell by calculating the proportion of bins that fall outside an expected coverage of zero by +/- 1, log base 2.

Segmentation (CBS)

Following GC bias correction, bin counts are segmented to reduce fluctuations in noise across chromosomes and identify longer regions of equal copy number. To this end, Ginkgo makes use of Circular Binary Segmentation (CBS) [24], which segments the genome by recursively splitting the chromosomes into segments based on a maximum t-statistic until a reference distribution estimated by permutation is reached. Once the CBS segmentation is complete, the breakpoints (segment boundaries) across all bins are determined, and the counts for all bins within each segment are reset to be the median bin count value within that segment.

The key step during segmentation is selecting the right reference sample for comparison. Using a diploid sample to normalize bin counts can eliminate additional biases uncorrected by GC normalization. Although Ginkgo supports uploading data from such a cell, this is not always available so Ginkgo provides alternatives for segmenting samples: (1) Independent segmentation, where samples are segmented independently by their own normalized bin count profiles; and (2) Sample with lowest IOD, where Ginkgo selects the sample with the lowest index of dispersion (IOD - the ratio between the read coverage variance and the mean) and uses that sample as a reference for all other samples. The sample with the lowest index of dispersion will likely be among the most evenly balanced ploidy and highest quality of all submitted cells.

Determining copy-number state

Since we are analyzing single-cell data, we expect every genomic locus to have an integer copy number (CN) value. Furthermore, the quantized nature of single-cell data means that the same number of reads per bin should separate every sequential CN state, e.g., ~50 reads for CN 1, ~100 reads for CN 2, ~150 reads for CN 3, etc. While biological and technical noise prevent read counts from segregating perfectly into distinct CN states, read counts should still be centered around integer CN states.

The most direct approach for determining the CN state of each cell is available for users that have *a priori* knowledge of the ploidy of each sample. For example, cells that are DAPI-stained prior to cell sorting can be gated based on their fluorescence activity, and ploidy can be determined by comparing its fluorescence activity to that of a reference cell with a known CN state. With these data, Ginkgo determines the copy number state of each sample by scaling the segmented bin counts such that the mean bin count is equal to the ploidy of the sample. Finally bin counts are rounded to integer copy number values. Advances in fluorescence activated cell sorting (FACS) will make this copy number prediction even more

accurate in time, although cells that are incorrectly sorted and placed into wells with more than one cell will show much higher fluorescence activity and will have an incorrectly inferred copy number state.

Since FACS data is not always available for analysis and has potential for error, Ginkgo provides an alternative to determine the copy number of each sample. As discussed earlier, before determining the CN state of a cell, the cell is binned, normalized, and segmented. This copy number profile with a mean of one is referred to as the raw copy number profile (RCNP). If the true genome-wide copy number of a sample were equal to X, the scaled copy number profile (SCNP) would then be the product of RCNP and X, and the final *integer* copy number profile (FCNP) would be the rounded value of the SCNP so all segments contain an integer value.

With these relationships, Ginkgo infers the genome-wide copy number X using numerical optimization. For a given cell, Ginkgo first determines the SCNP and FCNP for all possible values of X in the set [1.50, 1.55, 1.60, … , 5.90, 5.95, 6.00]. Ginkgo then computes the sum of square (SoS) error between the SCNP and the RCNP for each value of X and selects the value of X with the smallest SoS error. Once the multiplier is identified and applied, the scaled bins are rounded to generate the final integer copy number profile for each sample. Intuitively, this is equivalent to finding the copy number multiplier that causes the normalized segmented bin counts to best align with integer copy number values.

Clustering

Before visualization, the final step is to look outside the scope of individual cells and determine the overall population structure. Ginkgo first determines the distance (dissimilarity structure) between all cells. We provide six choices of distance metrics: Euclidean, maximum, Manhattan, Canberra, binary, and Minkowski. After computing the dissimilarity matrix, Ginkgo then computes a dendrogram by hierarchically clustering samples using one of four different agglomeration methods: single linkage, complete linkage, average linkage, and ward linkage. In addition, Ginkgo generates a phylogenetic tree by first computing the Pearson correlation between all samples and using these dissimilarity values to cluster the samples.

Masking bad bins

As described in the main text, there are a number of regions, specifically around the centromeres of certain chromosomes, where there is an accumulation of very high read depth compared to the expected depth. These regions consistently display high read depth in both bulk and single-cell sequencing data. Using data from 54 normal individual diploid cells from multiple individuals (not presented here), these bins (designated as “bad bins”) were determined in the human reference genome (hg19)

as follows. The bin counts were divided by the mean bin counts for each cell to normalize for differences between cells in total read count. For each chromosome, the mean of the bins over all cells is subtracted from each individual cell's normalized bin count to normalize for differences between chromosomes. The mean and standard deviation of the autosomes is then used to compute an outlier threshold corresponding to a p-value of 1/N, where N is the number of bins used. In practice, less than 1% of bins are identified as extreme outliers and masked for further processing.

