Genome-wide copy number analysis of single cells

Timour Baslan[1,2], Jude Kendall[1], Linda Rodgers[1], Hilary Cox[1], Mike Riggs[1], Asya Stepansky[1], Jennifer Troge[1], Kandasamy Ravi[1], Diane Esposito[1], B Lakshmi[3], Michael Wigler[1], Nicholas Navin[4,5] & James Hicks[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [2]Molecular and Cellular Biology Program, Stony Brook University, Stony Brook, New York, USA. [3]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [4]Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [5]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. Correspondence should be addressed to J.H. (hicks@cshl.edu).

Published online 3 May 2012; doi:10.1038/nprot.2012.039

Copy number variation (CNV) is increasingly recognized as an important contributor to phenotypic variation in health and disease. Most methods for determining CNV rely on admixtures of cells in which information regarding genetic heterogeneity is lost. Here we present a protocol that allows for the genome-wide copy number analysis of single nuclei isolated from mixed populations of cells. Single-nucleus sequencing (SNS), combines flow sorting of single nuclei on the basis of DNA content and whole-genome amplification (WGA); this is followed by next-generation sequencing to quantize genomic intervals in a genome-wide manner. Multiplexing of single cells is discussed. In addition, we outline informatic approaches that correct for biases inherent in the WGA procedure and allow for accurate determination of copy number profiles. All together, the protocol takes ~3 d from flow cytometry to sequence-ready DNA libraries.

INTRODUCTION

Copy number variation is an important source of genetic variation in humans and other organisms, and it is known to influence phenotypic traits[1]. Many studies have associated copy number variants with normal phenotypes, such as in human olfactory receptors and smell[2] and in the salivary amylase-encoding gene and diet[3]. Importantly, CNV has been linked with a wide range of deleterious phenotypes and disorders such as obesity[4], psychiatric disorders[5] and cancer[6]. In cancer, the commonality of copy number alterations has led to intense investigations of the copy number landscapes of tumors[7,8]. Thousands of tumors, across many cancer types, have been profiled using a variety of copy number detection techniques. These investigations have allowed the identification of disease-associated alterations that have subsequently been used to guide therapeutic decisions, for example, using amplification of the *ERBB2* locus to qualify patients for Herceptin[9].

The most commonly used method in interrogating the copy number landscape of genomes has been array comparative genomic hybridization (aCGH)[10]. aCGH technology, based on differential labeling of sample and reference (e.g., tumor and matched normal) DNA with fluorophores, hybridization to arrays containing oligonucleotide probes and subsequent analysis of fluorometric signal ratios, allows the calling of the copy number profile of discrete genomic intervals. The study of copy number alterations using aCGH has led to the identification of recurrent amplifications and deletions across many human cancers[11,12]. However, aCGH is not without limitations[13]. An important drawback of this technique, specifically in the study of cancer genomes, stems from the use of whole genomic DNA purified from tumor tissue in which genomically normal cells are almost always present. The presence of such 'nontumor' components dilutes CGH signals and can result in inaccurate copy number calling of certain genomic segments (for example, single copy deletions or duplications in polyploid tumors). Furthermore, even if tumor cells are enriched by means such as laser-capture microdissection, aCGH does not allow for the characterization of tumor heterogeneity in which multiple clones with distinct genomic profiles might be present in the tissue samples. Thus, methods that can obviate such shortcomings are of pivotal importance.

Major advances in genomic research have paralleled the emergence of next-generation sequencing technology[14–16]. The highly quantitative nature of the sequencing data and the ever-increasing output of next-generation sequencing machines have led to the adoption of sequencing technologies in all facets of genomic research. In the area of copy number analysis, for example, many laboratories have successfully leveraged the power of high-throughput sequencing in profiling genome copy number landscapes with notable advantages over aCGH[17,18]. The depth of data generated by cancer sequencing projects has allowed investigators to discern the phylogeny of tumorgenesis, and it has rekindled the cancer community's interest in a long known facet of tumor biology: intratumor heterogeneity. Somatic mutations identified in whole-genome sequencing efforts were found not to be present in all of the cells constituting the tumor mass; rather, they were present at varying percentages in the tumor cell populations[19,20]. With the notion that the study of tumor heterogeneity is necessary to understand tumor biology, numerous reports started to emerge that described the heterogeneous nature of cancer in greater detail[21–23].

In order to better understand and characterize tumor heterogeneity, we developed an approach, SNS, that allows for the genome-wide characterization of a single-cell copy number profile[24]. SNS combines flow sorting of single nuclei, WGA and next-generation sequencing to characterize copy number alterations. To study the evolutionary dynamics and population structure of tumors, SNS was used to sequence 100 single cells of two breast tumors, one with a matching liver metastasis. The data allowed, for the very first time, a comprehensive view of the evolutionary processes occurring in tumor cells. One tumor was shown to contain three distinct tumor subpopulations that were likely to have originated from a common precursor and later diverged phylogenetically. In the second tumor (the one with the matching liver metastasis), the data indicated that the primary tumor mass was formed by a single clonal expansion of a highly aneupolid cell, which later migrated, seeded

Box 1 | Multiplexing of single-cell libraries

With the increase in throughput afforded by the HiSeq machine, multiplexing of single cells is warranted. Furthermore, using simulations to estimate the number of reads required to reproduce an accurate copy number profile, we have empirically determined that ~2 million uniquely mapped reads are sufficient to quantify the copy number profile using the varbin algorithm with 50 thousand bins (data not shown). To multiplex samples, we use a collection of bar codes designed by our laboratory of 7 nt in length (eight bar code sequences that we have tested and verified are provided as **Supplementary Data**). Bar code distributions are generally uniform, with bar code ratio values (expected/observed) consistently between 0.8 and 1.2. Alternatively, the TruSeq indexing system can be used as well. For more details regarding the TruSeq indexing system, please refer to the Illumina website.

the metastasis and underwent very limited further genomic evolution. Furthermore, in both tumors, a subpopulation was identified comprising abnormal cells lacking evidence of a clear common precursor. Although it is not described in detail here, it is clear that the SNS methodology is not limited to nuclei but can be applied to whole cells isolated by flow cytometry using fluorophore detection of surface markers and/or endogenously expressed fluorescence proteins. To that end, since our initial report, we have successfully applied single-cell analysis to human circulating cells sorted by EpCAM fluorescence and to mouse cells expressing various fluorescent proteins (K.R. and J.H., unpublished data).

SNS offers a unique approach to characterizing cellular heterogeneity on the basis of genome-wide CNV. However, given that SNS relies on the quantification of CNV using sparse sequencing data, heterogeneity that might arise as a result of other genomic aberrations such as single-nucleotide variants and short insertions and deletions (indels) will be missed. Such is the case with certain hematological cancers such as acute myeloid leukemia, in which a subset of tumors is characterized by a cytogenetically normal genome. In that case, alternative approaches, such as deep exome sequencing, could offer a view of the underlying heterogeneity. Nevertheless, given that the vast majority of epithelial tumors show markedly rearranged genomes, SNS offers a valuable tool for dissecting clonal populations in tumors. Here we present a detailed explanation of the working protocol of SNS.

Overview of the procedure: benchwork

The experimental protocol for SNS involves three discrete steps: flow sorting of single nuclei, WGA of the DNA and library construction for sequencing on the Illumina platform. In SNS, flow-sorted nuclei are deposited into wells in a 96-well plate format. WGA is performed using the Sigma-Aldrich GenomePlex WGA4 kit. Single-cell amplification is based on a proprietary amplification method that randomly fragments the genome and uses a unique combination of primer extension preamplification and degenerate oligonucleotide primers/adaptors to generate DNA fragments of 200–1,000 bp in length, which are distributed across the genome and flanked by a universal adaptor sequence. The resulting library is then amplified using universal oligonucleotide primers with defined cycling parameters. Although multiple WGA kits are currently available from different vendors, in our experience, the GenomePlex kit offers the most robust and consistent results.

After amplification, WGA-amplified DNA is processed for sequencing library preparation just like normal genomic DNA. The WGA protocol attaches unique 30-nt termini at the ends of the DNA molecules. As such, before library construction, DNA is sheared to allow the removal of the adaptor sequences by sonication. Sonicated DNA is then processed using a standard Illumina library preparation protocol with end repair, 3′ A-overhang addition and adaptor ligation. Adaptor-ligated libraries are purified using agarose gel electrophoresis, which is robust and generates high-quality libraries. Alternatively, when processing many libraries, it is more suitable to purify sequencing libraries using the AMPure beads purification system offered by Agencourt, which is amenable to scaling. After purification, sequencing libraries are enriched using PCR. Initially, we sequenced each single cell on a lane of Illumina GAIIx instrument. Since then, with the increasing capacity of the HiSeq platform and newly developed in-house informatic tools, we have adopted multiplexing using DNA bar codes to sequence many single cells on a single lane (**Box 1**). Pooling of individually bar-coded samples is accomplished during the last steps of library preparation, after amplification and quantification using the Agilent Bioanalyzer instrument. After quantification using the Bioanalyzer, each sample is diluted to a concentration of 10 nM and samples are pooled, thus yielding a final library at 10 nM concentration. **Figure 1** shows the schematic for the experimental workflow.

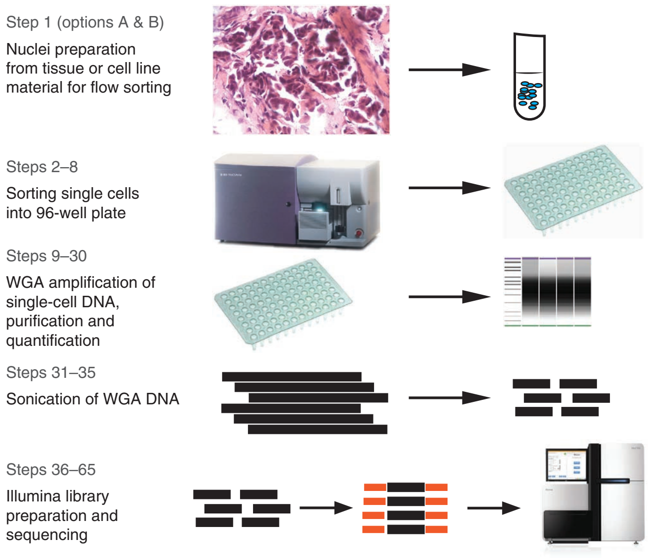
Step 1 (options A & B)
Nuclei preparation from tissue or cell line material for flow sorting
Steps 2–8
Sorting single cells into 96-well plate
Steps 9–30
WGA amplification of single-cell DNA, purification and quantification
Steps 31–35
Sonication of WGA DNA
Steps 36–65
Illumina library preparation and sequencing

Figure 1 | Schematic of the experimental workflow of SNS. Step numbering corresponds to the Steps of the PROCEDURE. The FACSAria image is courtesy of Becton, Dickinson and Company; reprinted with permission. HiSeq2000 image is courtesy of Illumina.

