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

Tyler Garvin[1,4], Robert Aboukhalil[1,4], Jude Kendall[1], Timour Baslan[1–3], Gurinder S Atwal[1], James Hicks[1], Michael Wigler[1] & Michael C Schatz[1]

We present Ginkgo (http://qb.cshl.edu/ginkgo), a user-friendly, open-source web platform for the analysis of single-cell copy-number variations (CNVs). Ginkgo automatically constructs copy-number profiles of cells from mapped reads and constructs phylogenetic trees of related cells. We validated Ginkgo by reproducing the results of five major studies. After comparing three commonly used single-cell amplification techniques, we concluded that degenerate oligonucleotide-primed PCR is the most consistent for CNV analysis.

Single-cell sequencing[1] has become an important tool for probing cancer[2], neurobiology[3], developmental biology[4–6] and other complex systems, enabling investigators to unravel genetic heterogeneity in samples and accomplish more complete phylogenetic reconstruction of subpopulations than is possible with bulk sequencing. Thousands of individual human cells have been profiled to map subclonal populations in cancerous tumors[7] and circulating tumor cells[8], to discover mosaic CNVs in neurons[3] and to identify recombination events in gametes[5,9]. One key application of single-cell sequencing is the identification of large-scale (>10 kb) CNVs[3,7,10]. In cancer, CNVs form a 'genetic fingerprint' from which one can infer the phylogenetic history of a tumor[11] and trace the progression of metastatic events[7]. Yet the analysis of single-cell sequence data is complex and demands tools that can make the approach more broadly accessible.

Although many computational tools exist for CNV analysis of bulk samples[12], currently there are no fully automated and open-source tools that address the unique challenges of single-cell sequencing data: (1) extremely low depth of sequencing coverage (<1×), which makes for noisy profiles and renders split-read, paired-end or single-nucleotide polymorphism density approaches ineffective; (2) whole-genome amplification (WGA) biases, such as the failure to amplify entire segments[13], which can markedly distort read counts; (3) inflated read counts ('bad bins') resulting from poorly assembled regions of the genome (for example, centromeres)[13]; (4) the need for new algorithms for calling copy numbers at single-cell, integer levels; and (5) the fact that current tools for exploring population structure are not built for single-cell data. In addition, several sources of cell-specific experimental errors, including GC content and other sequencing biases, need to be addressed.

Here we present Ginkgo, an open-source web platform for the automated and interactive analysis of single-cell CNVs (http://qb.cshl.edu/ginkgo and **Supplementary Software**). Ginkgo enables researchers to upload samples, select processing parameters and explore population-structure and cell-specific variants within a visual analytics framework in a web browser.

Ginkgo's user-friendly interface guides users through every aspect of the analysis, from inputting mapped reads through visualization and exploration of the single-cell copy-number profiles (**Fig. 1**). Briefly, mapped reads are binned by chromosome position, normalized for GC biases and other amplification artifacts and then segmented to identify chromosome regions with consistent copy-number states. Integer copy-number states are then assigned to each segment, which allows Ginkgo to calculate hierarchical tress and heat maps from the copy-number profiles of the collection of cells. This pipeline builds on our previous single-cell sequencing work[13] and contains several novel features: (1) an algorithm for determining absolute copy-number state from the segmented raw read depth, (2) a method for controlling quality issues in the reference assembly (Online Methods), (3) an option to integrate ploidy information from flow cytometry to more accurately call copy number and (4) a suite of interactive visual analytics tools that allows users to easily share results with collaborators and clinicians. Ginkgo provides functionality for five different species (human, chimp, mouse, rat and fly) and includes a wide array of tunable parameters for individual users' needs (Online Methods).

Once an analysis is complete, Ginkgo displays an overview of the data in a sortable data table, an interactive phylogenetic tree[14] of all cells used in the analysis and a set of heat maps detailing the CNVs that drove the clustering results. Clicking on a cell in the interactive phylogenetic tree or data table 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 a custom track of amplifications and deletions in the UCSC genome browser. Ginkgo also outputs several 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 for assessing coverage uniformity[15]. Subsets of interesting cells can also be selected by the user for direct comparison of copy-number profiles, Lorenz curves, GC bias and coverage dispersion.

