# SCCNV: A Software Tool for Identifying Copy Number Variation From Single-Cell Whole-Genome Sequencing

*Xiao Dong[1]\*, Lei Zhang[1], Xiaoxiao Hao[1], Tao Wang[2] and Jan Vijg[1,3]\**

[1] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, United States, [2] Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, United States, [3] Center for Single-Cell Omics in Aging and Disease, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Identification of *de novo* copy number variations (CNVs) across the genome in single cells requires single-cell whole-genome amplification (WGA) and sequencing. Although many experimental protocols of amplification methods have been developed, all suffer from uneven distribution of read depth across the genome after sequencing of DNA amplicons, which constrains the usage of conventional CNV calling methodologies. Here, we present SCCNV, a software tool for detecting CNVs from whole genome-amplified single cells. SCCNV is a read-depth based approach with adjustment for the WGA bias. We demonstrate its performance by analyzing data obtained with most of the single-cell amplification methods that have been employed for CNV analysis, including DOP-PCR, MDA, MALBAC, and LIANTI. SCCNV is freely available at https://github.com/biosinodx/SCCNV.

Keywords: single-cell whole-genome sequencing, single-cell whole-genome amplification, amplification bias, copy number variation, software development

## INTRODUCTION

Each single cell in a tissue or cell population has its own unique genome due to accumulating *de novo* mutations, such as single-nucleotide variations (SNVs), structural variations (SVs), copy number variations (CNVs) and aneuploidies. The frequency and spectrum of the mutations reflect the loss of genome integrity of a cell population, critically important to cancer and aging (Vijg and Dong, 2020). To detect the mutations unique to a single cell, single-cell whole-genome sequencing (SCWGS) is necessary. SCWGS requires whole-genome amplification (WGA), which is often biased, leading to uneven distribution of DNA content across the genome or differences between alleles. This essentially constrain the usage of variant callers designed for non-amplified bulk DNA. We recently developed a new software tool, SCcaller, that uses heterozygous SNPs to correct for the allelic bias hampering SNV calling (Dong et al., 2017).

CNV calling is typically based on variation of sequencing depth across the genome. However, for a single cell amplicon, variation of sequencing depth increases dramatically due to the locus-specific amplification bias (Navin et al., 2011; Zong et al., 2012; Chen et al., 2017). To solve this issue computationally, we developed SCCNV, a software tool to identify CNVs from SCWGS. SCCNV is also based on a read-depth approach: it controls not only bias during sequencing and alignment, e.g., bias associated with mappability and GC content,

but also the locus-specific amplification bias. We demonstrate the performance of SCCNV using SCWGS data of multiple experimental protocols, i.e., DOP-PCR (degenerative-oligonucleotide PCR), MDA (multiple displacement amplification), MALBAC (multiple annealing and looping–based amplification cycles), and LIANTI (linear amplification via transposon insertion) (Navin et al., 2011; Gundry et al., 2012; Zong et al., 2012; Chen et al., 2017; Dong et al., 2017).

## MATERIALS AND METHODS

## SCCNV

Our software tool for analyzing single-cell copy number variation (SCCNV) was written in Python. Its source code is freely available with a usage description and an example at Github[1] under the GNU Affero General Public License v3.0. It uses SCWGS data after alignment as input (i.e., a bam file per single cell). Of note, SCCNV cannot take sequencing data of a pool of single cells (a bam file composed of thousands of single cells data), e.g., the $10\times$ Genomics single-cell copy number data, as input.

First, SCCNV divides the genome into bins of equal size (500 kb as default), and counts the numbers of reads per bin of a cell. This step is relatively time-consuming, and we suggest users to use samtools on a high-performance computer cluster in parallel for all samples to be time-efficient (see instructions on Github). The remaining major steps of SCCNV do not require much computational resources – most modern desktop computers should work well.

SCCNV then normalizes mappability, which indicates the efficiency of the alignment to a genomic region. For a bin $b$ of a cell, SCCNV adjusts the raw number of reads, denoted by $NR_{raw}$, by dividing over the mappability $M$,

$$NR_{\mathrm{map},b} = NR_{\mathrm{raw},b}\big/M_b \tag{1}$$

where mappability $M$ is a value ranging from 0 to 1. SCCNV uses Encode Align100mer mappability score, downloaded from the UCSC genome browser, and calculates the mappability of each bin by using their weighted average.