Overview of the procedure: informatic analysis

To obtain copy number profiles of single cells, sequencing data is processed using a variety of computational and algorithmic tools that include Bowtie, SAMtools, Python and the SPlus/R software package. Sequence data are first mapped using the Bowtie algorithm with defined parameters. Once data are mapped, sequencing reads are processed through a series of tools using the SAMtools package to remove PCR duplicates and arrange the sequencing reads in a proper format suitable for downstream analysis. Only uniquely mapped reads are used in determining copy number profiles.

In determining the copy number profiles, uniquely mapped reads are processed using an in-house developed Python algorithm (Varbin) that counts sequence read density in genomic intervals (bins). The Varbin algorithm, provided in the **Supplementary Methods**, differs from previous sequence-based copy number detection algorithms in that, in contrast to previous tools that divide the genome into fixed bins[17,18], Varbin divides the genome into bins of variable length adjusted such that the number of potential uniquely mapping reads in each bin is normalized across the genome. To determine the bin sizes we used in previously published work[24], we simulated 200 million sequences from the human genome (HG18/NCBI36), all of 48 nt in length, while introducing single-nucleotide errors at a frequency similar to that encountered during Illumina sequencing. These simulated sequences were mapped back to the human genome (HG18) with defined parameters. Chromosomal bins were assigned on the basis of the proportion of mapped simulated sequence reads, with each bin containing an equal number. This resulted in ~50,000 distinct, nonoverlapping bins. We have since revised the method using the hg19 reference genome and with a nonrandom algorithm to specify bin boundaries with concordant results. Bin boundaries determined from hg19 simulations are provided here. Furthermore, owing to inherent mapping errors, some bins accumulate high read counts and appear as high focal amplifications. These bins are discussed in more detail in the PROCEDURE.

Crucially, because we simulated single-end reads from hg19 and mapped the sequencing reads using the Bowtie algorithm to define the variable bin boundaries, only Illumina single-end sequencing data that are mapped with Bowtie are useful for determining the copy number profile with the boundaries that are supplemented in this protocol. If BWA is preferred to Bowtie, then the simulations will have to be repeated to define a new set of bin boundaries. The same applies to paired-end sequencing data or sequence data obtained using a different platform (for example, ABI SOLiD).

The output file of the Varbin algorithm contains sequence counts in the assigned genomic bins. These data are processed to yield integer copy number values via a variety of algorithmic tools such as the Kolmogorov-Smirnov or circular binary segmentation (CBS), and Gaussian kernel smoothed density plots, which are usually done with software package SPlus or its nonproprietary version, R.

Finally, on occasion, we observe single-cell copy number profiles that contain unusually large homozygous chromosomal deletions or what appears to be 'shredding of chromosomes'. The nature of these profiles (i.e., whether they are biological or technical artifacts), is currently unknown and is under investigation. A discussion of these profiles, which we term 'Genome Sector Loss' is offered in the sections that follow.

The PROCEDURE takes the reader through an analysis example. The programs to use and example output files are provided in the **Supplementary Methods** and **Supplementary Data**, respectively. The **Supplementary Note** provides a brief overview, whereas **Supplementary Figure 1** provides a concise outline of all the steps with input-program-output labels.

Experimental design

Sample preparation and flow cytometry. In our initial report, we described the sequencing of single cells isolated from frozen tissue as well as single cells isolated from cell lines grown in tissue culture[24]. Although many techniques are available for the isolation of single cells (such as micromanipulation), we have empirically determined that flow cytometry offers a sensitive and reproducible approach. For sample preparation from frozen tissue, it is important to keep the tissue on dry ice in order to maintain the tissue's integrity for subsequent analysis. Generally, we remove a small sample of the tissue (1 mm × 1 mm) using no. 11 scalpels and transfer the piece to a Petri dish while maintaining the original tissue on dry ice. For first-time users, we generally recommend starting with cell line material to test the protocol. Before you set the gates for flow cytometry, we also recommend running a control sample (we use a diploid lymphoblastic cell line); at least 5,000 events of the examined sample (we usually collect 10,000) should be recorded in the DAPI channel to provide a clear picture of where the gates should be set (**Fig. 2**).

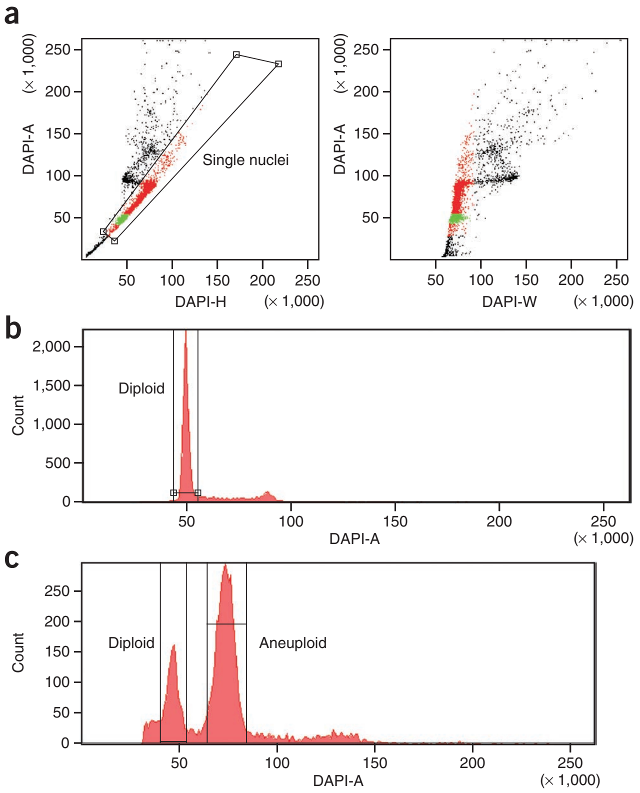

Figure 2 | Flow sorting of single nuclei on the basis of DNA content. (**a**) Dot plot view of DAPI-stained nuclei. Gate, drawn on the diagonal, excludes cellular debris and doublets and captures single nuclei. The black dots represent cellular debris and doublets, whereas the green and red dots represent diploid and nondiploid fractions from the single-nuclei gate, respectively. The dot plots are drawn with DAPI-H and DAPI-W to allow for enhanced precision in distinguishing subpopulations. (**b**,**c**) Examples of histograms drawn from single-nuclei gates illustrating a diploid profile (**b**) and an aneuploidy profile (**c**).

Box 2 | Crucial steps to take into consideration for FACS setup for sorting single cells

There are three crucial elements to FACS setup for single-cell sorting: the sample lines, flow cell and nozzle must be clean, droplet break-off must be stable, and the automatic cell deposition unit must be perfectly positioned. To achieve this, the following steps should be followed:

1. Perform 'flow cell clean' using FACSRinse and allow it soak for 10 min; then, perform 'flow cell clean' with $ddH_2$0 and soak for an additional 10 min. Run FACSClean through the sample line for 10 min at the highest flow rate setting and follow with $ddH_2$0 for another 10 min. Insert the nozzle, turn on the stream and allow it to stabilize for 30 min. Immediately before sorting, run $ddH_2$0 and record events for 5 min to verify that there are zero events.
2. Turn on Sort Test. Open the cover and sort block door. Visually inspect the side streams. The streams should be tight and steady. Because only the far left stream is used for sorting into plates, do not forget to check the left stream with the other turned off. Position the far left stream to the center opening of the splash shield. Close the sort block door and cover.
3. Determine the break-off point and drop delay using Accudrop. Drop formation and break-off must be stable. Sort precision should be set for 'single cell'; which selects high purity over yield.
4. Check the ACDU alignment. The ACDU is not designed to hold a PCR-size plate. A 96-well Falcon tissue culture plate is used as a holder for the PCR plate. Apply an adhesive plate seal to the surface of a 96-well PCR plate and smooth out the bubbles and wrinkles. Make sure the PCR plate is relatively flat and does not bow in the middle. Insert the PCR plate firmly into the Falcon plate and place it into the ACDU holder. Because the diameters of the wells of the PCR plate are much smaller than those of the tissue culture plate and the volume of lysis buffer is small, the sorted drops must be centered precisely in the middle of the wells. Set up the sort layout to sort 100 Accudrop beads per well onto the surface of the film. Deposit the beads, remove the plate from the ACDU and visually examine the position of the drops. Continue to adjust the ACDU until all wells are positioned correctly.

Flow cytometric determination of nuclear DNA content through DAPI staining also provides a means for identifying and isolating tumor subpopulations on the basis of ploidy. It is important to note that when handling different tumor samples, ploidy profiles (specifically aneuploid peaks) can differ depending on the tumor sample. As such, when sorting different samples, care must be taken in setting up appropriate gates. In addition, given that certain tumors have a high proliferative index, overlap between the G2/S-phase flow cytometry peak of diploid cells with that of the aneuploid peak might occur. Nonetheless, given the capacity of SNS to resolve genome profiles at the single-cell level, these cells (G2/S phase of diploids) are easily identified once the informatics analysis of sequencing data is performed. Given the high precision required for single-cell analysis, we also describe here important steps to consider when performing single-cell flow cytometry, such as droplet delay and break-off (**Box 2**).