Supplementary Notes and Figures

Supplementary Note 1: Replication of the single-cell analysis results

1. Navin et al.

This work profiled breast cancer in two separate studies. The first (dataset T10) examined heterogeneity in a polygenomic breast tumor. CNV analysis and hierarchical clustering of 100 single-cells revealed three distinct clonal subpopulations present in the tumor. The second study (datasets T16M/P) examined a monogenomic breast tumor and its suspected liver metastasis. CNV analysis and hierarchical clustering of 100 cells revealed that a single clonal expansion formed the primary breast tumor and seeded the metastasis. In the polygenomic breast tumor analysis, Ginkgo clusters all 100 samples into the same four distinct subpopulations of the original study, replicating the published population structure (**Supplementary Figure 1A**). In the monogenic breast tumor and its associated liver metastasis analysis, Ginkgo clusters all 100 samples into the same three distinct subpopulations as the original publication, linking the primary tumor to its metastasis (**Supplementary Figure 1B**).

2. McConnell et al.

This study profiled CNV events in human hiPSC-derived fibroblasts and 110 frontal cortex neurons. McConnell et al. found a wide degree of mosaic copy-number variation in neurons and discovered that a subset of neurons have highly aberrant genomes. Using Ginkgo, we successfully reproduce the major aneuploid events in all frontal cortex neurons.

3. Ni et al.

This study explored SNPs and CNVs in circulating tumor cells (CTCs) in patients with lung cancer. Through CNV analysis and hierarchical clustering of 29 CTCs across 7 patients with lung adenocarcinoma (ADC) or small-cell lung cancer (SCLC) the authors discovered that CNVs appear specific to cancer types and are reproducible from cell to cell and from patient to patient. Using default settings in Ginkgo, we generate similar CN profiles for all 29 samples and can reproduce the published clustering results (**Supplementary Figure 2**).

4. Hou et al.

This study sequenced the triads of first and second polar bodies (PB1 and PB2) and the oocyte pronuclei from same female egg donors to phase these genomes and determine their crossover maps and frequency. Additionally, genome-wide CN profiles were generated to explore aneuploidy in each sample. The authors identified a total of 47 CNVs in 25 aneuploid cells across 5 patients. We could replicate these results as Ginkgo uncovered 45 of the 47 CNVs in 23 of the 25 identified aneuploid cells. One sample, S0808 (containing the missing two cells/CNVs), did not have CN events matching the published results. We believe this was due to accidental mislabeling of samples/sample IDs upon being deposited to NCBI.

5. Lu et al.

In this study, single-cell sequencing of 99 sperm cells from an Asian male were used to examine meiotic recombination and aneuploidy. Across the 99 samples, the authors uncovered 5 aneuploid cells. As expected, our CNV analysis and hierarchical clustering using Ginkgo was able to separate the X and Y bearing chromosomes with the exception of two cells with extremely poor coverage uniformity and a high degree of read drop out that clustered separately (**Supplementary Figure 8**). In addition, we could cleanly replicate the CNV results as Ginkgo uncovered the same chromosomal aberrations in the five aneuploid cells as the original study.

6. Kirkness et al.

This work profiled genomic variants in sperm with the goal of demonstrating a technique to retrieve complete haplotypes of sequenced genomes, information that is averaged out and lost during conventional genome sequencing approaches. Using genotype data from 16 cells and low coverage single-cell sequencing of 96 sperm cells, the authors leverage the haploid nature of sperm to identify recombination events at a median resolution less than 100 kb.

7. Wang et al.

Much like *Kirkness et al.*, this work examined meiotic recombination and de novo mutations in sperm. Using a microfluidic system, the authors carried out single-cell sequencing on 91 sperm cells to generate a personal recombination map. This map was confirmed by the low-resolution afforded by bulk sequencing but individual samples were found to have significant differences from pedigree data at higher resolution. Finally, the authors used these results to test for meiotic drive, gene

conversion, and genome instability. Deep sequencing of 8 single-cells revealed additional unique de novo mutations.

