Title

DNA copy number profiling using single-cell sequencing.

Journal

Briefings in Bioinformatics, 19(5)

ISSN

1467-5463

Authors

Wang, Xuefeng
Chen, Hao
Zhang, Nancy R

Publication Date

2018-09-28

DOI

10.1093/bib/bbx004

Peer reviewed

OXFORD

Briefings in Bioinformatics, 19(5), 2018, 731–736

doi: 10.1093/bib/bbx004
Advance Access Publication Date: 4 February 2017
Paper

DNA copy number profiling using single-cell sequencing

Xuefeng Wang, Hao Chen and Nancy R. Zhang

Corresponding authors: Xuefeng Wang, Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA. Tel.: +1 813-745-6710; Fax: 813-745-6107; E-mail: xuefeng.wang@moffitt.org; Hao Chen, Department of Statistics, University of California, Davis, 1 Shields Avenue, Davis, CA 95616, Tel.: +1-530-554-1379; E-mail: hxchen@ucdavis.edu

Abstract

Currently, there is a lack of software for detecting copy number variations and constructing copy number profile for the whole genome from single-cell DNA sequencing data, which are often of low coverage and high technical noises. Here we introduce a new toolkit, SCNV, which features an efficient bin-free segmentation approach and provides the highest resolution possible for breakpoint detection and the subsequent copy number calling. SCNV can auto-tune parameters based on a set of normal cells from the same batch to adjust for the technical noise level of the data, facilitating its application to data gathered from different platforms and different studies.

Key words: change point detection; DNA copy number variation; single-cell sequencing; tumor heterogeneity

Introduction

Single-cell sequencing (SCS), characterization of the genome of individual cells, has become a widely used tool in stem cell, neuron and cancer studies [1–3]. In cancer studies, one question of tremendous interest is the detection of DNA copy number aberrations (CNAs) or copy number variations (CNVs), which are regions of the genome that undergo somatic mutations, such as deletions and duplications of some DNA segments. The study of CNV based on SCS provides new insights into cancer progression and evolution. For example, the recent large-scale analysis of >4000 single cells from oligodendrogliomas [4] has successfully identified three developmental categories of cancer cells. Tumor heterogeneity used to be a major hurdle for copy number assessment for bulk sample sequencing. The complexity of cell mixture is no longer an issue in the SCS-based analysis, as each cell is sequenced separately. With single cells as units, SCS data provide a new possibility to understand both intra-tumor and inter-tumor heterogeneity. However, there are considerable challenges for analyzing SCS data owing to the high technical noises existing in all SCS platforms, especially for low-pass sequencing. A typical SCS (DNA sequencing) workflow involves the following steps: isolation of individual cells, DNA extraction, whole-genome amplification (WGA), constructing sequencing libraries, sequencing using a next-generation sequencer and then bioinformatics data analysis. It is much more difficult to perform SCS when compared with sequencing from bulk cell population because of the extremely low amount of starting DNA materials and low tolerance for sample contamination and degeneration. There is currently no standardized method for single-cell isolation and it is still hard to capture a cell without including materials from other cells. The sequencing coverage bias is more prevalent in SCS studies owing to heavy amplification during the sample preparation, often demonstrating a nonlinear behavior. Also, each amplification platforms (further discussed in the following sections) have distinct patterns of uneven coverage and various dropout rates. Owing to these difficulties, it is impossible to make a systematic bias correction or normalization based on a single numeric model. In this study, we aim to develop a data-driven analysis toolkit for copy number profiling with SCS data across different platforms.

Xuefeng Wang is an Assistant Member of Biostatistics and Bioinformatics at H. Lee Moffitt Cancer Center and Research Institute and a visiting faculty member at Yale School of Medicine. His research focuses on computational and statistical modeling in cancer genomics.
Hao Chen is an Assistant Professor of Statistics at UC Davis. Her research interests include change-point analysis for high-dimensional and non-Euclidean data, categorical data analysis, allele-specific copy number variation analysis and single-cell sequencing data analysis.
Nancy R. Zhang is an Associate Professor of Statistics and co-director of PhD Program at the Wharton School Statistics Department, University of Pennsylvania. Her research interests include methods for change-point detection, model and variable selection and scan statistics.
Submitted: 20 October 2016; **Received (in revised form):** 6 December 2016

