

SOFTWARE

Open Access



# cnvCurator: an interactive visualization and editing tool for somatic copy number variations

Lingnan Ma<sup>1,2,3</sup>, Maochun Qin<sup>1</sup>, Biao Liu<sup>1</sup>, Qiang Hu<sup>1</sup>, Lei Wei<sup>1</sup>, Jianmin Wang<sup>1\*</sup> and Song Liu<sup>1\*</sup>

## Abstract

**Background:** One of the most important somatic aberrations, copy number variations (CNVs) in tumor genomes is believed to have a high probability of harboring oncotargets. Detection of somatic CNVs is an essential part of cancer genome sequencing analysis, but the accuracy is usually limited due to various factors. A post-processing procedure including manual review and refinement of CNV segments is often needed in practice to achieve better accuracy.

**Results:** cnvCurator is a user-friendly tool with functions specifically designed to facilitate the process of interactively visualizing and editing somatic CNV calling results. Different from other general genomics viewers, the index and display of CNV calling results in cnvCurator is segment central. It incorporates multiple CNV-specific information for concurrent, interactive display, as well as a number of relevant features allowing user to examine and curate the CNV calls.

**Conclusions:** cnvCurator provides important and practical utilities to assist the manual review and edition of results from a chosen somatic CNV caller, such that curated CNV segments will be used for down-stream applications.

**Keywords:** Somatic copy number variation, Next generation sequencing, Cancer genome analysis, Interactive data viewer

## Background

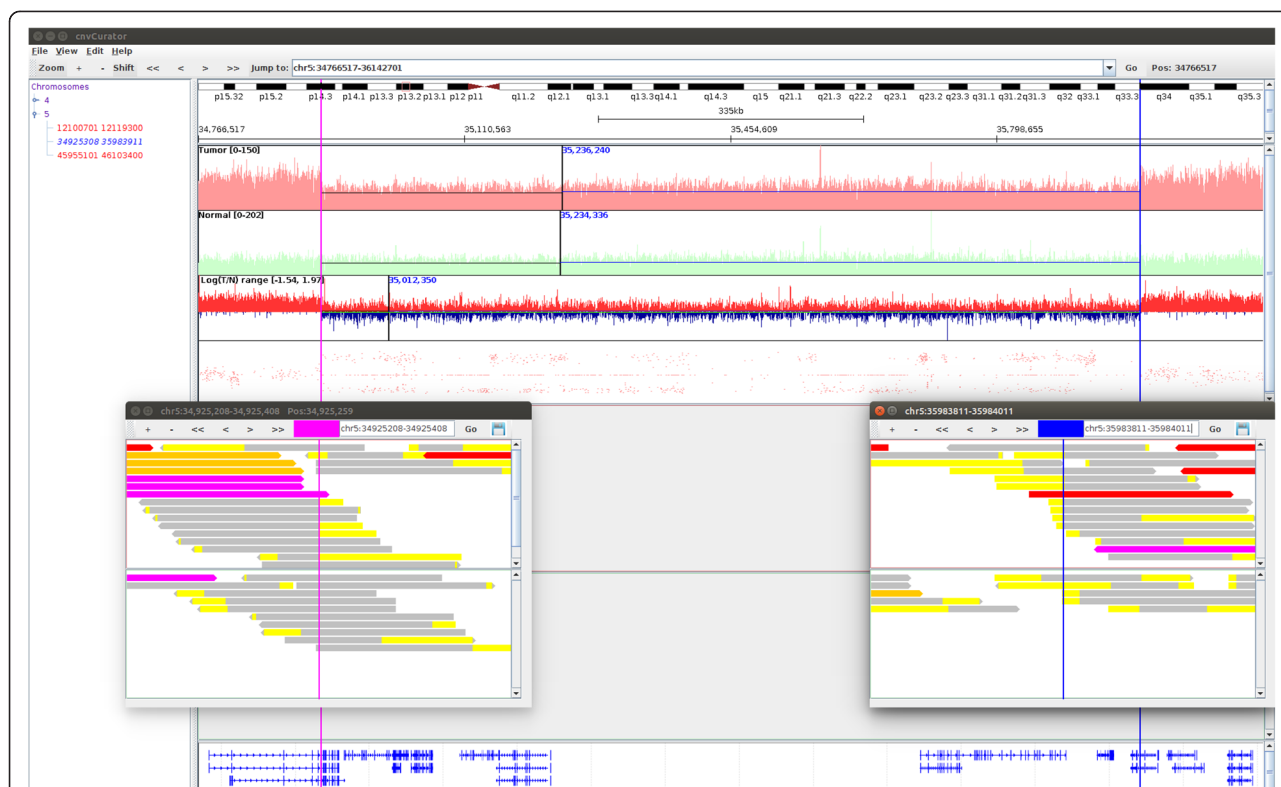
CNV was initially classified as gain or loss of a chromosome segment with a length greater than 1 kb, and then widened to include much smaller events (>50 bps) on accommodating the improved resolution of detection methods [1–3]. In cancer, CNV is one of the most important somatic aberrations, and has been identified as the driver event in many cancer types [1, 4, 5]. The widespread adoption of next-generation sequencing (NGS) provides an unprecedented opportunity to systematically screen for somatic CNVs in cancer genomes, and has been emerging as the primary means of interrogating the CNVs in recent investigations [6]. Accurate detection of CNVs from massive amounts of raw NGS data of the cancer genome requires sophisticated computational algorithms, with read depth, B Allele Frequency (BAF), split reads, and discordant read pairs derived from sequence read mapping as the primary input [7].