WGA of single cells. As controls for the WGA of single cells, the flow cytometry settings can be adjusted as to leave certain wells empty (i.e., no deposition of a single cell). The products of these wells, when quantified for DNA, do not yield any measurable quantity of WGA DNA, and when they are run on an agarose gel or a Bioanalyzer, they do not show the WGA product smear indicative of a successful WGA reaction.

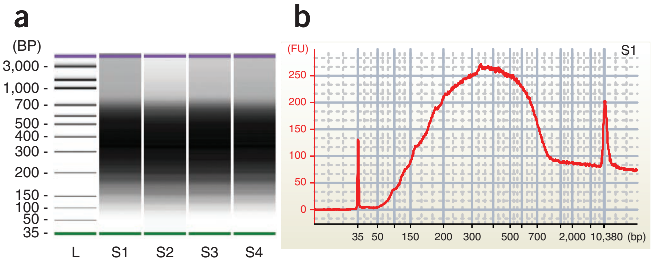

Figure 3 | WGA amplification profiles of single-cell DNA from four different single cells. (**a**) WGA DNA spreads (100–1,000 bp) of single-cell genomes as measured on the Bioanalyzer. S1-S2-S3-S4 refers to four different single cell–amplified products. (**b**) An example histogram of DNA spread from cell S1 as measured by the Bioanalyzer. FU, fluorescence units; L, ladder.

Sequence library construction. Samples selected for sequence library construction are analyzed by gel electrophoresis or by using the Bioanalyzer instrument to observe the WGA product spread between 100 and 1,000 bp (**Fig. 3**). In selecting samples for sequence library construction, we also take into account the DNA concentrations of the amplification products. Generally, diploid cells yield ~30 μl of material at ~200 ng μl^{-1} (ranging from 175 to 275 ng μl^{-1}). We avoid using WGA products of diploid-sorted cells that have concentrations above 300 ng μl^{-1} or below 175 ng μl^{-1}. Similarly, for aneuploid fractions, we observe DNA concentrations ranging from 250 to 400 ng μl^{-1}, and we proceed with only the samples in this concentration range. The reasoning behind the exclusion of such WGA products relates to concerns regarding nonuniform, incomplete amplification or an overamplification of single-cell genomes of the aforementioned products. We generally use 2 μg of WGA DNA to start with the library construction process; however, we have routinely been able to generate good libraries from as little as 0.5 μg of DNA using the method described in this protocol. This protocol and its associated timing information are intended for the construction of a single Illumina single-cell WGA library. Processing of multiple samples, for example for multiplexing purposes, is likely to increase the timing required.

Previously we reported sonication of WGA DNA using the Bioruptor ultrasonic disruptor[24]. However, we have switched to using the Covaris focus acoustics system, as it allows for higher throughput. Selection of the sonication programs depends on the desired insert length of the libraries. **Figure 4** illustrates the sonication profiles using multiple programs on the Covaris E210.

Informatic analysis. The central tenet of copy number analysis is based on the idea that sequenced molecules are a random sample of the genome, and that by computing local read density relative

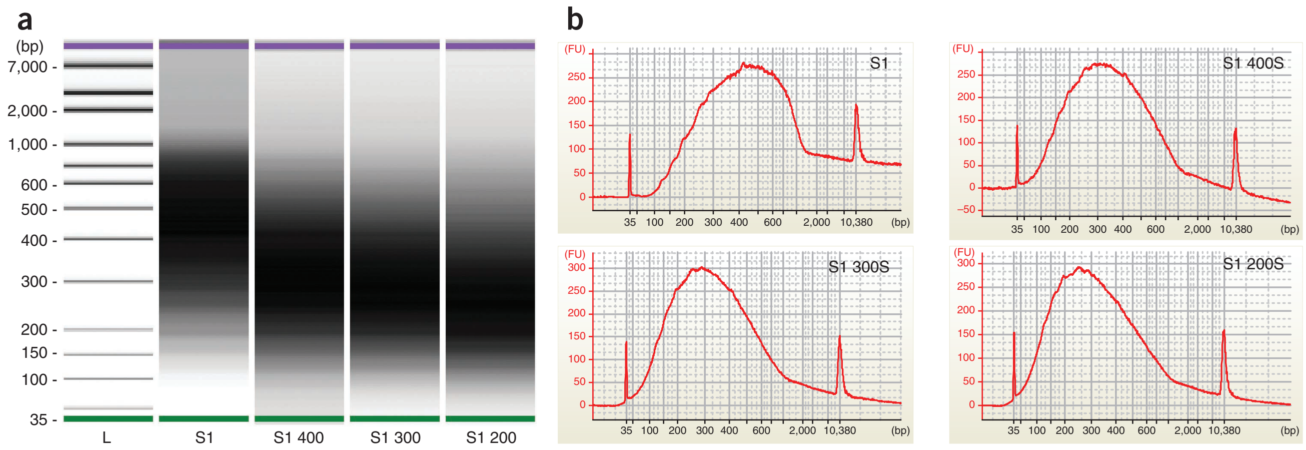

Figure 4 | Sonication profiles of WGA DNA using different sonication programs on the Covaris E210 instrument. (**a**) Sonication profiles as measured on the Bioanalyzer. S1: nonsonicated WGA DNA. S1 400 / S1 300 / S1 200: WGA DNA sonicated using 400±, 300± and 200± Covaris E210 programs, respectively. Profiles represent size distributions of DNA molecules (in bp) of the samples. The choice of which sonication program to use is dependent on the desired sequencing library length and the type of sequencing that will be implemented. We generally use the 300± program when sequencing 76-bp reads on the Illumina platform. (**b**) Histograms illustrating sonication profiles as measured by the Bioanalyzer. L, ladder.

to the average read density, it is possible to infer copy number. The method described here splits the genome into nonoverlapping regions (bins) that are expected to have the same average number of reads on the basis of a reference genome. This is accomplished by taking 50-bp sequences starting at each position in the reference genome, mapping them back to the reference, eliminating reads that map to multiple places in the genome (multimappers), and then setting bin boundaries such that each bin contains roughly the same number of uniquely mappable positions. The protocol does not give a uniform distribution. The main source of nonuniformity results from variation in GC content across the genome. To adjust for this, bin counts are normalized on the basis of GC content.

MATERIALS

REAGENTS

- Cells of interest: human tumor tissue, cell cultures grown in a cell culture dish of any kind, mouse tissue ▲ **CRITICAL** All experiments that use human tissue and animals should comply with institutional and governmental guidelines, and, where applicable, informed consent should be obtained from human subjects.
- Cell culture medium appropriate for cell type of interest
- Trypsin (Invitrogen, cat. no. 25200-056)
- DAPI (Invitrogen, cat. no. D1306)
- Whole genome amplification kit WGA4 (Kit includes proteinase K, library preparation enzyme, 10× single-cell lysis and fragmentation buffer, and so on. Sigma-Aldrich, cat. no. WGA4-50RXN) ▲ **CRITICAL** Sigma-Aldrich WGA4 kits yield relatively uniform distributions of amplification products across the genome to facilitate copy number analysis using the SNS method. It is imperative to use this kit to obtain reliable results.
- QIAquick 96-well purification kit (Qiagen, cat. no. 28181)
- Ethanol (100%, 200 proof; Ultra-Pure, cat. no. 200-CSPTP) **! CAUTION** Ethanol is flammable—keep it away from open flame.
- NP-40 (USB, cat. no. 19628) **! CAUTION** NP-40 contains materials that may cause respiratory tract, eye and skin irritation. It may be harmful if swallowed; handle it with appropriate care.
- $MgCl_2$ (VWR, cat. no. JT24440-1)
- NaCl (Fisher, cat. no. S271-10)
- Tris base (10 mM, pH 7.8; Fisher, cat. no. BP152-5)
- $CaCl_2$ (VWR, cat. no. JT1332-1)
- BSA (Sigma-Aldrich, cat. no. A7906-50G)
- QIAquick PCR purification kit (Qiagen, cat. no. 28106) **! CAUTION** Buffer PB contains irritant chaotropic salts. Take appropriate care when handling it.
- MiniElute PCR purification kit (Qiagen, cat. no. 28006) **! CAUTION** Buffer PB contains irritant chaotropic salts. Take appropriate care when handling it.
- QIAquick gel extraction kit (Qiagen, cat. no. 28704) **! CAUTION** Buffer QG contains irritant chaotropic salts. Take appropriate care when handling it.
- Agarose (Lonza, cat. no. 50004)
- GeneRuler 50-bp DNA ladder (Fermentas Life Sciences, cat. no. SM0373)
- FACSRinse (BD Biosciences, cat. no. 340346)
- FACSClean (BD Biosciences, cat. no. 340345)
- Accudrop beads (BD Biosciences, cat. no. 345249)
- Sucrose (USB, cat. no. 57-50-1; see REAGENT SETUP)
- Elution buffer (buffer EB; supplied with Qiagen PCR/gGel purification kits)
- T4 DNA polymerase (NEB, cat. no. M0203L)
- dNTPs, 10 mM each (supplied as 100 mM, see REAGENT SETUP; Roche, cat. no. 1 969 064)
- T4 DNA ligase buffer with 10 mM ATP (NEB, cat. no. B0202S)
- Klenow DNA polymerase (NEB, cat. no. M0210L; includes NEB buffer 2)
- T4 PNK (NEB, cat. no. M0201L)
- Agencourt AMPure (50 ml; Beckman Coulter, cat. no. A63880)
- dATP (1 mM; supplied as 100 mM, see REAGENT SETUP)
- Klenow fragment (3′–5′ exo –; NEB, cat. no. M0212L)
- Quick ligation kit (NEB, cat. no. M2200L)
- Sequencing oligo adaptors (IDT)
- Ethidium bromide (10 mg μl^{-1}; Sigma, cat. no. 057K8609) **! CAUTION** Ethidium bromide is a mutagen and potential carcinogen; handle it with care.
- TAE (50×; Invitrogen, cat. no. 24710)
- Phusion HF PCR master mix (NEB, cat. no. MO531L)
- Sequencing oligo primers (IDT)
- Agilent DNA high-sensitivity kit (Agilent, cat. no. 5067-4626)