8. Evrony et al.

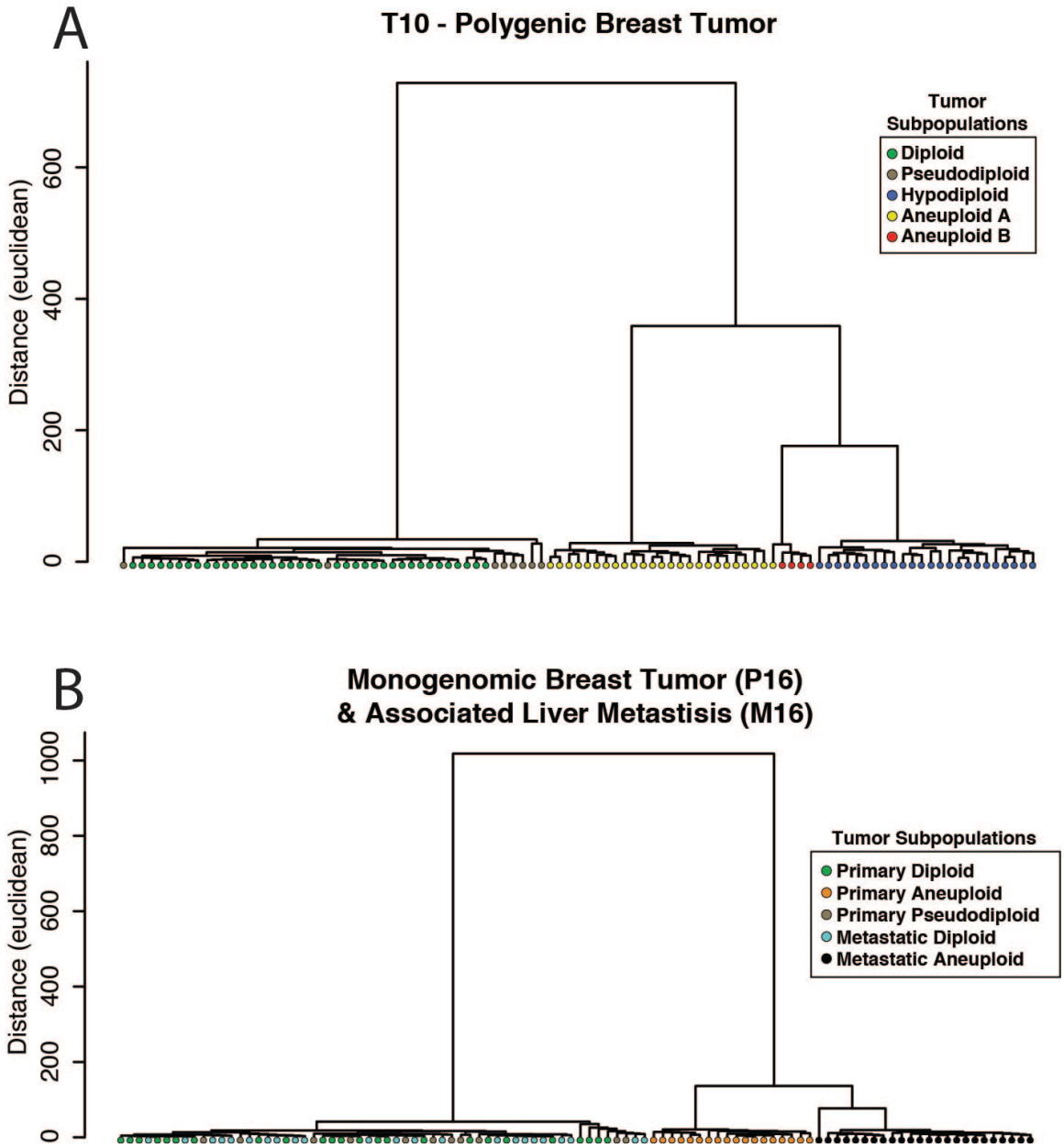
This work looked to unravel to what extent genetic mosaicism exists in the brain of an individual. To this end the authors sequenced 300 single neurons from the cerebral cortex and caudate nucleus of three normal individuals. L1 insertion profiling led to estimate that there are less than < 0.6 unique somatic L1 insertions per neuron with most neurons (~80%) lacking any detectable unique somatic mutations. Finally, the authors genotyped single cortical cells of a child with hemimegalencephaly to characterize the mosaicism of a somatic AKT3 mutation. The AKT3 mutation was found in both neuronal and nonneuronal cells indicating that the mutation occurred in nueroglial progenitor