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [2]Department of Molecular and Cellular Biology, Stony Brook University, Stony Brook, New York, USA. [3]Present address: Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to M.C.S. (mschatz@cshl.edu).

RECEIVED 11 NOVEMBER 2014; ACCEPTED 7 JULY 2015; PUBLISHED ONLINE 7 SEPTEMBER 2015; DOI:10.1038/NMETH.3578

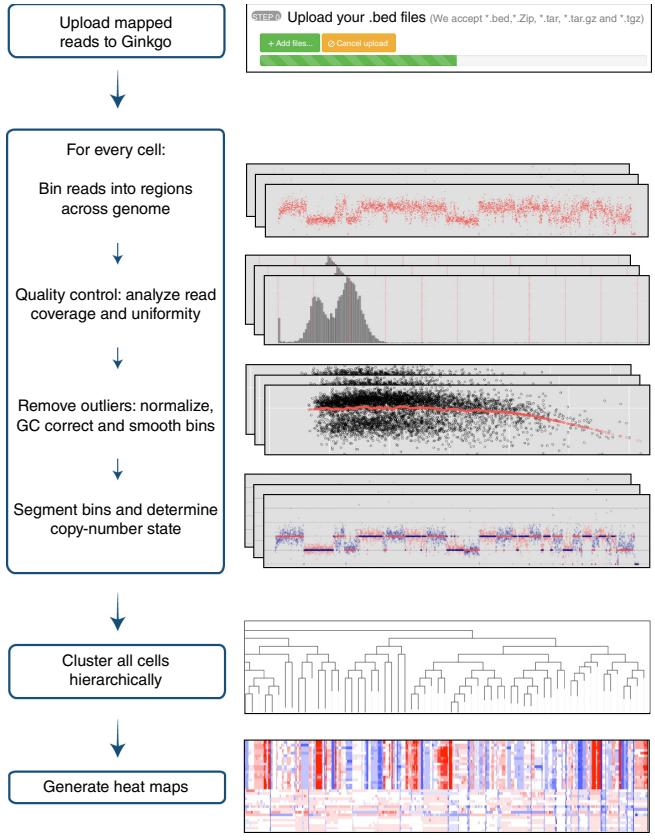

Figure 1 | Flowchart for performing single-cell copy-number analysis with Ginkgo. Usage and parameters are described in the Online Methods and on the Ginkgo website (http://qb.cshl.edu/ginkgo).

To validate Ginkgo, we set out to reproduce the findings of five single-cell studies that used either multiple annealing and looping-based amplification (MALBAC) or degenerate oligonucleotide–primed PCR amplification (DOP-PCR) (**Supplementary Note** and **Supplementary Table 1**). These studies addressed vastly different scientific questions, obtained data from a variety of tissue types and made use of different experimental and computational approaches at different institutions. Using Ginkgo, we replicated the vast majority of published CNVs for each cell in each of the data sets, with the exception of one cell from a study by Hou *et al.*[16]. We believe that this failed replication was due to mislabeling in the National Center for Biotechnology Information (NCBI) Sequence Read Archive. Moreover, Ginkgo was able to reproduce the distinct clonal subpopulations in the two data sets from Navin *et al.*[7] (**Supplementary Fig. 1**) and the patient clustering results from Ni *et al.*[8] (**Supplementary Fig. 2**) that were generated from called CNVs. Using simulated copy-number profiles, we confirmed that Ginkgo reliably identified copy-number changes (98.8% accuracy, 98.7% true positive rate and 1.2% false positive rate) and perfectly reproduced population structure through clustering of the individual samples (Online Methods and **Supplementary Table 2**).

Although Ginkgo corrects for many of the biases present in single-cell data, higher-quality data inevitably lead to higher-quality results. We set out to compare the biases and differences in coverage uniformity among the three most widely published WGA techniques—multiple-displacement amplification (MDA), MALBAC and DOP-PCR—using three distinct data sets with each method.