EQUIPMENT

- Scalpels (no. 11 blade; VWR, cat. no. 89176-382)
- Tissue culture plates
- Polystyrene round-bottom tube (5 ml; Falcon, cat. no. 352058)

- Polystyrene round-bottom tube with cell-strainer cap (5 ml; Falcon, cat. no. 352235)
- Flow Cytometer FACSAriaII (BD Biosciences)
- 96-well PCR tubes (Thermo Scientific, cat. no. AB-0731)
- 8-well PCR trip tubes (0.2 ml; AB Applied Biosystems, cat. no. N8010580)
- Agarose gel electrophoresis unit (Thermo Scientific)
- Heating block (50 °C)
- Thermal cycler (MJ Research, cat. no. PTC-225)
- Vacuum manifold (Qiagen, cat. no. 19504)
- 96-well elution plates (Thermo Scientific, cat. no. AB-0796)
- Adhesive tape (Marsh Bioproducts, cat. no. AB-0626)
- Centrifuge tubes (1.5 ml; Eppendorf)
- Centrifuge (VWR, cat. no. 80076-424)
- Minicentrifuge (VWR, cat. no. 80094-172)
- UV transilluminator **! CAUTION** UV radiation is harmful to the unprotected eye and skin.
- Sonicator Covaris E210 (Covaris, cat. no. 500008)
- Sonication tubes (Covaris, cat. no. S20045)
- DynaMag-2 magnet (Invitrogen, cat. no. 123-21D)
- DynaMag-96 side (Invitrogen, cat. no. 123-31D)
- NanoDrop ND-1000 spectrophotometer (Thermo Scientific, cat. no. ND-1000)
- Agilent 2100 Bioanalyzer (Agilent Technologies, cat. no. G2938C)
- Illumina Genome Analyzer and associated equipment (Illumina)
- Bowtie software package (http://bowtie-bio.sourceforge.net/index.shtml; ref. 25)
- SAMtools software package (http://samtools.sourceforge.net/; ref. 26)
- Python software package (http://www.python.org/)
- R software package (http://www.r-project.org/)
- DNAcopy CBS segmentor (http://www.bioconductor.org/packages/2.8/bioc/html/DNAcopy.html). This is an R package used to segment the bin count data into nonoverlapping regions of differing copy number[27]
- Pipette tips
- Hemocytometer

REAGENT SETUP

NST buffer Mix the following components in ddH$_2$O for a final volume of 800 ml: 146 nM NaCl, 10 mM Tris base (pH 7.8), 1 mM CaCl$_2$, 21 mM MgCl$_2$, 0.05% (wt/vol) BSA and 0.2% (vol/vol) NP-40. NST buffer can be prepared and stored at 4 °C for up to 5 months.

NST-DAPI buffer To the 800 ml of NST buffer, add 200 ml of MgCl$_2$ at a concentration of 106 mM. Afterward, dissolve 10 mg of DAPI in the mixture and store it at 4 °C protected from light. The solution is stable for up to 5 months.

dATP, 1 mM Make 1 mM dilutions of the original 100 mM stock in buffer EB and store them at − 20 °C for up to 6 months.

dNTP, 10 mM Mix each dNTP, to a final concentration of 10 mM each, in buffer EB and store them at − 20 °C for up to 6 months.

Sucrose loading dye Prepare a 40% (wt/vol) sucrose solution by adding 40 g of sucrose to 100 ml of H$_2$O. The solution can be stored at room temperature (20–25 °C) for up to 3 months.

Single-cell lysis buffer Single-cell lysis buffer is prepared by mixing 800 µl of H$_2$O with 100 µl of mixture 1; mixture 1 is prepared by mixing 6 µl of proteinase K with 96 µl of 10× single-cell lysis and fragmentation buffer (both components of mixture 1 are in Sigma-Aldrich WGA kit).

EQUIPMENT SETUP

Covaris E210 sonicator Set the sonicator parameters as follows to obtain DNA distributions of ~300 bp (range of 200–400 bp): duty cycle − 10%, intensity − 4, cycles/burst − 200 and time 80 s. Make sure that the water bath temperature is at 4 °C. ▲ **CRITICAL** It is imperative that the Covaris water bath temperature is at 4 °C to ensure proper sonication and reproducible results.

FACSAriaII The FACSAriaII is configured with a high–powered, air-launched 350-nm UV laser and 450/50 band-pass filter. It is equipped with an ACDU (automated cell deposition unit). A 70-µm integrated nozzle is used and the fluidics pressure is set to 70 psi. For DNA content analysis, set parameters for DAPI area, width and height. Change the threshold to DAPI and set a value of 5,000. We adjust photomultiplier tube voltages using a normal diploid human lymphoblastoid cell line stained with DAPI as the control.

PROCEDURE

Sample preparation and flow cytometry ● TIMING 4 h

1| *DAPI staining of unfixed nuclei for FACS.* To perform DAPI staining of nuclei from tissue follow option A, and to perform DAPI staining of nuclei from cell cultures follow option B.

(A) DAPI staining of nuclei from tissue

(i) Place a piece of frozen tissue in a 60-mm tissue culture plate. Add 0.2–1.0 ml of NST-DAPI buffer, depending on the size of the tissue. For fine needle aspirates or core biopsies use 0.2–0.5 ml buffer. For larger pieces of tissue use 1–2 mm^3 of tissue in 1 ml of buffer.

(ii) Use two fine-point disposable scalpels to cut and tease apart the tissue in the buffer until the pieces are very fine. Gently mix with a 1-ml pipette tip.

(iii) Transfer the sample to a 5-ml Falcon round-bottom tube, leaving behind as much of the solids as possible. Hold the sample on wet ice and protect it from light for at least 10 min and no longer than 3 h.

(iv) Do not vortex the nuclei. Vortexing will result in substantial damage to the nuclei.

(v) Before running it on the flow cytometer, filter the sample through a 5-ml Falcon round-bottom tube with a cell-strainer cap.

(B) DAPI staining of nuclei from cultured cells

(i) Collect cells either by trypsinization of monolayer cultures (resuspend in complete medium) or collection of suspension cultures in medium.

(ii) Use a hemocytometer to count the number of cells.

(iii) Transfer 0.5 to 1.0 × 10^6 cells to a 15-ml conical centrifuge tube.

(iv) Gently centrifuge the tube at 105*g* for 4 min at room temperature.

(v) Aspirate the medium, being careful not to disturb the cell pellet.

(vi) With your index finger, flick the tube until the pellet seems to be dispersed and not solid.

(vii) Add 1 ml of NST-DAPI buffer per 0.5 to 1.0 × 10^6 cells.

(viii) Transfer the mixture to a 5-ml Falcon round-bottom tube (polystyrene); hold it on wet ice and protect it from light for at least 10 min and no longer than 3 h.
(ix) Do not vortex the nuclei. Vortexing will result in substantial damage to the nuclei.
(x) Before running it on the flow cytometer, filter the sample through a 5-ml Falcon round-bottom tube with a cell-strainer cap.

2| Run the sample on the FACSAria II cell sorter (or any comparable cell sorter). For assistance in running the FACSAria II cell sorter, refer to the Users' Guide provided by BD Biosciences (part no. 640760 Rev.A, usually supplied with the instrument).

3| Create a dot plot that plots DAPI area on the *y* axis and DAPI pulse height on the *x* axis (**Fig. 2a**). For assistance in setting dot plots and gates for DNA content analysis, refer to Wersto *et al.*[28].

4| Set a gate (no. 1) on a population of single nuclei (which appears on the diagonal) and exclude doublets or debris.

5| Create a histogram derived from gate no. 1 (single nuclei), which plots the count on the *y* axis and the DAPI area (DNA content) on the *x* axis on a linear scale.

6| Record data on 10,000 counts of single nuclei.

7| Set gates on populations of interest (**Fig. 2b**,**c**).

8| Sort fraction(s) into a 96-well PCR plate prepared with 9 µl of single-cell lysis buffer and kept on ice. Plates with lysis buffer are prepared by making aliquots of 9 µl into each well.

WGA ● TIMING 6 h

9| Use a thermal cycler to incubate the plates for 1 h at 50 °C, followed by 4 min at 99 °C.

10| Quickly spin and cool the plate on ice.
■ **PAUSE POINT** Samples can be spun down and kept at −20 °C until further processing; however, we generally carry the WGA reaction all the way to Step 17 before the 96-well plate purification.

11| Prepare mixture 2 (3 µl per sample) by mixing the following components:

Component	Volume per sample (µl)	Volume for a 96-well plate (100 samples; µl)
Single-cell library preparation buffer (1×)	2	200
Library stabilization solution	1	100
Total	3	300

12| Add 3 µl of mixture 2 to each sample, quickly spin, and incubate the samples in the thermal cycler at 95 °C for 2 min.

13| Quickly spin and replace the samples on ice.

14| Add 1 µl of library preparation enzyme (part of WGA kit), quick spin, and incubate in a thermal cycler as follows: 16 °C for 20 min; 4 °C for 20 min; 37 °C for 20 min; 75 °C for 5 min, and a 4 °C hold.

15| Quickly spin and incubate the samples on ice.
■ **PAUSE POINT** Samples can be spun down and kept at −20 °C until further processing; however, we generally carry the WGA reaction all the way to Step 17 before the 96-well plate purification.

16| Prepare mixture 3 by mixing the following components:

Component	Volume per sample (µl)	Volume for a 96-well plate (100 samples; µl)
Amplification master mix (10×)	7.5	750
H_2O	48.5	4,850
WGA DNA Pol	5	500
Total	61	6,100

17| Add 60 µl of mixture 3, mix well, quickly spin and incubate in a thermal cycler as follows:

Cycle	Denature	Anneal/extend	Hold
1	95 °C for 3 min	—	—
2–26	94 °C for 30 s	65 °C for 5 min	—
27			4 °C

■ **PAUSE POINT** Samples can be quickly spun down and kept at −20 °C until further processing.

18| Quickly spin and proceed to QIAquick 96-well plate purification.

QIAquick 96-well-plate PCR purification ● TIMING 1 h

19| Place a QIAquick 96-well plate into a vacuum manifold with the vacuum turned off.