Then, SCCNV normalizes for GC content. For a cell, SCCNV calculates the percentile of GC content of each bin. For a bin $b$ of the cell, its number of aligned reads after normalizing GC content, $NR_{GC,b}$, is,

$$NR_{GC,b} = NR_{\mathrm{map},b} \times NR_{\mathrm{map,genome}}\big/NR_{\mathrm{map},b,\mathrm{percentile}} \tag{2}$$

where $NR_{map,genome}$ is the average $NR_{map}$ per bin of all bins from the cell; $NR_{map,b,percentile}$ is the average $NR_{map}$ per bin of bins in the same GC percentile as bin $b$.

After normalization for mappability and GC content, a pattern of sequencing read depth emerges that is consistent across different cells amplified using the same experimental protocol, i.e., the locus-specific amplification bias. Therefore, the bias is normalized across all cells in a particular batch and experiment. First, to make the $NR_{GC,b}$ comparable across cells, SCCNV

converts it to a raw copy number estimate, denoted by $CN_{raw,b}$ for bin $b$ of cell $c$, as follows,

$$CN_{\mathrm{raw},b,c} = NR_{GC,b,c}\big/NR_{GC,\mathrm{genome},c} \times \mathrm{ploidy} \tag{3}$$

where $NR_{GC,genome,c}$ is the median $NR_{GC,c}$ per bin in the genome of cell $c$; ploidy is 2 by default. Second, the adjusted copy number is estimated as,

$$CN_{\mathrm{adjusted},b,c} = CN_{\mathrm{raw},b,c}\big/CN_{\mathrm{raw},b,-c} \times \mathrm{ploidy} \tag{4}$$

where $CN_{raw,b,-c}$ denotes the average $CN_{raw}$ for bin $b$ across all cells except cell $c$. Of note, with this step SCCNV aims to discover the difference between the cell $c$ and the other cells. When analyzing CNVs of multiple tumor cells, it is not appropriate to use all tumor cells as input of SCCNV; instead, one should use one tumor cell with two or more normal diploid cells as the input.

Then SCCNV uses a sliding window approach to further minimize amplification noise. By default, a window includes 11,500-kb bins, i.e., 5.5 Mb of DNA sequence in total, with a 500-kb step size between two neighboring windows,

$$CN_{smoothed,b,c} = \frac{1}{11}\sum_{i=b-5}^{b+5} CN_{adjusted,i,c} \tag{5}$$

SCCNV then models the distribution of $CN_{smoothed,b,c}$ of all bins in autosomes of a cell $c$ as a normal distribution $N(\mu, \sigma_c^2)$. The $\mu = 2$, and $\sigma$ is estimated as,

$$\sigma_c = |CN_{\mathrm{smoothed},30.9\%,c} - \mu| + |CN_{\mathrm{smoothed},69.1\%,c} - \mu| \tag{6}$$

where $CN_{smoothed,30.9\%,c}$ and $CN_{smoothed,69.1\%,c}$ are the 30.9 and 69.1% percentiles of the $CN_{smoothed,b,c}$ of all bins in the autosomes, corresponding to the $\mu - 0.5\sigma$ and $\mu + 0.5\sigma$ percentiles, respectively. Here, we did not use the observed s.d. of $CN_{smoothed,b,c}$ of all the bins because the normal distribution was to estimate amplification noise, not real variation in copy number across the genome. When a cell has several large CNVs, the s.d. will be high, even if its amplification noise remains low.

Assuming equally likely priors, for a bin $b$ and a given possible copy number $k \in \{0, 1, 2, 3, 4\}$, its posterior probability is,

$$P(H_k|x) = f_k(x)\bigg/\sum_{i=0}^{4} f_k(x) \tag{7}$$

where $x$ is the $CN_{smoothed,b,c}$, and $f_i(x)$ is the probability density function of a normal distribution,

$$f_i(x) = \frac{1}{\sigma_c\sqrt{2\pi}} \exp(-\frac{(x-k)^2}{2\sigma_c^2}) \tag{8}$$

where the variance $\sigma_c^2$ is calculated according to Eq. (5). We only used $k \in \{0, 1, 2, 3, 4\}$ because the final copy number call was after multiple testing correction, i.e., Eq. (9) below, and we wished to minimize the number of hypotheses tested, i.e., five for a copy number of 0–4. However, this will result in an underestimation if the real copy number exceeds four. To resolve this issue, for bins with copy numbers $\geq 4$ and $\leq 100$, SCCNV reports the closest integer to the $CN_{smoothed,b,c}$.