Raw sequencing reads downloaded from NCBI were mapped to the human genome and downsampled to match the sample with the lowest coverage. Aligned reads were then binned into variable-length intervals across the genome that averaged 500 kb in length but contained the same number of uniquely mappable positions (Online Methods). We use these binned read counts

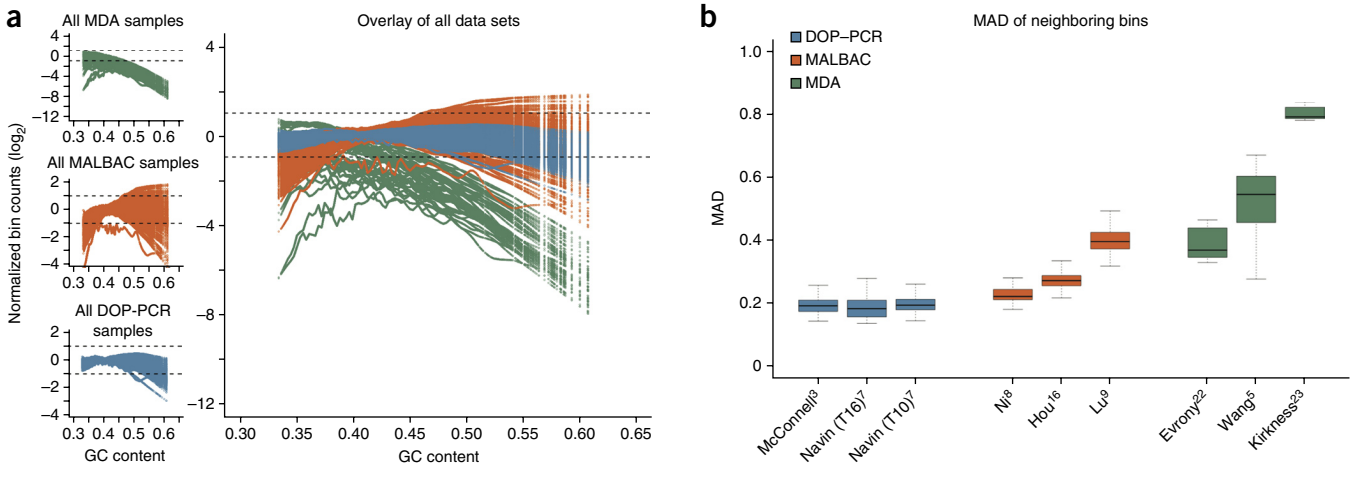

Figure 2 | **Assessment of data quality for different single-cell whole genome amplification methods using Ginkgo.** (**a**) LOWESS fit of GC content with respect to log-normalized bin counts for all samples in each of the nine data sets analyzed: three for MDA (top left, green), three for MALBAC (center left, orange) and three for DOP-PCR (bottom left, blue). Each colored line in a plot corresponds to the LOWESS fit of a single sample. The upper and lower dashed lines in each plot mark twofold increased and decreased values with respect to the average observed coverage. Note that the MDA plot has a different *y*-axis scale because of the large GC biases present in those data sets. (**b**) The MAD between neighboring bins. A single pairwise MAD value was generated for each sample in a given data set and is represented in the figure by a box and whisker. The bold line in the center of a box represents the mean, the box boundaries represent the quartiles and the whiskers represent the remaining data points. Names along the *x*-axis are the first authors of the referenced studies. T16 and T10 refer to types of breast cancer tumors as established by Navin *et al.*[7]. The high biases present in the MDA data sets made it difficult to compare DOP-PCR and MALBAC samples. **Supplementary Figure 3** shows this comparison more clearly.

to measure two key data-quality metrics: GC bias and coverage dispersion. Importantly, raw bin counts provide a view of data quality that is impartial to the different approaches to segmentation, copy-number calling and clustering.

GC content bias refers to preferential amplification of a given genomic region due to the local fraction of G and C nucleotides[17]. This bias introduces cell- and library-specific correlations between GC content and bin counts. In particular, when the GC content in a genomic region falls outside of a certain range (typically <0.4 or >0.6), read counts rapidly decrease (Online Methods). We found that the GC bias of MDA was very high compared with that of MALBAC or DOP-PCR (**Fig. 2a**). Only 45.9% of MDA bin counts fell within the expected coverage range, compared with 94.0% of MALBAC bin counts and 99.6% of DOP-PCR bin counts. It is important to note that regardless of the WGA approach used, each cell has unique GC biases that must be individually corrected.