Supplementary Tables

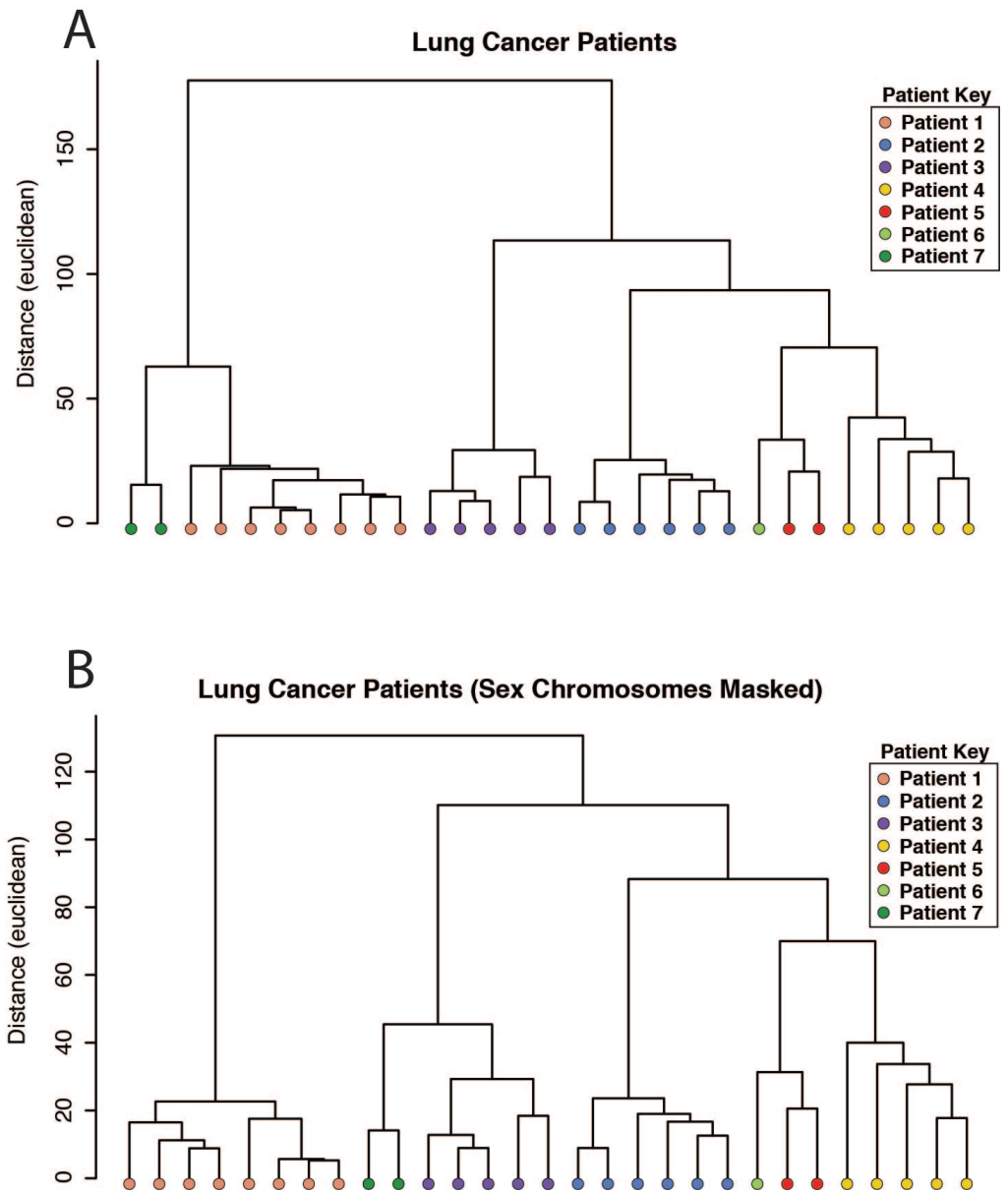
Study	WGA Method
Genome	- Human *(hg19, hg18)* - Chimpanzee *(panTro4, panTro4)* - Mouse *(mm10,mm9)* - Rat *(rn5)* - Drosophila *(dm3)*
Binning method	- Variable-size bins - Fixed-size bins
Bin sizes	- 10kb - 25kb - 50kb - 100kb - 175 kb - 250kb - 500kb
Segmentation method	- Independent *(using normalized read counts)* - Global *(using sample with lowest IOD)* - Custom *(user specifies segmentation .bed file)*
Clustering method	- Single - Complete - Average - Ward - Neighbor joining
Clustering distance metric	- Euclidian - Maximum - Manhattan - Canberra - Binary - Minkowski
Masking "bad bins"	- Due to incomplete/incorrect reference assemblies a small number of bins near telomeric and centromeric regions show consistently inflated copy number states across all cells, healthy or otherwise. This option removes these artificial amplifications from the analysis. *Currently only available for human.
Sex chromosome masking	- Binary option (whether or not to mask sex chromosomes during clustering)
FACS copy number estimation	- User uploads a file with ploidy estimates from FACS results that are used in tandem with our copy number estimation algorithm (not required).