Copy number assessment commonly starts with segmentation that separates the genome into non-overlapping regions with the hope that each region is homogeneous in copy number, i.e. the CNA events happen at the boundaries of the segments, which are often referred to as breakpoints. There are many segmentation algorithms for processing array-based data, and some for bulk sequencing data, while there is a lack of algorithms specifically designed for SCS data. Both bulk sequencing and single-cell sequence produce read-depth (RD) data, which is the number of reads mapped to the reference genome at each location. In the ideal scenario, the RD is proportional to the DNA copy number at the corresponding location. However, in reality, many other factors influence the RD, such as GC content and mappability. In SCS, as the amount of DNA in one cell is small, an amplification step is typically done before sequencing. This step further introduces biases to the RD. On the other hand, even with the amplification step, the RD in current SCS data is still much less than that from typical bulk sequencing. In other words, the signal-to-noise ratio is much lower in SCS data than that in bulk sequencing data. Moreover, in bulk sequencing data analysis, it is widely used to control noises using normal cells [5, 6]. In SCS experiments, when cells are extracted from a tumor, it is often unclear on the cell status, i.e. which cells are normal cells and which ones are cancer cells. We follow the protocol in Baslan *et al.* [7] to determine the cell status. These initial analysis procedures were embedded into our toolkit SCNV.

In this article, we will focus our discussion on a fine-resolution segmentation scheme provided in SCNV, based on a window/bin-free algorithm. The bin-based segmentation algorithm is a convenient method to handle the sparse and noisy single-cell data as binning naturally and effectively smooth the read counts. However, bin-based methods could fail in detecting break points that are close to the middle of the bin and the resolution is restricted by the bin size. More importantly, it is hard to decide an optimal bin size that retains good sensitivity and specificity across different scenarios of sequencing depth [8] and different SCS platforms. It is, therefore, essential to appropriately gauge the more sophisticated but more accurate alternative, bin-free algorithm, to single-cell data.

Copy number profiling with bulk and SCS

The first step in DNA copy number analysis is usually the detection of locations in chromosomes where the copy number changes, often referred to as segmentation in the field of genomics. Various methods for segmentation have been developed in the context of array data, with either comparative genomic hybridization or single nucleotide polymorphism probes. In all microarray studies, proper normalizations were done to remove technical and experimental artifacts before segmentation. The noisy intensity measurements are then translated into non-overlapping segments of nucleotide positions that are likely to share the same copy number in the segmentation step. Segmentation algorithms were developed based on a rich body of breakpoint detection or smoothing techniques such as Circular Binary Segmentation (CBS) [9, 10] and Hidden Markov Model (HMM) [11, 12].

Over the past few years, copy number profiling from next-generation sequencing (NGS) data are emerging as a key technique for assessing genome aberrations, especially in cancer research. The two most commonly used methods with NGS data are RD method and paired-end mapping (PEM), each with their strengths and weakness. The RD method is implemented by counting and comparing reads mapped to each genomic region, while the PEM method is based on mining discordantly mapped paired-end reads with distances significantly different from the average insert size. Compared with the PEM method, the RD method provides better performance in capturing large CNV events and is thus more suitable for the initial analyses in tumor sample profiling. The core segmentation algorithms developed for array data were applied to NGS data analysis with modifications that allow read counts or ratios as the measurement. However, systematic biases inherent to the sequencing data, such as mappability and GC content, need to be properly adjusted. In several recent studies, a matched normal sample is used to account for location-specific biases achieve satisfying results [4, 5].

SCS technologies allow us to study cancer genomics at a level previously impossible. As each cell is sequenced separately, SCS promises to bring unprecedented resolution and accuracy to cancer CNA analysis. In addition, the copy number at any location must be an integer, as the sample is pure now (versus a mixture in the bulk sequencing). Because extremely low-coverage sequencing could still capture major CNA events, such design allows population studies at affordable costs. Given that the dynamics of many cancers such as prostate cancers are dominated by copy number alterations, the genome-wide single-cell CNA detection of large numbers of tumor cells could serve as a valuable and feasible step to identify cancer driver events and to understand cancer progression and evolution.