SCCNV allows <1 false positive per cell. Therefore, it determines bin b as a copy number variant when,

$$P(H_k|x) \geq 1 - \frac{1}{\text{GenomeSize}(3.2\,GB)/} = 0.998$$
$$\text{WindowSize}(5\,Mb) \qquad (9)$$

## Sensitivity and False Positive Rate

To determine copy number, SCCNV is based on a statistical test described in equations (8) and (9) for a normal distribution and multiple testing correction separately. With a given value of coefficient of variation (CV) of $CN_{smoothed,b,c}$, sensitivity and FPR can be estimated as follows. Sensitivity equals the difference between two cumulative distribution functions (CDFs) of Eq. (8) at the upper and lower boundaries, which SCCNV provides the correct CNV call after the correction in Eq. (9). The percentage of FP out of all bins is equal to the sum of (a) CDF at the lower boundary of SCCNV providing an incorrect CN gain call; and (b) 1 – CDF at the upper boundary of SCCNV providing an incorrect CN loss call. Then FPR was estimated as the ratio of % of FP to the sum of % of FP and % of TN.

For example, under the assumption that the true copy number is 2, if SCCNV calls CN = 2 when $CN_{smoothed,b,c}$ is between 1.8 and 2.2, sensitivity = CDF($x$ = 2.2, $\mu$ = 2) – CDF($x$ = 1.8, $\mu$ = 2), in which CDF is the cumulative distribution function of Eq. (8). If SCCNV calls (a) CN = 1 when $CN_{smoothed,b,c}$ is between 0.8 and 1.2, and (b) CN = 3 when $CN_{smoothed,b,c}$ is between 2.8 and 3.2, then%FP = CDF($x$ = 1,2, $\mu$ = 2) + 1 – CDF(2.8, $\mu$ = 2).

## Testing Datasets and Preprocessing of Data

Four SCWGS datasets were obtained for demonstrating and validating the performance of SCCNV (Zong et al., 2012; Lodato et al., 2015; Chen et al., 2017; Dong et al., 2017). The datasets included 8.2 TB SCWGS of 63 single human fibroblasts, neurons and cells of a tumor cell line amplified using eight different protocols, i.e., DOP-PCR (Sigma), Rubicon, MALBAC, LIANTI, and MDA (including Qiagen, GE, Lodato et al's MDA and



**FIGURE 1** | Major steps in SCCNV. A example of copy number estimates of **(A)** Raw sequencing depth; **(B)** after normalizing mappability and GC content; **(C)** after normalizing amplification bias; and **(D)** final results is presented. The example is a normal neuronal nucleus amplified with MDA (SRA id: SRR2141574). Each dot presents a 500 kb bin in the genome. Red and blue colors indicate bins of different chromosomes in Red and blue colors present bins of different chromosomes in their lexicographic order.

SCMDA). **Supplementary Table 1** lists all the single-cell data used in this study.

Sequence alignment was performed using BWA and GATK as follows (Li and Durbin, 2009; McKenna et al., 2010). Raw sequencing data of each sample (single cell and bulk DNA) were obtained from the SRA database and subjected to quality control using FastQC (version 0.11.4;[2]) and trimming using Trim Galore (version 0.4.1;[3]) with default parameters. Then they were aligned to the human reference genome (version hg19) using BWA MEM (version 0.7.12; option: -t number of CPUs -M reference genome fasta file) (Li and Durbin, 2009). PCR duplications were removed using picard tools (version 1.119;[4]). The alignments were subjected to indel realignment and basepair recalibration using GATK (version 3.5; using options, RealignerTargetCreator, IndelRealigner, BaseRecalibrator, and PrintReads) (McKenna et al., 2010). The step above was used for generating an analysis-ready bam file for other types of variants, e.g., single nucleotide variants, small insertions and deletions, and this step is optional for large CNVs or aneuploidies using SCCNV. Reads with mapQ < 30 were discarded. The number of reads per bin of each sample was calculated using samtools (version 1.3; option: bedcov) (Li et al., 2009). SCCNV (version 1.0) was used to estimate CNV of each cell.