20| Aliquot 300 µl of buffer PB (part of Qiagen kit) into all wells.

21| Transfer the PCR amplification mixture (from Step 18) into the wells and mix well by pipetting several times.

22| Turn on the vacuum and allow the PCR mixture to flow through until the membranes are dry.

23| Wash two times with 900 µl of buffer PE (part of Qiagen kit) per well.

24| Vacuum until dry.

25| Remove the QIAquick 96-well plate from the vacuum manifold; shake it to remove excess fluid and mount it on a waste collection plate.

26| Centrifuge at 1,470*g* for 5 min at room temperature.

27| Mount the QIAquick 96-well plate onto a fresh collection plate.

28| Add 50 µl of buffer EB (part of Qiagen kit) per well and incubate for 1 min.

29| Centrifuge at 1,470*g* for 5 min at room temperature to elute DNA.

30| After elution (elution generally yields ~30 µl of DNA), use a NanoDrop to determine WGA DNA concentrations. Samples can be run on agarose gel or Bioanalyzer to determine the amplification profile (for further details, refer to Agilent's Bioanalyzer manual).
▲ **CRITICAL STEP** Generally, we achieve a ~90% success rate in amplifying single-cell genomes from 96-well plates. The NanoDrop readings and Bioanalyzer profiles of successfully amplified single-cell DNA should have readings and profiles

similar to those mentioned in the Experimental design section. Only WGA DNA products with the aforementioned parameters are selected for library construction.
? TROUBLESHOOTING
■ **PAUSE POINT** Samples can be stored at –20 °C until further processing.

Sonication ● TIMING 30 min

31| Prepare 2 μg of WGA DNA in a total volume of 75 μl (bring up to volume with buffer EB).

32| Transfer the mixtures to Covaris microtubes.

33| Sonicate DNA as follows: duty cycle –10%, intensity –4, cycles/burst –200 and time 80 s.
▲ **CRITICAL STEP** Make sure that the water bath temperature is at 4 °C.

34| Transfer the Covaris Microtubes to tube holders.

35| Quickly spin to collect material and then transfer it to fresh PCR tubes to proceed with library preparation.
■ **PAUSE POINT** Samples can be briefly spun down and stored at –20 °C until further processing.

End repair of sonicated WGA DNA to generate blunt ends ● TIMING 45 min

36| Prepare the following master mix in a 1.5-ml centrifuge tube for each sample. Mix carefully by pipetting up and down:

Reagent	Volume (μl)
T4 DNA ligase buffer with 10 nM ATP	10
T4 DNA polymerase	5
T4 Polynucleotide kinase (PNK)	5
dNTP mix (10 mM each)	4
Klenow DNA polymerase	1

37| Transfer 25 μl of end repair mix from Step 36 to each sample from Step 35 and mix well by pipetting. Incubate in a thermal cycler for 30 min at 20 °C.

38| Purify each sample using the QIAquick PCR purification kit according to the manufacturer's protocol.

39| Elute each sample in 30 μl of buffer EB.
■ **PAUSE POINT** Samples can be briefly spun down and stored at –20 °C until further processing.

3′ A-overhang addition to blunted DNA ends ● TIMING 45 min

40| Prepare the following master mix in a 1.5-ml centrifuge tube. Mix well by pipetting:

Reagent	Volume (μl)
Klenow buffer (10×; NEB buffer 2)	5
ATP, 1 mM	10
Klenow fragment (3′–5′ exo-)	10

41| Transfer 25 μl of master mix from Step 40 to each sample from Step 39; mix well by pipetting and incubate at 37 °C for 30 min.

42| Purify each sample using the MiniElute PCR Purification kit according to the manufacturer's instructions.

43| Elute each sample in 17 µl of buffer EB.
■ **PAUSE POINT** Samples can be briefly spun down and kept at −20 °C until further processing.

Illumina adaptor ligation to DNA fragments ● TIMING 25 min

44| Prepare the following master mix in a 1.5 ml-centrifuge tube. Mix well by pipetting:

Reagent	Volume (µl)
Quick ligation buffer	25
Quick ligase	2

45| Add 6 µl of 10 µM PE adaptor mix (10 µM each of PE5/7) to each 17-µl DNA sample from Step 43.
▲ **CRITICAL STEP** Six microliters of 10 µM PE adaptor mix is the amount that has been crucially determined to work effectively when using an input DNA quantity from 500 ng to 2 µg. If less DNA is used, adaptor mix quantity would have to be adjusted. However, given that the process of WGA from single cells yields DNA on the order of 4–5 µg, there should be plenty of DNA from which to construct sequencing libraries.
▲ **CRITICAL STEP** Samples can be bar-coded and the resulting libraries can be multiplexed and run together on an Illumina lane; refer to **Box 1** for multiplexing.

46| Add 27 µl of ligation master mix from Step 44, mix well, and incubate at 20 °C for 15 min.

47| Purify each sample using the MiniElute PCR purification kit according to the manufacturer's instructions.

48| Elute each sample in 16 µl of buffer EB.
■ **PAUSE POINT** Samples can be briefly spun down and kept at −20 °C until further processing.

Size selection and gel purification of DNA adaptor ligation products ● TIMING 2.5 h

49| Prepare a 2% (wt/vol) agarose gel using 1× TAE prepared with ethidium bromide.
▲ **CRITICAL STEP** Alternatively, samples can be purified using Agencourt AMPure beads; refer to **Box 3** and **Figure 5** for purification using AMPure beads.

50| Add 4 µl of sucrose loading dye to the eluted DNA samples from Step 48 (total volume 20 µl).

51| Load 20 µl of the adaptor-ligated DNA samples into the gel wells with a DNA ladder. When loading multiple samples on the same DNA gel, make sure to leave at least one empty well between samples and one empty well between DNA ladders and samples to avoid cross-contamination.

52| Run the agarose gel at 100–120 V for ~1 h to obtain clear separation of the ladder's 200–250–300 bands.

53| Place the gel on a UV transilluminator, and by using a clean scalpel for each sample, mark the positions of the 200–300-bp products. Turn off the transilluminator and cut the gel slices at the marked positions.

54| Use the Qiagen gel extraction kit to purify the DNA for the agarose gel slices according to the manufacturer's instructions.

55| Elute the DNA in 30 µl of buffer EB.
■ **PAUSE POINT** Samples can be briefly spun down and stored at −20 °C until further processing.

PCR enrichment of adaptor-ligated DNA products ● TIMING 2.5 h

56| Add 5 µl of PE5/PE7 mixture at 10 µM (each primer) to each 30-µl DNA sample from Step 55.

57| Add 30 µl of Phusion DNA polymerase and mix by pipetting thoroughly.

Box 3 | Bead purification of sequencing libraries

Purification of DNA sequencing libraries using agarose gel electrophoresis has traditionally been the standard practice. However, when it is adapted to purify a large number of samples (for example, for multiplexing purposes), agarose gel purification becomes a limiting step because of time- and labor-intensive considerations associated with scaling. An alternative method that is rapidly being adopted in sequencing laboratories is magnetic bead purification. Many bead purification products are offered through a variety of vendors; we have adopted the Agencourt AMPure XP purification system offered by Beckman Coulter Genomics. The Agencourt AMPure XP system uses solid-phase paramagnetic bead technology in selectively enriching DNA fragments of 100 bp and larger while efficiently removing excess nucleotides, salts and enzymes. Furthermore, depending on volumetric ratios of beads to purification reactions, the Agencourt systems allows for selective enrichment of DNA fragments of particular lengths (**Fig. 5**). We routinely purify libraries using 30 μl of beads for paired-end 76 Illumina sequencing runs. Below is a description of a working protocol for the purification of a single library using the DynaMag-2 magnet (the DynaMag-2 magnet can accommodate up to 16 samples). If more samples are to be processed, for example 96 samples, the DynaMag-96 Side magnet can be used.

Adaptor ligation reactions for purification using the AMPure system

Adaptor ligation reactions for purification using the AMPure system are set up differently from ligation reactions intended for gel electrophoresis purification (as described in Steps 44–46 of the PROCEDURE). For libraries to be purified using AMPure beads, we perform the ligation reaction in a total volume of 75 μl (35 μl of quick ligation buffer, 6 μl of 10 μM adaptors (10 μM of each PE5/PE7), 2 μl of quick ligase, and 32 μl of DNA library). Also, after the completion of the ligation reaction (15 min at 20 °C), the mixture is heated at 65 °C for 15 min to deactivate the DNA ligase. Libraries are then purified using AMPure beads without prior reaction cleanup using QIAquick columns (Steps 47 and 48).

Purification of sequencing libraries using Agencourt AMPure XP system

1. Aliquot 30 μl of AMPure beads into a clean microcentrifuge tube(s).
▲ CRITICAL STEP Make sure the beads are at room temperature for at least 30 min before purification.
▲ CRITICAL STEP Make sure the magnetic beads are in suspension by gently shaking the bottle before making aliquots.
2. Transfer the adaptor ligation mixture(s) to the microcentrifuge tube(s) containing the AMPure beads and mix well by pipetting up and down 10 times.
3. Allow the beads/ligation reaction mixture(s) to incubate at room temperature for 5 min.
4. Place the microcentrifuge tube(s) onto DynaMag-2 magnet and allow the mixture(s) to stand for 5 min for efficient collection of the beads.
5. Maintain the reaction mixture on the magnet and aspirate or pipette the cleared solution carefully and discard (solution should be clear after magnetic separation, compared to brown when beads are in suspension).
6. Add 200 μl of 70% (vol/vol) ethanol to the beads in the microcentrifuge tube(s) on the magnet and gently mix by inverting the magnet a couple of times.
7. Aspirate or pipette the 70% (vol/vol) ethanol and discard it while maintaining the tube(s) on the magnet.
8. Repeat for a total of two washes.
9. Allow the beads to dry for ~3 min.
▲ CRITICAL STEP Make sure not to overdry the magnetic beads as that will result in lower yields of DNA.
10. Off the magnet, add 30 μl of buffer EB and mix thoroughly by pipetting.
11. Allow the mixture to stand at room temperature for 5 min.
12. Transfer the mixture back to the magnet and incubate it for ~3 min to separate the beads from the solution.
13. Transfer the eluant to a fresh PCR tube and proceed with library amplification.