As a further measure of data quality, we calculated the median absolute deviation (MAD) of all pairwise differences in read counts between neighboring bins for each sample, after normalizing the cells by dividing the count in each bin by the mean read count across bins. The MAD is resilient to outliers caused by copy-number breakpoints, as transitions from one copy-number state to another are relatively infrequent. Instead, pairwise MAD reflects the bin count dispersion due to technical noise. As expected on the basis of previous comparisons[15,18], MDA data displayed high levels of coverage dispersion, with a mean MAD two to four times that of the DOP-PCR data sets (**Fig. 2b**). In addition, the MALBAC and MDA data sets showed large differences in data quality between studies, whereas the DOP-PCR data sets showed consistently flat MAD across all three studies (**Supplementary Fig. 3**).

We found that DOP-PCR outperformed both MALBAC and MDA in terms of data quality. As previously reported[15,18–21], MDA displayed poor coverage uniformity and low signal-to-noise ratios. These characteristics, coupled with overwhelming GC biases, make MDA unreliable for accurate determination of CNVs compared with the other two techniques examined. Furthermore, although both DOP-PCR data and MALBAC data can be used to generate CNV profiles and identify large variants, DOP-PCR data have a substantially lower coverage dispersion and smaller GC biases than MALBAC data. Our results indicate that given the same level of coverage, data prepared using DOP-PCR can reliably call CNVs at higher resolution with better signal-to-noise ratios and are more reliable for accurate copy-number calls than are data obtained with MDA or MALBAC.

METHODS

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

Accession codes. The following accessions were referenced in this study: SRP017516, SRA053375, SRA056303, SRA060945, SRP029757, SRA091188, SRX021401, SRX037035, SRX037132 and SRP030642.

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

ACKNOWLEDGMENTS

We thank N. Navin and P. Andrews for their helpful discussions and for assisting with data access. The project was supported in part by the US National Institutes of Health (award R01-HG006677 to M.C.S.), the US National Science Foundation (DBI-1350041 to M.C.S.), the Starr Cancer Consortium (I7-A723 to G.S.A.), the Breast Cancer Research Foundation (BCRF) (to M.W. and J.H.), the Simons Foundation, Simons Center for Quantitative Biology (SFARI award number 235988 to M.W.), the Susan G. Komen Foundation (llR13265578 to J.H.), the Prostate Cancer Foundation (Challenge Award to J.H.), the Cold Spring Harbor Laboratory (CSHL) Cancer Center (Support Grant 5P30CA045508) and the Watson School of Biological Sciences at CSHL through a training grant (5T32GM065094) from the US 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 read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
2. Wigler, M. *Genome Med* **4**, 79 (2012).
3. McConnell, M.J. *et al. Science* **342**, 632–637 (2013).
4. Blainey, P.C. *FEMS Microbiol. Rev.* **37**, 407–427 (2013).
5. Wang, J., Fan, H.C., Behr, B. & Quake, S.R. *Cell* **150**, 402–412 (2012).
6. Gundry, M., Li, W., Maqbool, S.B. & Vijg, J. *Nucleic Acids Res.* **40**, 2032–2040 (2012).
7. Navin, N. *et al. Nature* **472**, 90–94 (2011).
8. Ni, X. *et al. Proc. Natl. Acad. Sci. USA* **110**, 21083–21088 (2013).
9. Lu, S. *et al. Science* **338**, 1627–1630 (2012).
10. Navin, N. *et al. Genome Res.* **20**, 68–80 (2010).
11. Henrichsen, C.N., Chaignat, E. & Reymond, A. *Hum. Mol. Genet.* **18**, R1–R8 (2009).
12. Alkan, C., Coe, B.P. & Eichler, E.E. *Nat. Rev. Genet.* **12**, 363–376 (2011).
13. Baslan, T. *et al. Nat. Protoc.* **7**, 1024–1041 (2012).
14. Smits, S.A. & Ouverney, C.C. *PLoS One* **5**, e12267 (2010).
15. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. *Science* **338**, 1622–1626 (2012).
16. Hou, Y. *et al. Cell* **155**, 1492–1506 (2013).
17. Ross, M.G. *et al. Genome Biol.* **14**, R51 (2013).
18. Navin, N.E. *Genome Biol.* **15**, 452 (2014).
19. Cai, X. *et al. Cell Rep.* **8**, 1280–1289 (2014).
20. Chen, M. *et al. PLoS One* **9**, e114520 (2014).
21. de Bourcy, C.F. *et al. PLoS One* **9**, e105585 (2014).
22. Evrony, G.D. *et al. Cell* **151**, 483–496 (2012).
23. Kirkness, E.F. *et al. Genome Res.* **23**, 826–832 (2013).