One fundamental challenge in SCS data processing is to cope with the high noise in the data. Numerous technical errors can be introduced during three major steps: cell isolation, genome amplification and sequencing. As illustrated in the Supplementary Figure S3, loss of coverage and extremely non-uniform RD across the genome are often observed in raw SCS coverage data. Many platforms provide more coverage and thus much higher RD around telomeric and centromeric regions. The uneven coverage has a larger impact on CNV calling than SNV calling because copy number inference with low-pass sequencing data is directly based on changes in RD. The noise model is currently poorly explored, as to date no gold standard method has been developed to benchmark or even simulate the single-cell DNA data. Meanwhile, WGA and sequencing platforms and protocols are rapidly evolving. WGA is a critical component of SCS data because replicating DNA from a single cell introduces a substantial amount of technical noise. Three most widely used WGA methods are multiple-displacement amplification (MDA), degenerate oligonucleotide primed polymerase chain reaction amplification (DOP-PCR) and hybrid methods such as looping-based amplification (MALBAC) [13, 14]. DOP-PCR and MALBAC methods and their variants generate lower physical coverage of the genome than MDA, but more uniform amplification [13, 15]. A recent report found that DOP-PCR had more consistent results than MALBAC for copy number profiling [16], although it provides lower genome coverage among the two. All these methods produce uneven coverage in certain genomic regions, e.g. near-centromeric regions. The optimal method that provides both good coverage and uniformity is yet to be developed. Therefore, parameters that affect sensitivity and specificity of the segmentation algorithm need to be carefully tuned to specific types of amplifiers. To reduce the number of false CNV calls, it was recently suggested to use the overlap results from different segmentation methods such as CBS and HMM [17] in single-cell CNV detection. However, the optimal algorithm parameters need to be determined and validated

experimentally, which makes it hard to be applied to other studies directly.

In the following, we present a refined data processing and analysis pipeline that allows for more efficient and accurate tumor CNA profiling analysis for SCS data, enabling users to better exploit the information contained in the current single-cell DNA sequencing data.

New software SCNV and benchmark results

The proposed toolkit, SCNV, provides a complete spectrum of functions for single-cell CNA analysis starting from BAM files, where all raw single sequencing reads (FASTQ files) are properly aligned to the reference genome. If the cell status is unknown, which is the usual case, an initial copy number assessment modified from the protocol proposed by Baslan *et al.* (2012) is conducted. This initial calling step is bin-based and provides a fast and approximate profiling on the CNA status of each cell. Based on these preliminary results, we can identify normal cells and potential tumor cells. This also provides an easy way to detect poor-quality samples. A more detailed description of this step is described in 'Single-cell copy number profiling analysis overview' section.

With the normal cells and tumor cells from the same batch, we then use a bin-free algorithm to do a refined CNV calling. We adapt the algorithm, SeqCBS [5], which was designed for bulk sequencing data, to SCS data. We made the following three main adjustments. (1) We pool a certain number of normal cells, e.g. 20 normal cells, together as the control to the tumor cells. This is because the coverage for SCS data is low, usually of $<0.1\times$ with total reads around 1–10 million, using a composite of normal cells, as the control produces more reliable results than a single-sample control. (2) An additional calibration step is introduced to reduce falsely detected CNV events. We use a number of normal cells from the same batch to tune a number of cutoffs we designed for SCS data. This calibration step could adjust the method to the right noise level and eliminate many false discoveries. (3) Copy numbers estimated in each segment are discretized because they are always integers in single cells. A schematic diagram of the data analysis pipeline implemented in SCNV is shown in Figure 1.

We use spike-in experiments to study the achievable performance of the proposed approach on low-coverage SCS data. The design of the spike-in experiment is shown in Figure 2. To mimic the noise level in real SCS experiments, we extract reads from randomly selected cells from a true tumor sample. The average RD (for the entire genome) is around $0.02\times$. The simulation set-up is as follows: we selected 400 segments with no copy number change, each with a length of 10 Mb. At the location of 5 Mb, we inserted a loss segment (randomly selected from the known loss segments pool) with length varying from 0.1 Mb to 2 Mb (the X-axis in Figure 3). Sensitivity is calculated as the proportion of segments with detected copy number signals out of the 400 segments; specificity is calculated as one minus the proportion of false calls out of all calls. Figure 3 shows the performance of the algorithm with and without having the calibration step. It is clear from the plots that the calibration step improves the specificity a lot with little sacrifice in sensitivity for short signals (<1 Mb) and almost no sacrifice in sensitivity for signals slightly longer (>1.5 Mb).