Supplementary Table 1: Data analysis parameters in Ginkgo

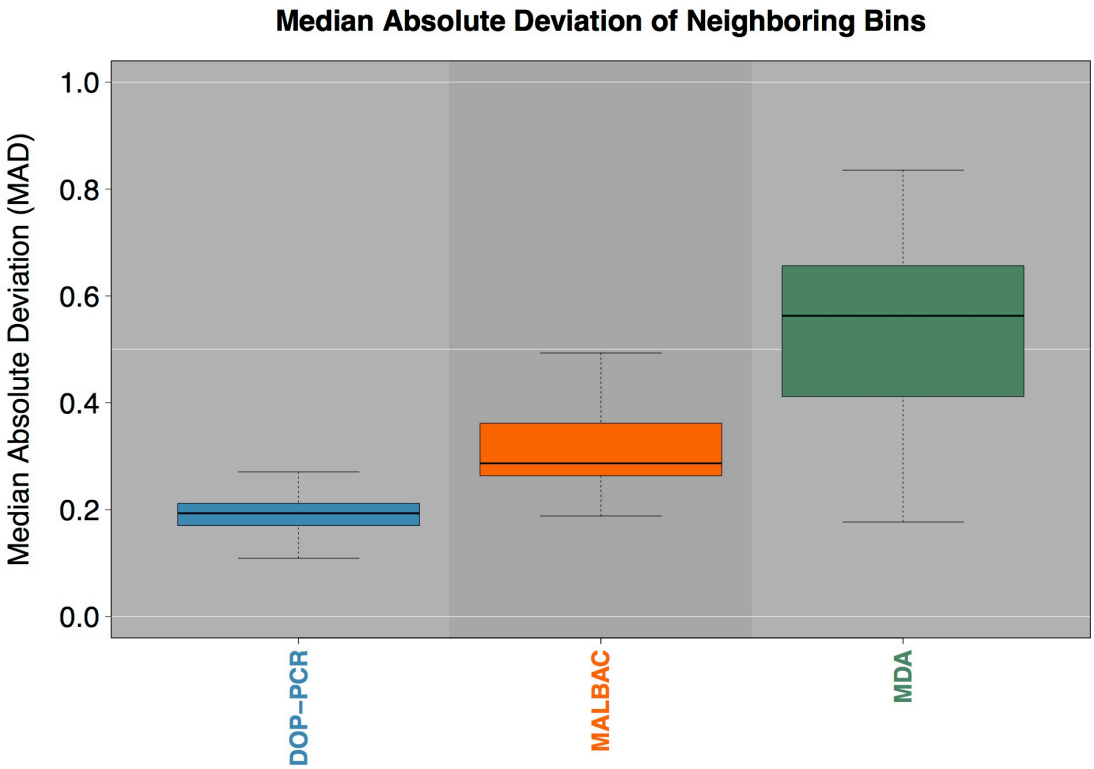
Supplementary Figures



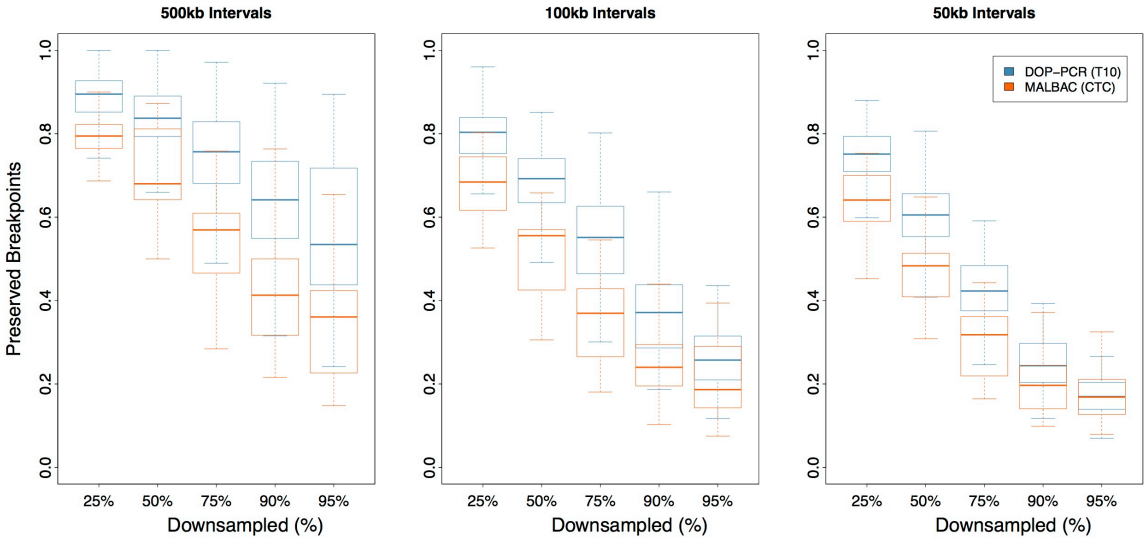
Supplementary Figure 1 | Ginkgo successfully reproduces the clonal structure determined by Navin et al.: Phylogenetic trees generated through hierarchical clustering by copy-number using (A) 100 polygenomic breast tumor samples (T10) and (B) 52 monogenomic breast tumor (T16P) and 48 liver metastasis (T16M) samples.