## RESULTS

### Major Steps in SCCNV

As illustrated in **Figure 1**, SCCNV is composed of four major steps. It first calculates sequencing depth of the genome in bins of equal size (500-kb by default). Second, it normalizes the depth based on two features of the reference genome, including mappability and GC content. These two features are usually also considered by conventional CNV callers for bulk DNA sequencing. Next, it does further normalization across single cells of a same experimental batch. This step minimizes locus-specific bias due to WGA. Finally, it smooths the data (5 Mb by default) and infers copy number of each bin. Intermediate results between any two connecting steps can be generated by SCCNV for users to monitor its performance. We provide an example about the intermediate results of a normal neuronal nucleus amplified with MDA (SRA id: SRR2141574) in **Supplementary Figures 1–4**.

### Performance on Real Datasets

To evaluate the performance of SCCNV, we obtained four SCWGS datasets from the SRA database, which includes 8.2 TB high-depth WGS data of 63 single human fibroblasts, neurons and cells of a tumor cell line amplified using eight different protocols, i.e., DOP-PCR (Sigma), Rubicon, MALBAC, LIANTI, and MDA (including four MDA protocols, Qiagen, GE, Lodato et al's MDA and SCMDA) (**Supplementary Table 1**; Zong et al., 2012; Lodato et al., 2015; Chen et al., 2017; Dong et al., 2017). The data were processed as described in the Materials and Methods.

We used the CV of sequencing depth across all genomic bins on autosomes as an indicator of performance, because it directly determines sensitivity and FPR of copy number calling step in SCCNV. We show sensitivity and FPR of the copy number calling in **Figures 2A,B**, respectively. As the CV decreased from 0.135 to 0.041, the sensitivity increased from 0 to 100% and the FPR decreased from $3.8 \times 10^{-3}$ to $3.9 \times 10^{-11}$.

For the real datasets, we calculated the CV of raw data and normalized data after each step to demonstrate the performance of normalization in SCCNV (**Figure 3**). Almost all raw data (CV: $0.475 \pm 0.135$, avg. $\pm$ s.d.) are beyond the detection threshold, i.e., CV = 0.135. Each step of normalization decreased the CV by a significant fraction: on average, mappability normalization by 5%, GC content normalization by 33% percent, across-cell normalization by 22%, and smoothing by 55%. This shows the contributions of each normalization step to performance increase in the final variant calling. After all the normalization steps, the CVs are $0.107 \pm 0.076$ (avg. $\pm$ s.d.), corresponding to a



**FIGURE 2 |** Sensitivity **(A)** and False positive rate **(B)** of SCCNV. Y-axis value represents the coefficient of variation of the sequencing read depth (or normalized read depth) of all 500 kb bins across a single-cell genome (see section "Materials and Methods").

**FIGURE 3** | Decreased coefficient of variation by normalization steps in SCCNV. Each line presents the average CV of normalized sequencing depths of multiple single cells amplified using the same experimental procedures. See **Supplementary Table 1** for citations and SRA ids of each experimental batch.

sensitivity of 68.6% and an FPR of $3.6 \times 10^{-4}$ on average. Of note, different amplification protocols have significantly different performance when using SCCNV, likely due to differences in DNA amplification linearity among the protocols. As expected, LIANTI outperformed all the others (Chen et al., 2017). Protocols that included PCR steps, i.e., DOP-PCR, MALBAC and Rubicon, ranked in the middle. Although known as suffering from the least artifactual SNVs (Dong et al., 2017; Zhang et al., 2019), MDA-based protocols were ended last (**Figure 3**).

## DISCUSSION

Identification of copy number variation and aneuploidy has been one of the major areas of genomics methods development. Several statistical models have been developed for analyzing initially microarray data and later sequencing data of bulk DNA, for example, Circular Binary Segmentation (CBS), Mean Shift-Based (MSB) model, Shifting Level Model (SLM), Expectation Maximization (EM) model, and Hidden Markov Model (HMM) as discussed in Zhao et al. (2013). Based on these models, multiple computational software tools have been developed, e.g., CBS, Copynumber, CNVnator, and HMMcopy (Olshen et al., 2004; Abyzov et al., 2011; Ha et al., 2012; Nilsen et al., 2012). To call CNVs, most of the methods rely on assessing either sequencing read depth or alternate allele fraction at heterozygous SNPs across the genome of one sample, i.e., across-genome normalization. Some of the methods have been applied directly for analyzing single-cell sequencing data with specific filtering for cells with too much bias after WGA.