This spike-in experiment also provides a clue to the best resolution the SCS-based CNA assay could reach given the current coverage. Our method provides acceptable performance for detecting the CNV event as short as 0.3 Mb. Therefore, current coverage is dense enough to detect most CNA events in cancer; but deeper sequencing should be considered if finer resolution is desirable. As the SCS technology is fast evolving and becomes more accessible, the proposed algorithm provides a valuable tool for users to investigate the coverage to power relationship in the experiment preparation stage.

For illustration purpose, we applied the method to a brain cell population [17] and the prostate cancer data we used for simulation. Single cells are from the same individual in each data set. The brain cell data were downloaded from the National Center for Biotechnology Information Sequence Read Archive using accession number SRP041670. In the prostate data, we pooled 20 normal cells as control and calibrated the model through an additional 11 normal cells. We then applied the calibrated model to nine tumor cells from the same batch. A histogram of raw estimates of the copy numbers is shown in Supplementary Figure S4A. These raw estimates all peak around integer values, which is expect for single cells, so we can easily round them to their nearest integers to get absolute copy number state. The full copy number profiles for these nine tumor cells are shown in Supplementary Figure S1. Similarly, in analyzing the brain cell data, we pooled 18 diploid cells provided in the study (from the same individual) as a composite control. We analyzed four brain cells with reported copy number events and compared results with the findings in the original publication (Supplementary Figure S2). We obtained consistent results. But note that our segmentation analysis (based on SCNV) was performed on raw data, without removing any low-quality regions or correcting for GC content and mappability bias. A limitation of the current package is that the segmentation was designed for profiling one cell at each time. More sophisticated cross-sample models can be developed and implemented in the future to enhance information usage, as cells from the same tumor share similar CNV events.

Single-cell copy number profiling analysis overview

In this section, we discuss in more details the method we implemented in SCNV. Assuming that all raw SCS reads (FASTQ files) are properly aligned to the reference genome (as BAM formatted files), the analysis scheme starts from an initial copy number assessment from a control-free segmentation procedure. Reads are first binned in every 100 kb, and the initial segmentation partitions the genome into segments based on the number of reads in the bins using the CBS algorithm implemented in the DNAcopy [9]. We choose CBS for the raw data segmentation because it is computationally efficient compared with many other existing alternatives and it has been pilot tested and shown to be promising in previous single-cell studies [7, 18]. Note that, even with higher coverage sequencing data, we still recommend the 100 kb windows size to start with because it preserves sufficient resolution for an exploratory analysis of breakpoints while remaining computationally efficiency. To control for mappability, the genome was divided into small regions with an equal number of mappable positions. A lowess regression based normalization is further applied to adjust for GC content bias. Detailed steps are described in a protocol described in [7]. This initial segmentation step has a 2-fold purpose. First, it translates the noisy read intensity into smoothed region and provides a fast and approximate profiling on the CNA status of each cell, which aids in identifying potential normal cells for controls in later steps. Second, it provides an easy way to perform quality checks, where poor quality samples are indicated by disorderly and unsystematic changes in segment values (RD