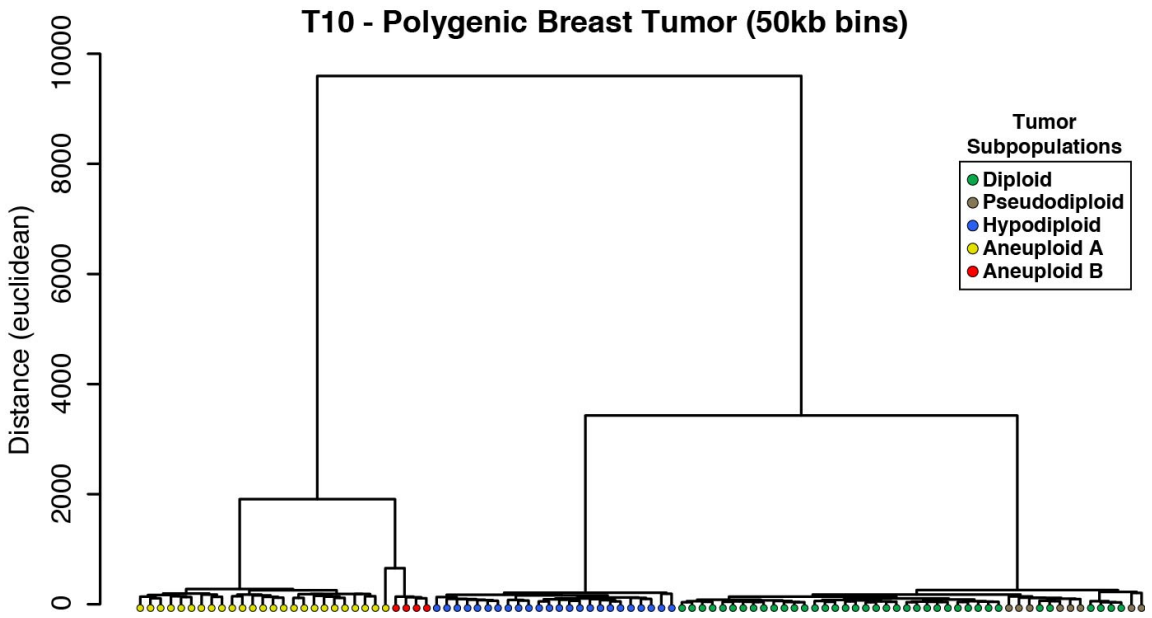
Supplementary Figure 2 | (A) Ginkgo correctly reproduces the patient clustering results of Ni et al. Hierarchical clustering of 29 samples derived from 7 different patients with either adenocarcinoma (patients 2-6) or small cell lung cancer (patients 1, 7). (B) When sex chromosomes are masked, cells still cluster by patient, but patients no longer cluster by cancer subtype. In particular, after masking sex chromosomes, patient 3 is intermixed between patients 1 and 7 and there is no clear association between cancer types.



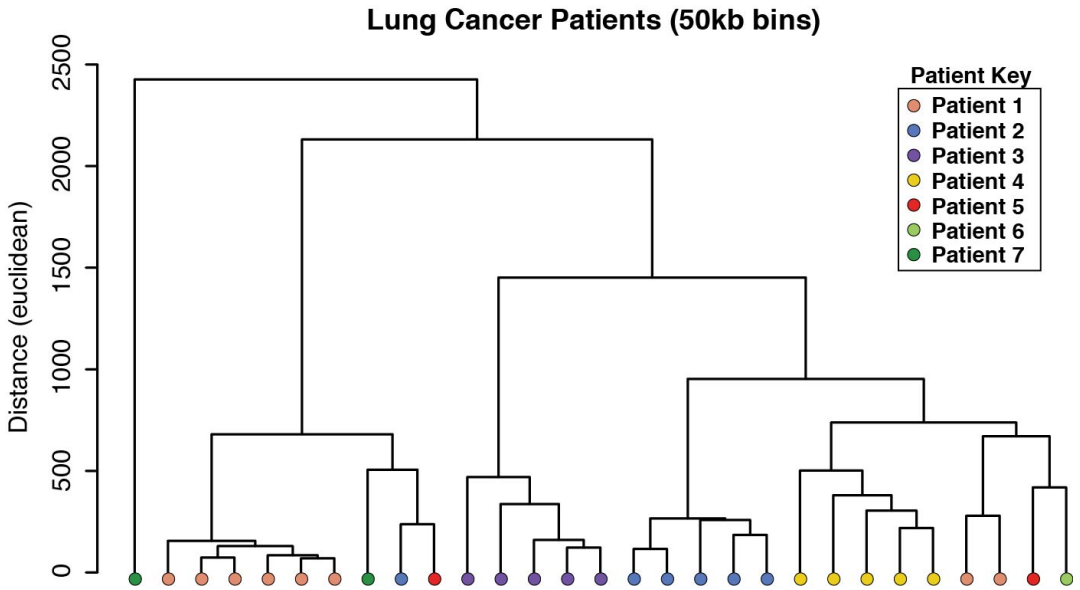
Supplementary Figure 3 | The median absolute deviation (MAD) of neighboring bins: A single pair-wise MAD value is generated for each sample in a given dataset and represented by a box and whisker plot. The DOP-PCR datasets show the lowest mean MDA as well as the lowest variance across samples. While certain MDA samples outperform the MALBAC dataset, they show much large variability in data quality than MALBAC.