ONLINE METHODS

Code availability. The source code for Ginkgo is available open source as **Supplementary Software** and is maintained at https://github.com/robertaboukhalil/ginkgo. It is also preinstalled at http://qb.cshl.edu/ginkgo. It provides a large number of user-specified parameters to control analysis and interpretation (**Supplementary Table 1**). Several parameters must be set according to the experimental design (genome, bin size, sex chromosome masking and flow cytometry copy-number estimation), but others allow the researcher to explore the analysis using different metrics depending on the goals of the study. A more complete description is provided below.

Binning method. Copy-number analysis begins with binning of uniquely mapping reads into fixed- 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. The use of fixed-length bins is generally discouraged because they lead to read dropout 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 by Navin *et al.*[7] to sample 101-bp stretches of the reference assembly at every position along the genome. These simulated reads are mapped back to the genome using Bowtie[24], and only uniquely mapping reads are analyzed. For a given bin size, we assign reads to bins such that each bin has the same number of uniquely mappable reads. Consequently, intervals with higher repeat contents and low mappability will be larger than intervals with highly mappable sequences, although they will both have the same number of uniquely mappable positions.

When variable-length bins are used 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, users are encouraged to use larger bins.

Masking bad bins. A number of regions, specifically around the centromeres of certain chromosomes, attain very high read depths compared with the expected depth in both bulk and single-cell sequencing data. Using data from 54 normal individual diploid cells from the breast tissue of multiple individuals, we determined these 'bad bins' in the human reference genome (hg19) as follows[25]. We divided the bin counts by the mean bin counts for each cell to normalize for differences between cells in the total read count. For each chromosome, we subtracted the mean of the bins over all cells from each individual cell's normalized bin count to normalize for differences between chromosomes. We then used the mean and s.d. of the autosomes to compute an outlier threshold corresponding to a P value of $1/N$, where N is the number of bins used. Fewer than 1% of bins were identified as extreme outliers and masked for further processing.

GC bias correction. Once reads are placed into bins, Ginkgo normalizes each sample and corrects for GC biases before segmentation. The normalization process begins with division of 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[26] using the R function LOWESS (smoother span, 0.5; three iterations; 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, $\log_2$.

Segmentation. After GC bias correction, bin counts are segmented to reduce fluctuations in noise across chromosomes and identify longer regions of equal copy number. Ginkgo makes use of circular binary segmentation (CBS), which segments the genome by recursively splitting the chromosomes into segments on the basis of a maximum t-statistic until a reference distribution estimated by permutation is reached[27]. Once the CBS segmentation is complete, the breakpoints (segment boundaries) across all bins are determined, and the counts for all bins in each segment are reset as the median bin count value in 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 data uploads from such a cell, this is not always possible, 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 IOD (index of dispersion, or 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 IOD will likely be among the submitted cells with the most evenly balanced ploidy and highest quality.

Determining copy-number state. The quantized nature of single-cell data means that every genomic locus should have an integer copy-number value, and the same number of reads per bin should separate every sequential copy-number state—for example, ~50 reads for copy number 1, ~100 reads for copy number 2, ~150 reads for copy number 3, etc. Although biological and technical noise prevent read counts from segregating perfectly into distinct copy-number states, read counts should still be centered on integer copy-number states.

The most direct approach for determining the copy-number state of each cell is available for users who have *a priori* knowledge of the ploidy of each sample. For example, cells that are DAPI-stained before cell sorting can be gated on the basis of their fluorescence activity, and ploidy can be determined through comparison of the cell's fluorescence activity with that of a reference cell with a known copy-number 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 flow cytometry 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.