58| Incubate the mixture in thermal cycler as follows:

Cycle	Denature	Anneal	Extend	Hold
1	98 °C for 3 0 s	—	—	—
2–12	98 °C for 10 s	65 °C for 30 s	72 °C for 30 s	—
13			72 °C for 5 min	—
14				4 °C

59| Purify the PCR amplification products using the QIAquick PCR purification kit according to the manufacturer's instructions.

60| Elute samples in 30 μl of buffer EB.
■ PAUSE POINT Samples can be spun down and stored at −20 °C until further processing.

61| Measure DNA concentrations using the NanoDrop spectrophotometer.

62| After obtaining DNA concentrations, make a 30-µl total volume dilution of the library at 10 ng µl^{-1}.

63| Run 1 µl of the 10 ng µl^{-1} library dilution on the Agilent Bioanalyzer.

64| Using the 'Peak Range' option on the Bioanalyzer, gate on the DNA library peaks between 150 and 350 bp.
? TROUBLESHOOTING

65| If necessary, dilute the samples to 10 nM, pool different libraries if applying multiplex sequencing, and send the samples for sequencing.

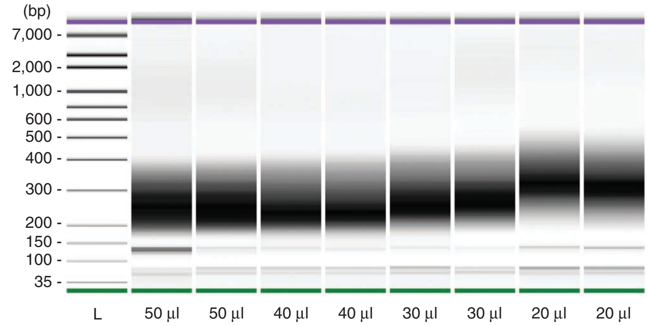

Figure 5 | Sequencing library size profiles following AMPure bead purification using different volumes of beads. The amount of beads indicated was added to a ligation reaction of 75 µl volume (50–40–30–20 µl of beads). Replicas are shown. Profiles illustrate the differing sequence library size profiles obtained when using different volumes of AMPure beads. We generally perform library purification using 30 µl of beads and sequence 76 bp on the Illumina platform. L, ladder.

Informatic analysis: prepare a reference genome for use with Bowtie ● TIMING variable

66| To begin the analysis, first the reference genome must be configured for use with bowtie. Download the hg19 reference genome from the UCSC Genome Browser: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/. Download the file named *chromFa.tar.gz*. Detailed instructions are on the web page.

67| Change the sequence of the pseudoautosomal regions on chrY to N's; a sample Python program to do so is included in the **Supplementary Methods**, labeled *hg19.chrY.psr.py*.
▲ CRITICAL STEP The sequence of the pseudoautosomal regions on chrY is an exact copy of the corresponding regions on chrX. As you will use reads that map to exactly one place in the reference genome, it is necessary to eliminate one of the two copies of the pseudoautosomal regions in order to normalize that region with the rest of the X chromosome.

68| In order to use Bowtie it is necessary to prepare an index file for the reference genome. The command to create the Bowtie index is named *hg19.bowtie.build.bash* in the **Supplementary Methods**. The files with the 'hap' annotation in their names are haplotype variants. These are not used in our copy number analysis. Also, we use our modified version of chrY rather than the reference version.

Informatic analysis: computing bin boundaries ● TIMING variable

69| A 'bin boundaries' file for 50,000 bins in hg19 is provided in the **Supplementary Data** with the title *hg19.bin.boundaries.50k.bowtie.k50.sorted.txt*. If this file is used, it is not necessary to complete Steps 69–76. Otherwise, to compute the bin boundaries, make 'reads' files from the reference genome. For the chromosomes to be used for copy number analysis, start at position one in the chromosome sequence and take the first 50 bases. Also create read ID strings and quality score strings in a format readable by Bowtie. These can be Illumina format or fastq format. Output these to a file and continue likewise at positions 2 and 3 and 4 and so on until the end of the chromosome is reached. A sample program (*hg19.generate.reads.k50.py*) is provided in the **Supplementary Methods**. The program uses the input file: *[chromlist.txt]*, which is also provided in the **Supplementary Methods**. This creates separate files, each with 150 million reads. For hg19 chromosomes 1 through 22 and X and Y there will be 21 files of this size.
▲ CRITICAL STEP Mapping three billion reads can take up to 500 h of computer time. If multiple computers are available, this mapping step can be split up and the parts distributed and run concurrently.

70| By using Bowtie, map the reads created using the same mapping parameters expected to be used when mapping real data. An example command is:

```
/filepath/bowtie-0.12.7/bowtie -S -t -n 2 -e 70 -m 1 --best --strata --solexa1.3-
quals hg19 /filepath/sequence.part.0.k50.txt /filepath/sequence.part.0.k50.sam.
```

A sample Python program for creating and submitting Sun Grid Engine jobs *(bowtie.qsub.py)* is provided in the **Supplementary Methods**.

71| Create a file listing the sizes of the chromosomes to be used for the copy number analysis. A sample program is provided in the **Supplementary Methods** *(hg19.chrom.sizes.py)*. The necessary file for further processing is also provided in **Supplementary Data** *(hg19.chrom.sizes.txt)*.

72| Genome positions with reads that map back to where they came from and nowhere else in the genome are called ‘mappable positions’. The goal is to create a set of bins, each having the same number of mappable positions. Summarize the list of mappable positions in a file with one row for each contiguous block of mappable positions. These blocks are called ‘goodzones’. A sample program for creating the list of goodzones from the mapped read files is included in the **Supplementary Methods** *(hg19.bowtie.goodzones.k50.py)*. The file listing the goodzones is also provided in the **Supplementary Data** *(hg19.goodzones.bowtie.k50.bed)*. This file is used to compute the bin boundaries.
▲ CRITICAL STEP If it is desired to create a file with more or fewer bin boundaries (e.g., 5,000 or 100,000), such a file can be computed from this goodzones file without having to recreate and remap three billion reads from the reference genome. Just start at the next step in the protocol.

73| From the goodzones file, compute the number of mappable positions on each chromosome. A sample program is provided in the **Supplementary Methods** *(hg19.chrom.mappable.bowtie.k50.py)*. The output file from this program is also provided in the **Supplementary Data** *(hg19.chrom.mappable.bowtie.k50.txt)*.

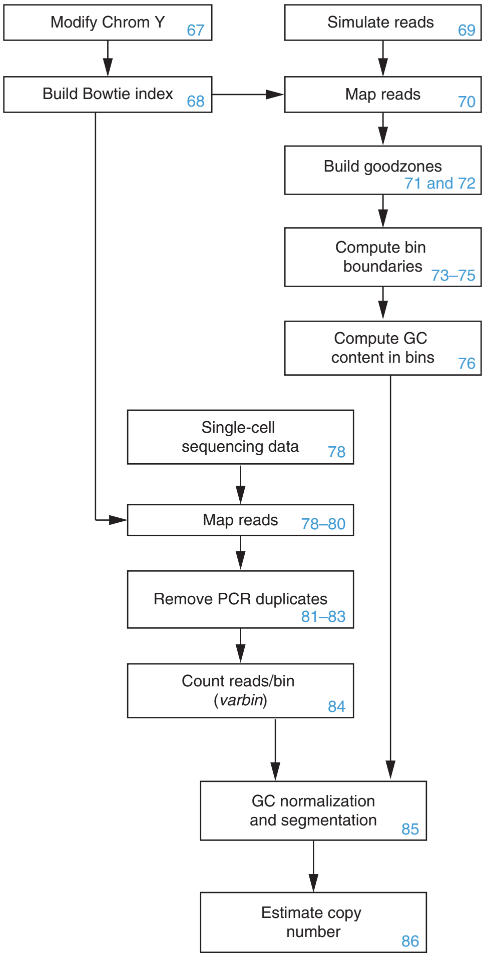
Modify Chrom Y 67
Simulate reads 69
Build Bowtie index 68
Map reads 70
Build goodzones 71 and 72
Compute bin boundaries 73–75
Compute GC content in bins 76
Single-cell sequencing data 78
Map reads 78–80
Remove PCR duplicates 81–83
Count reads/bin (*varbin*) 84
GC normalization and segmentation 85
Estimate copy number 86

Figure 6 | Schematic of the informatics workflow of SNS. Blue numbers refer to the Steps of the PROCEDURE.

74| After deciding on how many bins are desired, compute the bin boundaries from the goodzones file and the number of mappable positions in each chromosome. A number of bins are allocated to each chromosome proportional to the number of mappable positions on that chromosome relative to all the chromosomes being used in the copy number analysis. Furthermore, the number of mappable positions for each bin is computed as mappable positions divided by the number of bins, rounding up when the fractional bin accumulated passes 1 and adding one mappable position to the last bin on the chromosome if necessary. A sample program is provided in the **Supplementary Methods** *(hg19.bin.boundaries.50k.py)*.
▲ CRITICAL STEP The choice of the number of genomic bins to be used in the analysis depends on a number of factors. Segmentation algorithms generally perform better with more data points. However, the variance due to sampling is very high if the median bin count is low—below 20 reads per bin, for example. Another consideration is variation because of small-scale differences in the genome. We normalize the bin counts for each sample on the basis of GC content. This is sufficient at the scales we have been using (for example, using 50,000 or 240,000 bins), however, at a much smaller scale, for example using 2.5 million bins, GC normalization alone might not be sufficient to correct for WGA biases that might be independent of GC content. For the 50,000 bins supplied in this paper, we generally achieve a median read count of 35 reads, which is sufficient to allow genome-wide copy number determination.

75| Sort the bin boundaries file:

```
sort -k 3,3n hg19.bin.boundaries.50k.bowtie.k50.txt>
hg19.bin.boundaries.50k.bowtie.k50.sorted.txt
```

The input data for the sorting is provided in the **Supplementary Data** *(hg19.bin.boundaries.50k.bowtie.k50.txt)*. Use this sorted bin boundaries file for subsequent processing.