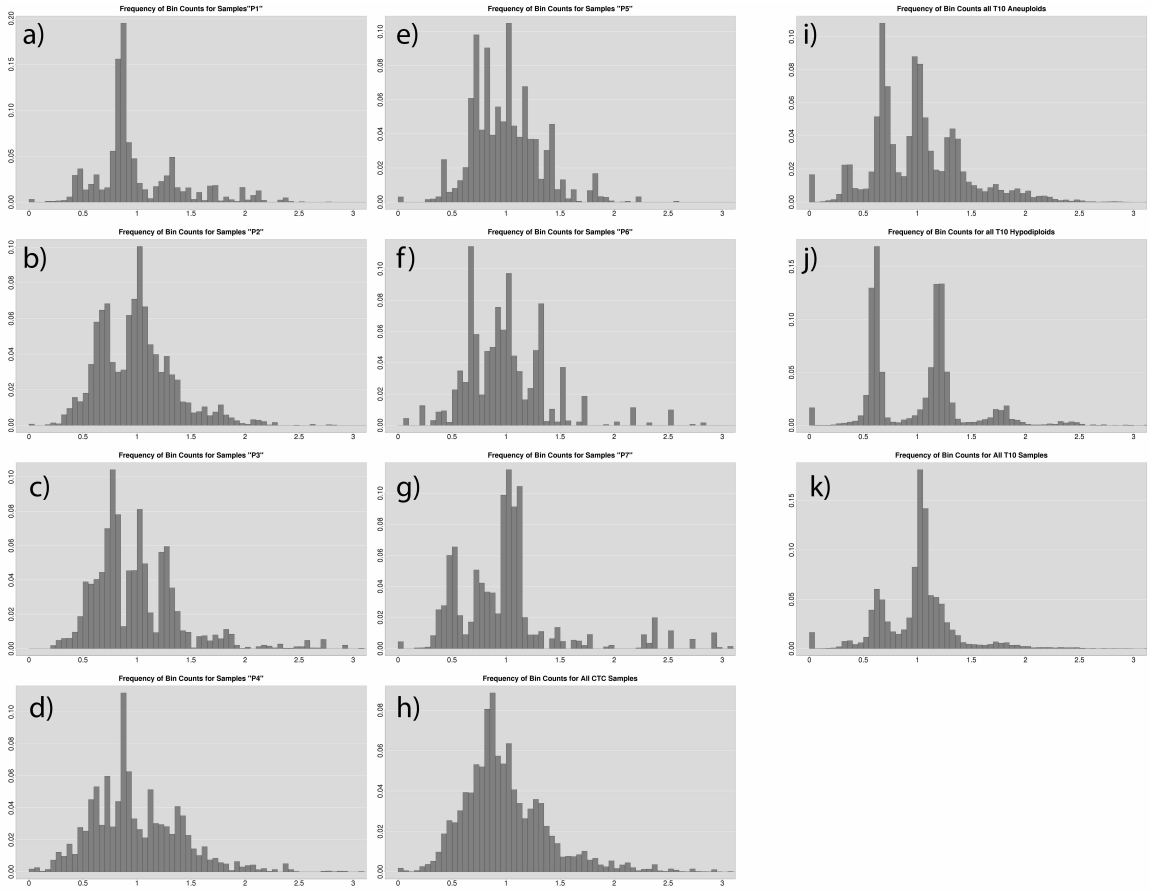
Supplementary Figure 4 | Comparing breakpoint conservation between T10 (blue) and CTC (orange). The fraction of breakpoints conserved between the full intact dataset and the dataset downsampled 25%, 50%, 75%, 90%, or 95% =using interval sizes of **(A)** 500kb, **(B)** 100kb, and **(C)** 50kb. At all levels of downsampling and all intervals sizes, the T10 DOP-PCR data retains a larger fraction of breakpoints than the CTC MALBAC data.