Because flow cytometry data are not always available for analysis and have potential for error, Ginkgo provides an alternative to determine the copy number of each sample. Because data for each cell are binned, normalized and segmented, this copy-number profile has a mean of 1 and 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 the RCNP and *X*, and the final integer copy-number profile (FCNP) would be the rounded value of the SCNP so that all segments contained 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-squares (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 has been identified and applied, the scaled bins are rounded to generate the FCNP 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. The final step before visualization 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 distance metrics: Euclidean, maximum, Manhattan, Canberra, binary and Minkowski. After computing the dissimilarity matrix, Ginkgo computes a dendrogram through neighbor joining or 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 sex chromosomes. Careful consideration of gender must be given when analyzing patients from mixed populations, as the combined set of the X and Y chromosomes makes up a large fraction of the human genome that can distort the clustering results. Indeed, when we used Ginkgo to examine the data set from Ni *et al.*[8] with sex chromosomes masked, we could still discriminate between individual patients' tumors, but we could no longer discriminate between lung adenocarcinoma (ADC) and small-cell lung cancer (SCLC) cells (**Supplementary Fig. 2b**); the SCLC patients were exclusively female and, with one exception, the ADC patients were all male. Ginkgo comes prepackaged with the ability to selectively mask sex chromosomes to prevent gender biases from dominating the clustering.

Single-cell data sets analyzed. We validated Ginkgo by reproducing major findings of several single-cell sequencing studies that used three different WGA techniques: MALBAC, DOP-PCR/WGA4 and MDA. We analyzed the data characteristics of nine data sets across five tissue types (**Supplementary Table 2**). The Ginkgo parameters for these data sets are described in the main text, and additional parameters are noted below.

We mapped reads to hg19 using Bowtie and kept only uniquely mapped reads (mapping quality score ≥ 25). Mapped read counts ranged from 1,538,234 (Ni *et al.*[8]) to 30,638,853 (Lu *et al.*[9]) with a mean of 15,827,886. To perform an unbiased comparison, we randomly downsampled all samples to 1,538,234 reads to match the lowest available coverage.

To compute the GC biases across all nine data sets, we calculated the LOWESS fit of the $\log_2$-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 monitored the bias of each cell by calculating the proportion of bins that showed a twofold change from the expected coverage in either direction (by ± 1, $\log_2$).

Detailed comparison of MALBAC and DOP-PCR protocols. WGA using MDA introduces a large degree of bias compared with MALBAC or DOP-PCR, limiting its applicability to CNV analysis. Therefore, in the remaining comparisons we focused on MALBAC and DOP-PCR. For a fine-grained comparison of the two techniques, we compared the T10 data set from Navin *et al.*[7] and the circulating tumor cell (CTC) data set from Ni *et al.*[8], because of the similar biological and technical conditions used in the studies and similar published analyses. Both data sets contain information from aneuploid cancer cells, were sequenced to similar depths (CTC mean read count of 4,133,466; T10 mean read count of 6,706,119) and were used to generate phylogenetic clusters of samples on the basis of CNVs. We began by comparing the coverage dispersion and investigated the minimum coverage and bin size needed to reproduce the published results.

Coverage dispersion. When we used the MAD criteria described above, the DOP-PCR–based T10 data set showed markedly better bin-to-bin correlation than the MALBAC-based CTC data set as judged by a lower MAD of adjacent and offset bin counts (**Supplementary Fig. 4**). For adjacent bins, the first quartile of the CTC MAD comparison was higher than the third quartile of the T10 MAD comparison. As we increased the bin offset, we observed greater variation in the CTC data, as shown by the separation of the mean MAD between the T10 and CTC data sets. We interpreted this to mean that there is more local trending in amplification efficiency in MALBAC than in DOP-PCR data.

Minimum coverage requirement. We next explored whether WGA protocols differ with respect to the minimum coverage required to observe the same population or clonal substructure identified at full coverage. To this end, we downsampled all data sets and analyzed each in Ginkgo to determine (1) how well segment breakpoints were conserved and (2) how well the phylogenetic relationships were maintained. With all degrees of downsampling (from 25% to 99%), the T10 data showed better breakpoint conservation than the CTC data, but as expected, increased degrees of downsampling led to substantial erosion of breakpoint boundaries in both data sets (**Supplementary Fig. 5**).