76| Compute the GC content in each bin. This will be used in Step 78 for GC normalization. This consists of computing the percentage of G and C bases in each bin from the reference genome. A sample program is provided in the **Supplementary Methods** *(hg19.varbin.gc.content.50k.bowtie.k50.py)*. The output file is also provided in the **Supplementary Data** *(hg19.varbin.gc.content.50k.bowtie.k50.txt)*. **Figure 6** illustrates the schematic for genome configuration and bin boundary definition as well as the steps downstream necessary to infer genome copy number.

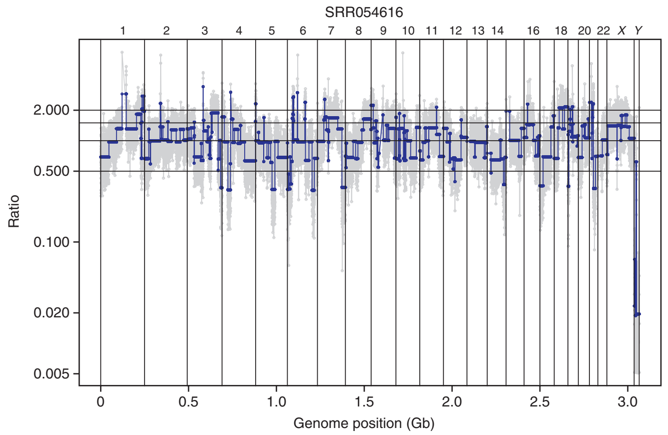

Figure 7 | Genome plot of normalized bin counts and segmentation illustrating the genome-wide copy number profile of the single-cell example SRR054616. The blue line shows the seg.mean values from the CBS.

Informatic analysis: sequence mapping and data analysis
● TIMING variable

77| If there are multiple bar-coded samples in a lane of sequence data, these must first be allocated to separate files. In our system, the bar codes are the first eight bases of each read. The eighth base position is always a T, so only the first seven positions are needed to identify the samples. A file listing the bar code sequences and bar code IDs is used to determine which are valid bar-coded sequences and to which output file they are allocated. A sample program to do this is provided in the **Supplementary Methods** *(barcode.split.sr01.py)*. A sample bar code file is also provided in **Supplementary Data** *(barcode.8.txt)*. Once the sequence data is split into separate files for each sample, carry out processing as described below.

78| Map the reads to the reference genome:

```
/filepath/bowtie-0.12.7/bowtie -S -t -n 2 -e 70 -3 0 -5 0 -m 1 --best --strata hg19
/filepath/SRR054616.fastq /filepath/SRR054616.sam
```

An example data set can be downloaded from the NCBI Short Read Archive (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=SRR054616). This data set is from a single cell from Navin *et al.*[24]. The accession ID is SRR054616.

The -3 and -5 parameters indicate how many bases to trim from the 3′ and 5′ ends of each read. The example data set was not bar coded so it is not necessary to trim bases from the 5′ end. These reads are only 36 bases in length. As the bin boundaries were computed using 50-base reads, it is desirable to map a number of bases that is as close to 50 as possible. If all the samples in a project were sequenced at 36-bp length, then it would be desirable to recompute bin boundaries with 36 base reads from the reference genome. Bases from the 3′ end of the reads can be trimmed if more than 50 bases are available for mapping to match the computation of the bin boundaries. On more recent sequencing runs it is typical to have 100-base reads. If many reads have the WGA primer sequence at the 5′ end of the read immediately after the bar code sequence, an additional 30 bases can be trimmed. The 5′ parameter would then be 38 and the 3′ parameter would be 12, leaving 50 bases to be mapped. A Sun Grid Engine script to map reads for the sample cell is provided in the **Supplementary Methods** *(SRR054616.bowtie.qsub)*.

79| Convert the output to .bam file format:

```
/filepath/samtools-0.1.16/samtools view -Sb -o /filepath/SRR054616.bam
/filepath/SRR054616.sam
```

80| Sort the .bam file:

```
/filepath/samtools-0.1.16/samtools sort /filepath/SRR054616.bam
/filepath/SRR054616.sorted
```

81| Remove reads that are likely to be PCR duplicates:

```
/filepath/samtools-0.1.16/samtools
rmdup -s /filepath/SRR054616.sorted.bam
/filepath/SRR054616.rmdup.bam
```

82| Create a .bam file index:

```
/filepath/samtools-0.1.16/samtools index
/filepath/SRR054616.rmdup.bam
```

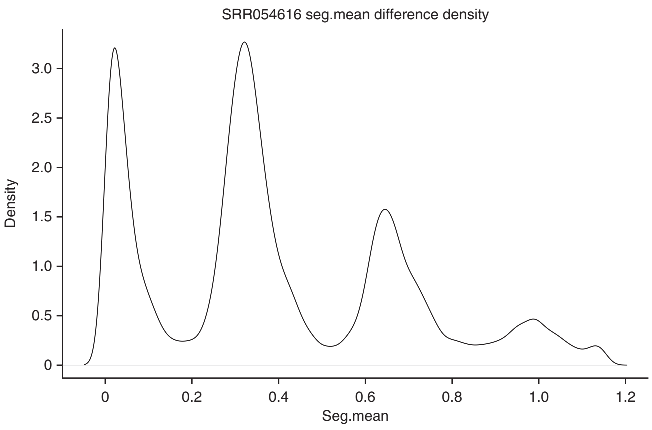

Figure 8 | Density plot of segment value differences. The density plot shows the Gaussian kernel smoothed density of differences in seg.mean values for differences between all segments called by the segmentor weighted by segment length. The second peak represents the mode of the seg.mean difference between segments one copy number apart.

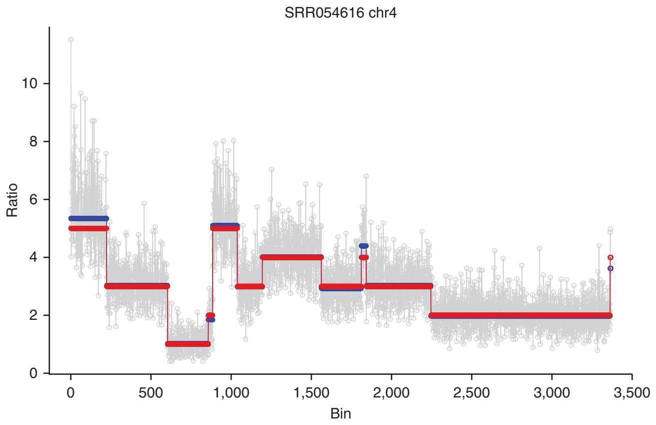

Figure 9 | A close-up view of a region on chromosome 4 illustrating normalized bin count for each bin on the chromosomal segment. The blue line is the seg.mean as called by the CBS algorithm. The red line is the estimated copy number.

83| Create a .sam file from the sorted .bam file with duplicates removed:

```
/filepath/samtools-0.1.16/samtools
view -o /filepath/SRR054616.rmdup.sam
/filepath/SRR054616.rmdup.bam
```

84| Count the number of reads in each bin:

```
/filepath/Python-2.7.1/python /filepath/
varbin.50k.sam.py /filepath/SRR054616.
rmdup.sam /filepath/SRR054616.varbin.50k.txt /filepath/SRR054616.varbin.50k.stats.txt
```

The output files of the varbin algorithm are provided in the **Supplementary Data**. A sample Python program for doing this is provided in the **Supplementary Methods** *(varbin.50k.sam.py)*

85| Run the R script provided in the **Supplementary Methods** *(SRR054616.cbs.r)* to perform the GC content normalization and CBS and plot the graphs:

```
/usr/bin/R CMD BATCH /filepath/SRR054616.cbs.r /filepath/SRR054616.cbs.r.out
```

The R script brings in the data file, adds one to each bin count, normalizes the bin count on the basis of GC content using LOWESS smoothing, uses the CBS segmentor to find nonoverlapping regions of differing copy number and outputs genome plots (**Fig. 7** and **Supplementary Data**). **Figure 7** shows a genome plot of normalized bin counts and segmentation. There is one gray point for each normalized bin count. The blue line shows the seg.mean value from CBS. The high peaks near the centromeres are artifacts of inaccurate genome assembly in the highly repetitive regions near some of the centromeres and telomeres. These bins can be masked using the file *(hg19.50k.k50.bad.bins.txt)* provided in the **Supplementary Data**. This file is a list of 50,000 zeroes and ones, with one indicating that the bin is to be masked. These 'bad bins' are from empirical observation of a number of samples sequenced in our laboratory. These are provided as an example.