A few new tools for single cell data were also developed recently under the same rationale (assessing one sample at a time, or across-genome normalization), such as AneuFinder, baseqCNV, Ginkgo and SCOPE (Garvin et al., 2015; Bakker et al., 2016; Fu et al., 2019; Wang et al., 2019). SCCNV was developed based on our observation that the locus-specific amplification bias is often the same in different cells within one experimental batch and amplified using the same protocol (e.g., **Supplementary Figure 3**); and we showed that normalization across multiple samples (cells) significantly contributed to the increase in variant calling performance for single cells amplified using most WGA protocols (**Figure 3**). Following the same across-sample normalization rationale, another software tool, SCNV, was developed (Wang et al., 2018). It differs from SCCNV that SCCNV performs normalization based on empirical data directly (Eq. 4) without any assumption on its distribution. Of note, with across-sample normalization, SCCNV essentially aims to identify differences among different cells in one input batch and, therefore, it is important to input cells of interest (e.g., tumor cells) together with cells with a standard diploid genome.

## CONCLUSION

We developed SCCNV to identify copy number variations from whole-genome amplified single cells. We demonstrated its step-wise performance using most of the recent SCWGS datasets generated with 8 different amplification protocols.

## DATA AVAILABILITY STATEMENT

Raw sequencing data of each sample were obtained from SRA database (SRA SRP067062, SRA SRA060929, SRA SRP102259, SRA SRP041470, and SRA SRP061939).

## AUTHOR CONTRIBUTIONS

XD, LZ, and JV conceived the study. XD and TW developed the method. XD and XH analyzed the data. XD, LZ, and JV wrote the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

## ACKNOWLEDGMENTS

## CODE AVAILABILITY STATEMENT

The source code of SCCNV can be found in the https://github. com/biosinodx/SCCNV.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene. 2020.505441/full#supplementary-material

## REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Bakker, B., Taudt, A., Belderbos, M. E., Porubsky, D., Spierings, D. C., De Jong, T. V., et al. (2016). Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.* 17:115.

Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., et al. (2017). Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science* 356, 189–194. doi: 10.1126/science.aak9787

Dong, X., Zhang, L., Hao, X., Wang, T., and Vijg, J. (2019). SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing. *bioRxiv* [Preprint], doi: 10.1101/535807

Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A. Y., Wang, T., et al. (2017). Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* 14, 491–493. doi: 10.1038/nmeth.4227

Fu, Y., Zhang, F., Zhang, X., Yin, J., Du, M., Jiang, M., et al. (2019). High-throughput single-cell whole-genome amplification through centrifugal emulsification and eMDA. *Commun. Biol.* 2:147.

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G. S., Hicks, J., et al. (2015). Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* 12, 1058–1060. doi: 10.1038/nmeth.3578

Gundry, M., Li, W., Maqbool, S. B., and Vijg, J. (2012). Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic Acids Res.* 40, 2032–2040. doi: 10.1093/nar/gkr949

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., et al. (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* 22, 1995–2007. doi: 10.1101/gr.137570.112

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Lee, E., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350, 94–98. doi: 10.1126/science.aab1785

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., Mcindoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94. doi: 10.1038/nature09807

Nilsen, G., Liestol, K., Van Loo, P., Moen Vollan, H. K., Eide, M. B., Rueda, O. M., et al. (2012). Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genom.* 13:591. doi: 10.1186/1471-2164-13-591

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557–572. doi: 10.1093/biostatistics/kxh008

Vijg, J., and Dong, X. (2020). Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell* 182, 12–23. doi: 10.1016/j.cell.2020.06.024

Wang, R., Lin, D.-Y., and Jiang, Y. (2019). SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing. *bioRxiv* [Preprint], doi: 10.1101/594267

Wang, X., Chen, H., and Zhang, N. R. (2018). DNA copy number profiling using single-cell sequencing. *Brief Bioinform.* 19, 731–736. doi: 10.1093/bib/bbx004

Zhang, L., Dong, X., Lee, M., Maslov, A. Y., Wang, T., and Vijg, J. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9014–9019. doi: 10.1073/pnas.1902510116

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinform.* 14(Suppl. 11):S1. doi: 10.1186/1471-2105-14-S11-S1

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338, 1622–1626. doi: 10.1126/science.1229164