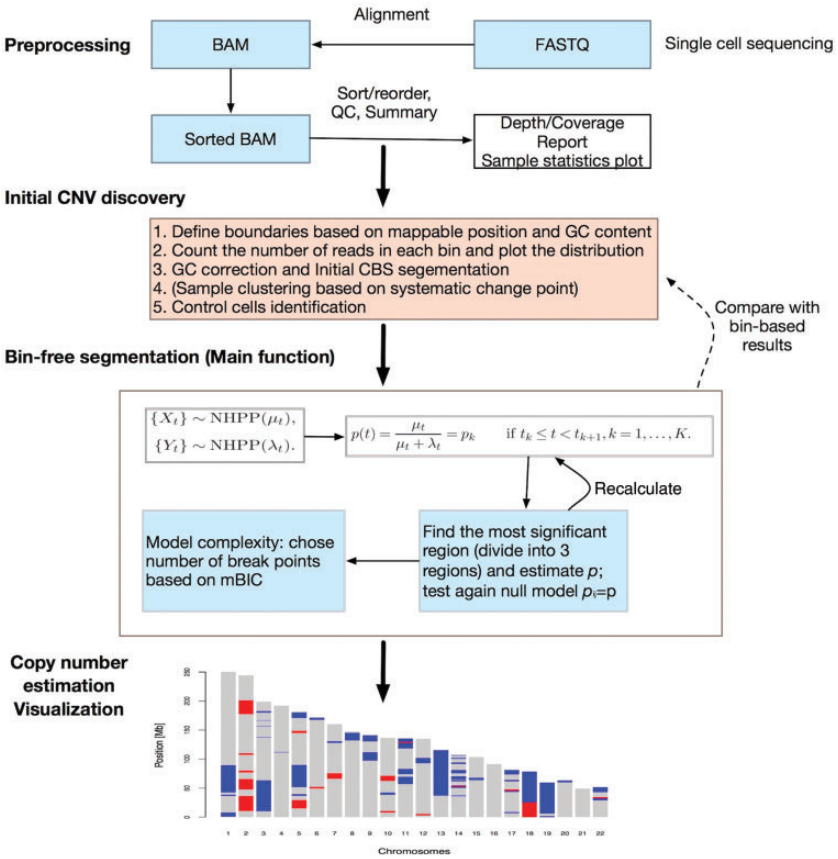

Figure 1. SCS CNV profiling analysis overview (NHPP: non-homogeneous Poisson process; mBIC: modified BIC). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

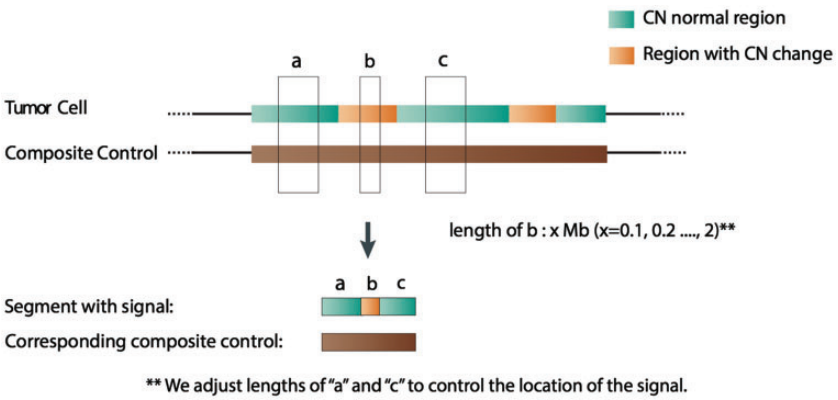

Figure 2. Schematic representation of the spike-in simulation. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

ratio). To establish a quality control threshold, a total variation score is calculated for each cell sample based on the L2 norm of the 'error' vector, i.e. the Euclidean distance of the vector of segment values to the vector of the baseline values. Another useful visual tool for checking the global CNA profile is through the smoothed density plot of the segment values. Usually, a high-quality sample exhibits a few (two or more) peaks in the density plot, with centers around the true possible copy numbers in the genome, while poor quality samples do not have these nice patterns.

After identifying normal cells from this initial step, we pool a set of normal cells together to serve as the composite control. We adapted the model proposed in Shen and Zhang (2013) designed for bulk sequencing. In the model, the read counts are modeled by a Poisson process along the genome with the locational mean reflecting the underlying true copy number and location-specific biases. Making use of the relationship between Poisson distribution and Binomial distribution—a Poisson distributed random variable ($X \sim Poi(\lambda)$) conditioning on the sum of itself and another independent Poisson random variable ($Y \sim Poi(\mu)$) follows Binomial distribution ($X|X+Y \sim Bin(X+Y, \lambda/(\lambda+\mu))$—the read count from the tumor sample conditioning on the sum of itself and the read count of the same genome location from the matched normal sample follows Binomial distribution with the success probability the ratio of the true copy