Nevertheless, these downsampling experiments showed that MALBAC and DOP-PCR are remarkably robust with respect to preserving the overall clonal or population structure, even at extremely low coverage, although additional, smaller CNVs can be discovered with deeper coverage[28]. The clonal structure of the T10 data set remained fully intact across all downsampling experiments even as the mapped reads were downsampled by

doi:10.1038/nmeth.3578

99% (from ~608 reads per bin to ~6 reads per bin). The population structure of the CTC data set was preserved when downsampled by 95% (from ~597 reads per bin to ~30 reads per bin); after downsampling to 99%, one cell from one patient was incorrectly clustered.

Although the depth of coverage in both studies was originally very low (<0.15×), our downsampling results indicate that Ginkgo can correctly determine the phylogenetic relationship between samples even when sequenced to a depth of coverage of only 0.01×. If generally applicable, which we have not proven here, this approach will allow for 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 affects the resolution at which CNVs can be called. Up to this point in the study we had used 500-kb bins to reproduce the results of Navin *et al.*[7] and Ni *et al.*[8] following the procedure by Ni *et al.*[8]. 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 the bin size from 500 kb to 10 kb (**Supplementary Table 3**) for both data sets until the hierarchical clustering of the copy-number profiles produced different results.

The T10 data set retained its hierarchical structure until bin sizes dropped below 25 kb (**Supplementary Fig. 6**), whereas the CTC data set lost its original hierarchical structure at a bin size of 100 kb. In the T10 data set, when bin sizes dropped to 10 kb, a few hypodiploid cells clustered incorrectly. In the CTC data set, as bin sizes approached 100 kb, cells from two patients (4 and 7) began to overlap. With 50-kb bins, there was widespread overlap between nearly all patients' cells, and only the cells from two patients clustered correctly (**Supplementary Fig. 7**). This indicates that at the same level of coverage, DOP-PCR can resolve smaller CNVs than MALBAC can, but more comparably structured studies are needed.

Detecting integer copy-number states. Preliminary analysis of bin counts indicated 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 supported this claim, as the ability to properly discriminate between CTC patients on the basis of copy-number state was lost at a bin resolution that was easily resolved with the T10 data set. To understand the effects of noise further, we evaluated each data set to discriminate distinct copy-number states.

Because the copy-number states of individual cells are integers, we expected the data to be centered at integer values. If the data are 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 data sets (**Supplementary Fig. 8**). We also mapped the distributions of bin counts for representative cells: excellent, typical and lower-quality cells, as well as the highest-quality population average (**Supplementary Fig. 9**). All T10 profiles had distinct peaks representative of integer copy-number values. Although a few cells in the CTC data set had distinct peaks, many of the CTC profiles had considerably worse resolution with substantial blurring between copy-number states. Furthermore, the scaled read-count distributions illustrated the substantial difference in signal-to-noise ratio between the T10 and CTC data sets (**Supplementary Fig. 10**).

Analysis of copy-number accuracy. To test the accuracy of the copy-number and clustering analysis by Ginkgo, we simulated single-cell sequencing of 90 cells with 100 total copy-number events per cell. We modeled the cells after a population comprising nine distinct clonal populations, with ten cells per population (**Supplementary Fig. 11a**). We began by generating three primary clonal populations by introducing 80 copy-number events compared to the parent diploid cell. Next, for each of the three primary clones, we generated three subclonal populations by introducing an additional 20 nonoverlapping copy-number events to the original clones. Overall, this resulted in nine distinct subclones belonging to three larger clonal populations with a total of 100 CNVs with respect to the human reference genome (hg19).

The genome positions of CNVs were nonoverlapping and generated from a uniform random distribution across the genome. The lengths of CNVs were generated from an exponential distribution with a mean of 5 Mb and ranged between 200 kb and 20 Mb to approximate the CNVs observed in the genuine data. The copy-number states of the CNVs were generated from a Poisson distribution with a mean of 2.5, excluding the value 2.