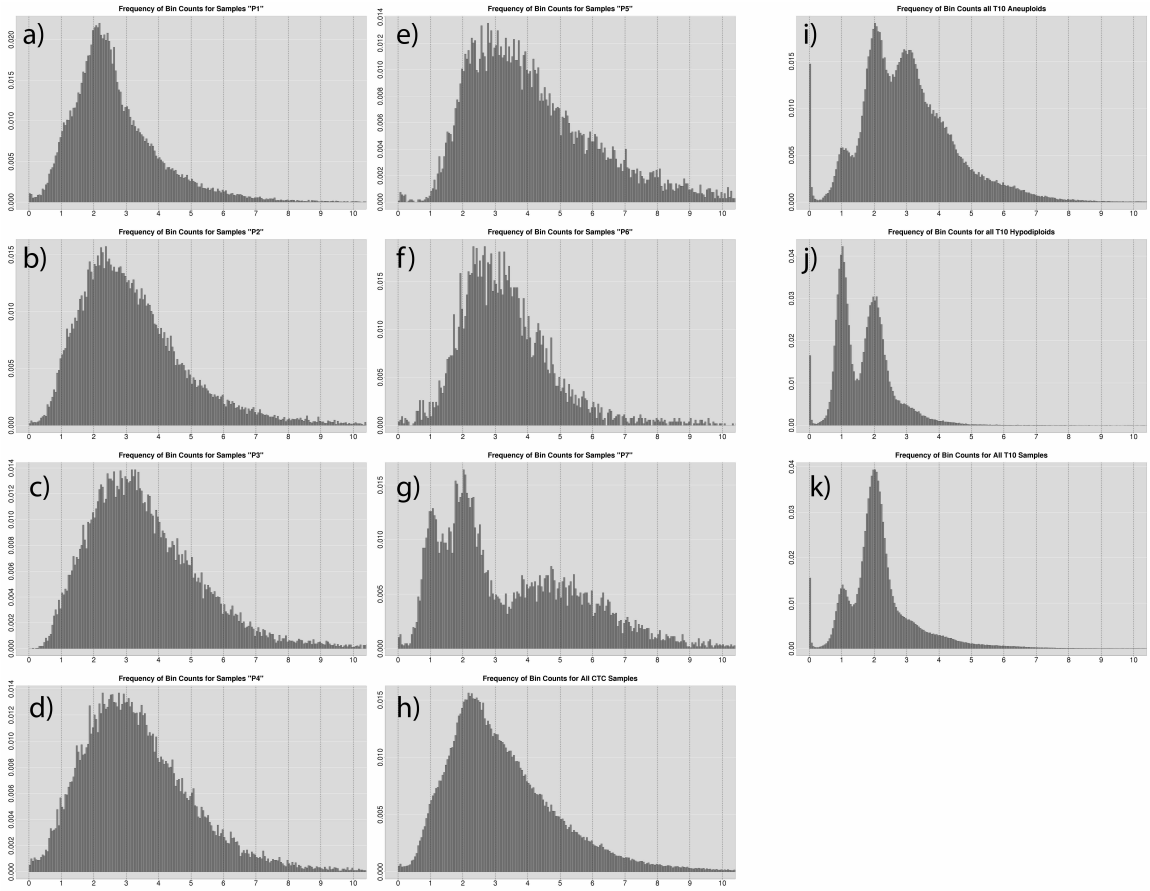
Supplementary Figure 5 | A phylogenetic tree of the T10 cell population at a bin size of 50kb built using the copy number profiles of individual cells. All of the clonal populations remain intact and represent the 500kb clustering results.



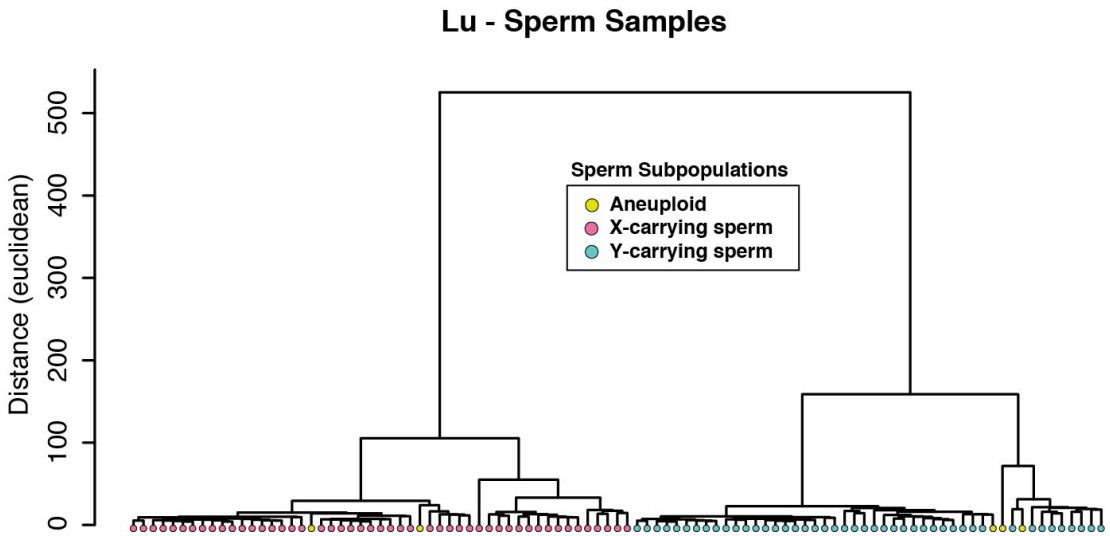
Supplementary Figure 6 | A phylogenetic tree of the CTC single cell samples across all 7 patients at a bin size of 50kb using the copy number profiles of the individual cells. Only cells from patients 2 and 4 can be correctly clustered at this bin resolution. The cells of the remaining patients no cluster together.



Supplementary Figure 7 | Histograms of the normalized bin counts across **(a-g)** patients 1-7 respectively, **(h)** all CTC patients, **(i)** all T10 hypodiploid samples, **(j)** all T10 aneuploid samples, and **(k)** all T10 samples.



Supplementary Figure 8 | Histograms of the scaled bin counts across **(a-g)** patients 1-7 respectively, **(h)** all CTC patients, **(i)** all T10 hypodiploid samples, **(j)** all T10 aneuploid samples, and **(k)** all T10 samples.



Supplementary Figure 9 | Ginkgo clusters the Lu *et al.* sperm samples by x- and y-carrying sperm. The original study identified 5 aneuploid cells (shown in yellow) with copy-number aberrations. Ginkgo is able to identify the same variants in the 5 aneuploid cells.