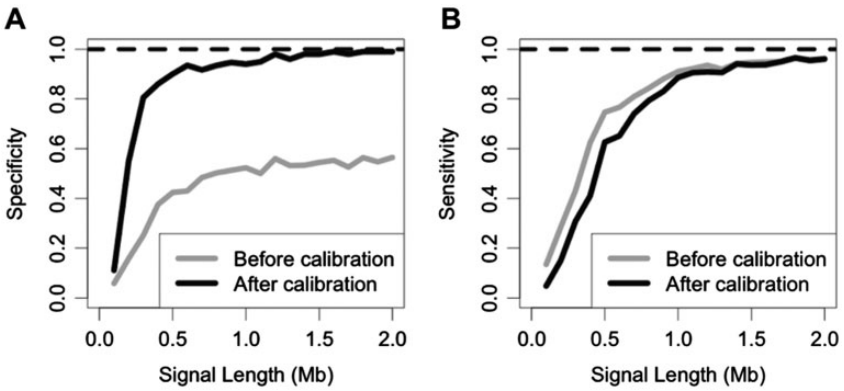

Figure 3. Performance: (**A**) specificity and (**B**) sensitivity, of SCNV bin-free CNA calling method in *in silico* spike-in experiments with and without the calibration step under different signal lengths.

number of the tumor sample at that specific location and that number plus 2, where 2 is the true copy number of the normal sample because the location-specific biases cancel out. Therefore, the change in copy numbers in the tumor sample exhibits as the change in the success probability of the Binomial process. This whole idea works properly for whole-genome bulk sequencing. However, owing to the high noise level, the Poisson distribution assumption does not fit the data for SCS well. In particular, the data exhibit over-dispersion, dropout and other amplification biases, and directly applying the approach for bulk sequencing result in an elevated rate of false discoveries. To address this problem, we use a small set of normal samples to calibrate the model first: we treat these normal cells as if they were cancer cells and run the algorithm with the composite control selected earlier—using the original stopping rule. We then select thresholds for the test statistics so that 95% of the detected CNV events could be removed. These thresholds are then applied in analyzing the tumor samples. This approach does not find the most suitable distribution to model the noise in the SCS data, but it has two key advantages: (1) the approach inherits the nice property of Binomial distribution that the distributional parameters can be estimated analytically and thus the approach scans the genome fast; and (2) the approach can be adapted to incorporate data from different platforms even though they may have different noise patterns.

We also used the spike-in experiments to examine the sensitivity and specificity of the bin-free approach in SCNV when the signal is not in the middle of the sequence. We place the signal at the one-quarter position of the segment, i.e. the length of the normal region is 2.5 Mb to the left of the signal region and 7.5 Mb to the right. The results are shown in Supplementary Figure S5. Comparing with Figure 3, we see that the location of the signal does not affect the sensitivity and specificity much.

Cell ploidy estimation

The estimation of the tumor's purity and it overall ploidy is an important component of bulk sequencing analysis in cancer, in which the two parameters can be jointly estimated based on observed copy number profiles using software such as ABSOLUTE [19] or our previously developed tool CLOSE [20]. In SCS, only the ploidy parameter needs to be inferred because the purity parameter can be fixed at one unless the cell is contaminated. However, the equation is still unable to be solved directly for the ploidy because only RD or relative ratio can be observed while bulk sequencing has additional information such as loss of heterozygosity. In some SCS experiments, the ploidy and copy-number state can be determined experimentally by flow cytometry assay comparing a cell's fluorescence activity with a reference cell. When such data are not available, SCNV provides a function to infer ploidy based on a damped sine wave plot. This function is a modified version of the numeric optimization approach as implemented in Ginkgo [16]. The rationale behind this method is that the copy-number state of all segments in a single cell should be an integer value; thus, the ploidy can be estimated by minimizing a cost function measuring the difference between the scaled copy number (multiplied by a grid multiplier) and its integer rounded copy number. If a cell is diploid, the expected plot of the empirical cost function against grid multipliers shows a typical damped sine wave (Supplementary Figure S4B). The magnitude of cost functions can also be used as a side index to evaluate the sequencing quality. An alternative approach is to interrogate the pairwise copy number differences of all segmented mean values [7], which is, however, sensitive to the smoothing or bandwidth parameter selection in constructing the density plot of differences.