We generated ten cells from each of the nine subclones (90 cells in total) by simulating reads from the subclone reference sequences. For each cell, we simulated 200,000 101-bp, single-end reads from the subclone reference sequence using dwgsim (https://github.com/nh13/DWGSIM) (dwgsim -n 101 -z -1 -e 0.01 -d 1 -r 0 -1 101 -2 0). For each cell, the simulated reads were then mapped to the hg19 human reference genome using the command "bowtie hg19.fa -S -t -m--best -strata" and filtered for only uniquely mappable high scoring reads (quality > 25). The SAM output was then converted to BED format, and all 90 cells were uploaded and analyzed directly in Ginkgo with variable-length 50-kb bins.

Ginkgo is able to accurately reproduce the population structure through hierarchical clustering (**Supplementary Fig. 11b**). In addition, we examined Ginkgo's ability to call CNVs by examining the false negative and false positive rates for all 90 cells at three different read counts (2 million, 1.5 million and 1 million) across three different bin sizes (100 kb, 50 kb and 25 kb) (**Supplementary Table 2**). We measured a 0.15% false negative rate and a 0.08% false positive rate, excluding those bins that were partially spanned by a copy-number alteration. When the entire genome was considered, including partially spanned bins, Ginkgo still had only an ~2% false negative rate and ~1.2% false positive rate. Hence, as expected, errors were almost exclusively concentrated at the boundaries of CNVs where the precise end of the event could not be determined because of the extremely low coverage available or partial spanning of a bin.

We compared these results to the widely used CNVnator[29] algorithm (http://sv.gersteinlab.org/cnvnator) for bulk sequencing CNV analysis and found that Ginkgo performed CNV calls

with higher accuracy (**Supplementary Table 2**). Furthermore, CNVnator and other bulk sample analysis programs do not attempt to assign integer copy-number states, but in this analysis we measured Ginkgo's accuracy with this stricter requirement, whereas for CNVnator we could evaluate only whether an amplification or deletion had been identified. Ginkgo also has numerous features for evaluating population-wide CNV relationships (heat maps and hierarchical clusters, multisample GC and Lorenz plots, etc.) that are also not present in CNVnator or other bulk sample programs that we could not evaluate. Finally, in a practical sense, we found Ginkgo to be substantially faster than CNVnator, requiring a few hours via a simple web interface, rather than many days in a very difficult-to-install console program for the 90-cell evaluation.

We further evaluated Ginkgo's accuracy by means of a detailed comparison to the results presented by McConnell *et al.*[3] (**Supplementary Note 1**). That study profiled CNV events in human induced pluripotent stem cell–derived fibroblasts and 110 frontal cortex neurons and found a wide degree of mosaic copy-number variation. They reported a total of 148 CNVs across 45 of the 110 sequenced cortical neurons using DOP-PCR–based single-cell sequencing and their own analysis pipeline. When we processed the same data with Ginkgo, we found 99.7% bin-level concordance between the two analysis pipelines, including a very high correlation (R^2 = 0.996) between the copy-number assignments of the predicted CNVs (**Supplementary Fig. 12a**). We investigated the disagreement between the pipelines and found that it was primarily due to differences in analyzing repetitive sequences (**Supplementary Fig. 12b**) or differences in internal thresholds (**Supplementary Fig. 12c**). In a final assessment, we found that Ginkgo was able to correctly identify the major populations in the study by Lu *et al.*[9] and separated X chromosome–carrying sperm, Y chromosome–carrying sperm and aneuploid cells (**Supplementary Fig. 13**).

24. Langmead, B. *et al. Genome Biol.* **10**, R25 (2009).
25. Baslan, T. *et al. Genome Res.* **25**, 714–724 (2015).
26. Cleveland, W.S. *Am. Stat.* **35**, 54 (1981).
27. Olshen, A.B. *et al. Biostatistics* **5**, 557–572 (2004).
28. Daley, T. & Smith, A.D. *Bioinformatics* **30**, 3159–3165 (2014).
29. Abyzov, A. *et al. Genome Res.* **21**, 974–984 (2011).

doi:10.1038/nmeth.3578