86| (Optional) Run the R script provided in the **Supplementary Methods** *(SRR054616.copynumber.r)* to estimate copy number. This will only work if there are enough regions of the genome at varying copy numbers to allow the algorithm to work. For genomes that are near diploid we assume the majority of the genome is copy number two and estimate other regions based on the segment ratio relative to two:

```
/usr/bin/R CMD BATCH /filepath/SRR054616.copynumber.r
/filepath/SRR054616.copynumber.r.out
```

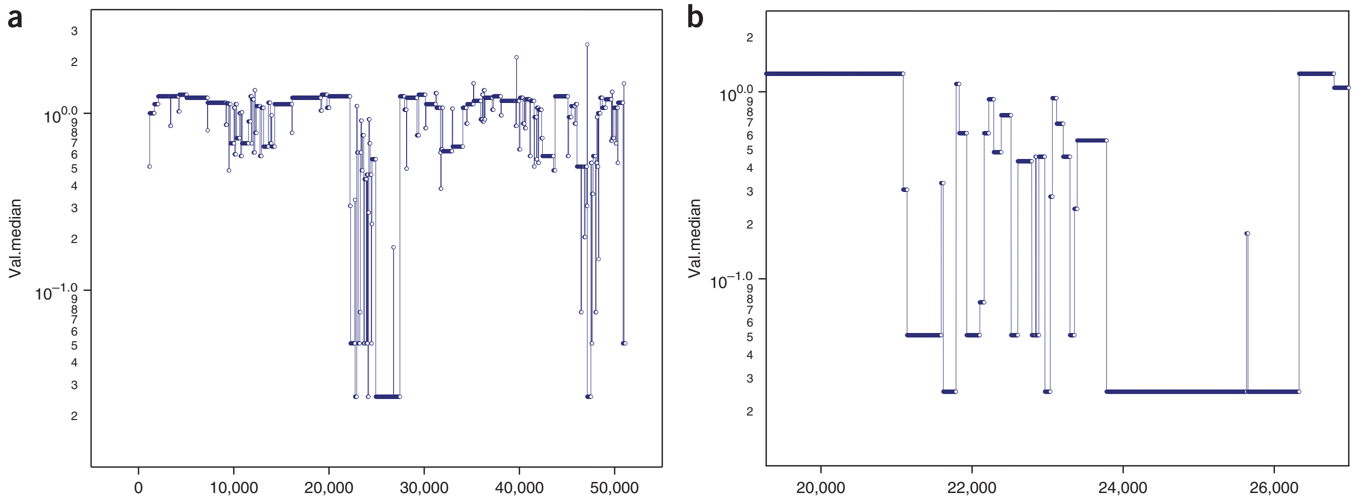

Figure 10 | Representative illustration of genome sector loss (GSL), where large homozygous deletions and patterns consistent with chromosomal 'shredding' are evident. (**a**) Whole-genome view of a single cell collected out of a 'normal' diploid flow sorting gate. (**b**) A close-up view of the profile of the same cell on chromosomes 7 and 8.

Box 4 | Genome sector loss (GSL)

In ~5% of single-cell profiles, we observe an as-yet-unexplained phenomenon in which one or more chromosomes has been either completely lost (homozygous loss), or it appears 'shredded' as if multiple regions up to 20 Mb in length from a single chromosome have been randomly lost from the nucleus. We observe this phenomenon to varying degrees in all types of samples, whether from cell culture, normal or malignant tissue. Such GSL can affect any chromosome and the breakpoints are not shared among different cells from the same source or sorting session. The profiles of these cells are highly disordered and appear distinct from those reported for 'pseudodiploid' cells in our initial publication[24]. In the absence of a biological or physical explanation for these cells, we consider them at least moribund, and although we include them in our lineage analysis, they do not contribute to the clonal lineage trees.

It is not clear whether the cause of GSL lies in the sorting process, perhaps by shear stress on the nuclei as they pass through the nozzle, or if it has a biological explanation related to abortive cell division or the observed fragmentation of chromosomes during programmed cell death (apoptosis)[29]. It is also tempting to relate the observation of shredded chromosomes among the GSL profiles to the recently reported events leading to 'chromothripsis', in which segments of shredded chromosomes reform in a highly rearranged yet viable state[30]. The potential explanations for GSL are currently under investigation. In any case, the phenomenon affects only a small minority of the profiled nuclei.

This script will output a density plot of segment value differences and plots of each chromosome showing the adjusted bin counts, segmentation values and copy number estimates (**Fig. 8** and **Supplementary Data**). The density plot shows the Gaussian kernel smoothed density of differences in seg.mean values for differences between all segments called by the segmentor weighted by segment length. The second peak represents the mode of the seg.mean difference between segments one copy number apart. This is used to estimate copy number for the genome. **Figure 9** shows a close-up view of a region on chromosome 4 illustrating the normalized bin count for each bin on the chromosome. The blue line is the seg.mean as called by the CBS algorithm. The red line is the estimated copy number.

Supplementary Data provides the output files from the copy number R script. As mentioned earlier in the text, occasionally, we observe single-cell copy number profiles that contain large homozygous deletions or what appears to be 'shredding' of chromosomes. **Figure 10** provides an illustration of those profiles and **Box 4** provides a discussion.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
30	Failure to amplify single-cell genome	Generally, problems with amplification stem from flow sorting problems	Make sure flow cytometry parameters are set properly to capture single cells in 96-well plates
		96-well plate not properly aligned for sorting; single-cell deposition device was not checked or device position was moved following alignment	Perform test sort using beads to determine that drops are deposited precisely in the center of each of the 96 wells. If necessary, use the instrument device positioning feature to make adjustments
		Break-off is not stable; sample line, flow cell or nozzle is not clean	Perform proper cleaning of instrument (refer to **Box 2**) and check for air bubbles in the sample line and the flow cell
		Break-off is not stable; room temperature has changed considerably (ambient air temperature affects the size and flight of sort droplets)	If the flow cytometry facility experiences temperature fluctuations, check break-off and drop delay settings regularly and adjust accordingly
		Break-off is not stable; fluidics pressure is not stable	Check sheath and sample pressures. Check in-line filters and tubing connections. Call instrument service engineer
		Drop delay is incorrect (the drop delay value determines which drop will be deflected); break-off has drifted	Monitor break-off and repeat drop delay if any minor changes are observed

(continued)

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
		Drop delay is not correct (the drop delay value determines which drop will be deflected); instrument sort setting was changed at some point following sort setup	Changing some sort setting values will alter the drop delay. Perform drop delay determination again
64	Low yield of enriched DNA library (too many adaptor-adaptor linkers)	Low ratio of DNA to adaptor-adaptor ligation products	Lower the amount of adaptors used in ligation. Alternatively, DNA libraries with a lot of adaptor-adaptor amplification product contaminants can be re-purified and amplified using limited cycles (e.g., four or five cycles)

● TIMING

Steps 1–8, sample preparation and flow cytometry: 4 h
Steps 9–18, WGA: 6 h
Steps 19–30, QIAquick 96-well-plate PCR purification: 1 h
Steps 31–35, DNA sonication: 30 min
Steps 36–39, end repair of sonicated WGA DNA: 45 min
Steps 40–43, 3′ A-overhang addition: 45 min
Steps 44–48, adaptor ligation to DNA: 25 min
Steps 49–55, size selection and library gel purification: 2.5 h
Steps 56–65, library enrichment and quantification: 2.5 h
Steps 66–86, informatic analysis; variable and depends on computer processing power

ANTICIPATED RESULTS

In our previous report[24], we performed, as a proof of concept, SNS on multiple single nuclei isolated from the human breast cancer cell line SK-BR-3; we compared the genome-wide copy number profiles to profiles obtained from sequencing bulk DNA from a million cells as well as profiles determined using DNA on aCGH. Copy number profiles from the different samples are highly concordant and reproducible with R^2 correlation values of ~0.9. In addition, SNS was performed on single nuclei from a diploid immortalized fibroblast cell line (SKN1) with a normal 'flat' copy number profile. The results from SNS on single SKN1 nuclei, illustrating a normal flat profile, again prove the reproducibility of the approach. Notably, when analyzing many single cells from cancer tissue specimen, as done in Navin *et al.*[24], the clustering of the copy number profiles yields evolutionary trees of tumor progression that are legible and interpretable. Furthermore, the quantitative nature of the data that are produced with the SNS method (**Figs. 7** and **9**) allows for the accurate identification of genomic copy number alterations and will help in furthering our understanding of cancer biology. Since our initial report, we have applied SNS to many additional breast tumors as well as tumors of different anatomical origins and we reproducibly obtain quantitative and intelligible genome-wide copy number profiles.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS We acknowledge all the members of the Hicks and Wigler labs for discussions during the technology development. We thank W. McCombie, E. Ghiban and L. Gelley for their advice and technical assistance, and A. Gordon for informatics support and assistance with figures. N.N. is supported by grants from Texas STARS and the Alice Kleberg Reynolds Foundation. This work was supported by grants to M.W. and J.H. from the Department of the Army (W81XWH04-1-0477), the Breast Cancer Research Foundation, and the Simons Foundation. M.W. is an American Cancer Society Research Professor.

AUTHOR CONTRIBUTIONS T.B., J.K., N.N., J.H. and M.W. designed methods and analyzed data. T.B., L.R., H.C., M.R., A.S. and J.T. performed experiments. K.R., D.E. and B.L. contributed to the technology development. J.K., N.N., T.B., J.H. and M.W. developed informatic analysis. T.B., J.K. and J.H. wrote the paper.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nprot.2012.039. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Beckmann, J.S., Estivill, X. & Antonarakis, S.E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet.* **8**, 639–646 (2007).
2. Hasin, Y. *et al.* High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet.* **4**, e1000249 (2008).
3. Perry, G.H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).

4. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
5. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
6. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
7. Bignell, G.R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
8. Russnes, H.G. *et al.* Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci. Transl. Med.* **2**, 38ra47 (2010).
9. Shiu, K.K., Natrajan, R., Geyer, F.C., Ashworth, A. & Reis-Filho, J.S. DNA amplifications in breast cancer: genotypic-phenotypic correlations. *Future Oncol.* **6**, 967–984 (2010).
10. Shinawi, M. & Cheung, S.W. The array CGH and its clinical applications. *Drug Discov. Today* **13**, 760–770 (2008).
11. Santarius, T., Shipley, J., Brewer, D., Stratton, M.R. & Cooper, C.S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).
12. Pinkel, D. & Albertson, D.G. Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.* **37** (suppl.): S11–17 (2005).
13. Praulich, I. *et al.* Clonal heterogeneity in childhood myelodysplastic syndromes—challenge for the detection of chromosomal imbalances by array-CGH. *Genes Chromosomes Cancer* **49**, 885–900 (2010).
14. Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
15. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
16. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
17. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).
18. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
19. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
20. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
21. Campbell, P.J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
22. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
23. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**, 356–361 (2011).
24. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
25. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
26. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **15**, 2078–2079 (2009).
27. Venkatraman, E.S. & Olshen, A.B. A faster cicular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–63 (2007).
28. Wersto, R.P. *et al.* Doublet discrimination in DNA cell-cycle analysis. *Cytometry* **46**, 296–306 (2001).
29. Nagata, S., Nagase, H., Kawane, K., Mukae, N. & Fukuyama, H. Degradation of chromosomal DNA during apoptosis. *Cell Death Differ* **10**, 108–116 (2003).
30. Stephans, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).