Conclusion

In this article, we have presented a bioinformatics procedure for DNA copy number profiling with SCS data. The proposed method as implemented in SCNV provides the highest resolution possible for breakpoint detection and subsequent copy number calling with SCS. Importantly, it auto-tunes parameters through normal cells to adjust for the technical noise level of the data, facilitating its application to data gathered from different labs or under different conditions. As far as we know, SCNV is currently the first software that can perform bin-free segmentation for SCS data. We hope to provide a more accurate way to understand both intra- and inter-tumor heterogeneity. Still, many challenges remain. It is technically difficult to normalize noisy RDs generated from single-cell WGA. More research should be done in the analysis at the global analysis, where a hierarchical clustering can potentially be applied based on systematic change points that are detected across all samples. The global analysis can be conducted based on the bin-based segmentation for the preliminary profiling and quality control purpose, and then based on the finer bin-free segmentation for more accurate genome profiling that could serve as inputs for downstream analysis, such as to infer the tumor clonal

structure and the evolutionary trajectory through clustering and constructing the phylogenetic tree.

A statistically more rigorous method may also be developed to incorporate the over-dispersion issues in the Poison process explicitly in the model. Perhaps the greatest challenge comes from the fact that the SCS technology reinvents itself more rapidly than methods can be developed and thoroughly tested, or even than a valid benchmark data set can be set up for simulation or comparison purpose. Nevertheless, we hope this article provides a useful framework and generic pipeline for implementing single-cell CNV segmentation, making the SCS data more accessible to analysis. It will also be an interesting topic to extend our method to single-cell transcriptome sequencing. The gene expression level and copy number estimation in single-cell RNA sequencing are confounded because they are both based on RD counting. However, large-segment CNA signals can be differentiated from gene expression levels based on the assumption that the chance of all neighboring genes is up- or down-regulated is small. When the sequencing cost further reduces and deep sequencing becomes more accessible, both RD and allelic imbalance (B-allele frequencies) can be incorporated into a unified model to further improve the segmentation and copy number estimation [19, 20].

Key Points

- Single-cell DNA sequencing is promising for inferring CNAs and tumor clonality and heterogeneity.
- The new software SCNV provides an accurate and computational efficient tool for single-cell copy number profiling, which is the first tool that performs binless segmentation with SCS.
- Further research is needed to differentiate technical artifacts from true CNV signals and tumor heterogeneity under various WGA platforms and to perform global copy number profiling of cancer genomes.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

The NIH grant (P20 CA192994 to X.W., in part) and NSF award (DMS-1513653 to H.C., in part).

References

1. Navin N, Kendall J, Troge J, *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**:90–4.
2. Cai X, Evrony GD, Lehmann HS, *et al*. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep* 2014;**8**:1280–9.
3. Editorial. Method of the year 2013. *Nat Methods* 2014;**11**:1.
4. Tirosh I, Venteicher AS, Hebert C, *et al*. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* 2016;**539**:309–13.
5. Shen JJ, Zhang NR. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann Appl Stat* 2012; 476–96.
6. Chen H, Bell JM, Zavala NA, *et al*. Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res* 2015;**43**:e23.
7. Baslan T, Kendall J, Rodgers L, *et al*. Genome-wide copy number analysis of single cells. *Nat Protoc* 2012;**7**:1024–41.
8. Gusnanto A, Taylor CC, Nafisah I, *et al*. Estimating optimal window size for analysis of low-coverage next-generation sequence data. *Bioinformatics* 2014;**30**:1823–9.
9. Venkatraman E, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 2007;**23**:657–63.
10. Olshen AB, Venkatraman E, Lucito R, *et al*. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004;**5**:557–72.
11. Shah SP, Xuan X, DeLeeuw RJ, *et al*. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* 2006;**22**:e431–9.
12. Wang K, Li M, Hadley D, *et al*. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007;**17**:1665–74.
13. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 2016;**17**:175–88.
14. Navin NE. Cancer genomics: one cell at a time. *Genome Biol* 2014;**15**:1.
15. Huang L, Ma F, Chapman A, *et al*. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu Rev Genomics and Hum Genet* 2015;**16**:79–102.
16. Garvin T, Aboukhalil R, Kendall J, *et al*. Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* 2015;**12**:1058–60.
17. Knouse KA, Wu J, Amon A. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Res* 2016;**26**:376–84.
18. Zhang C, Zhang C, Chen S, *et al*. A single cell level based method for copy number variation analysis by low coverage massively parallel sequencing. *PLoS One* 2013;**8**:e54236.
19. Carter SL, Cibulskis K, Helman E, *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;**30**:413–21.
20. Wang X, Chen M, Yu X, *et al*. Global copy number profiling of cancer genomes. *Bioinformatics* 2016;**32**;926–8.