A number of computational methods have been developed to identify CNVs using NGS data [7, 8]. However, recent evaluations show each algorithm has its own strengths and weaknesses, and none of the CNV calling methods performed well in all situations [9]. The concordance of different CNV calling tools is especially low in real applications [10–12], suggesting that caution and care are needed to interpret and report the calling results, and additional post-processing might be needed to achieve maximum accuracy.

An ideal CNV report should accurately quantify the copy numbers in all genomic segments and delineate their breakpoints across the whole genome. In cancer study, CNV calling is particularly challenging due to tumor purity, heterogeneity, and aneuploidy [13–15]. Inaccurate segmentation could create false positive segments and miss true breakpoints, which will greatly affect downstream interpretation and application of CNV calls. Therefore, a post-processing procedure including manual review and curation of the CNV calling results is often required to reduce incorrect predictions in cancer genome analysis [16–18]. Manual review is a process to

\* Correspondence: Jianmin.Wang@RoswellPark.org; Song.Liu@RoswellPark.org  
<sup>1</sup>Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY 14263, USA

Full list of author information is available at the end of the article



**Fig. 1** An example of CNV call supported by multiple evidences. The interface of cnvCurator shows segment index on the left and various data tracks on the right. From top to bottom, the tracks include tumor read depth (red line), normal read depth (green line), logR ratio of tumor to normal (red for positive and blue for negative values, respectively), BAF (red dot), read alignment for tumor (in red box), read alignment for normal (in green box), and transcript annotation (blue). The shown region in the main window is base pairs 34766517–36142701 on chromosome 5. The two vertical shading lines (purple on the left, and blue on the right) in the main window delimit the boundary of the called CNV segment (Chr5: 34925308-35983911), with the breakpoint alignments displayed in the two separate pop-up windows. Within each pop-up window, the vertical black line delimitates the exact position of the breakpoint, and the red box and the green box display the tumor bam and the norm bam, separately. For the read displaying, split reads are in yellow–gray with the soft-clipped part in yellow and the matching portion in gray, and the others are discordant read pairs (purple - mate unmapped; red - mate in different chromosome; blue - mate in different strand; orange - mate in discordant distance)

identify false segmentation, missing breakpoints, and incorrect copy number status by visually checking relevant information underlying the CNV calling, such as read depth information of both tumor and normal samples, BAF of germline heterozygous SNVs (single nucleotide variations) in tumor, and signatures of reads and read-pairs (e.g., soft-clipped reads and discordant read pairs) related to breakpoints at a segment boundary. After an error is identified, an edition process needs to be performed to update the CNV boundaries as well as the related quantitative information.

Although several CNV detection methods provide certain types of graphical output for inspection, none of them provides an interactive viewer for result review and curation. While existing viewers such as Integrative Genomics Viewer [19] have sophisticated interfaces for interactive visualization, they are not tailored to CNV-specific applications. Furthermore, they generally do not provide any functions to facilitate manual curation of CNV calling results. Here, we present cnvCurator, a platform-independent

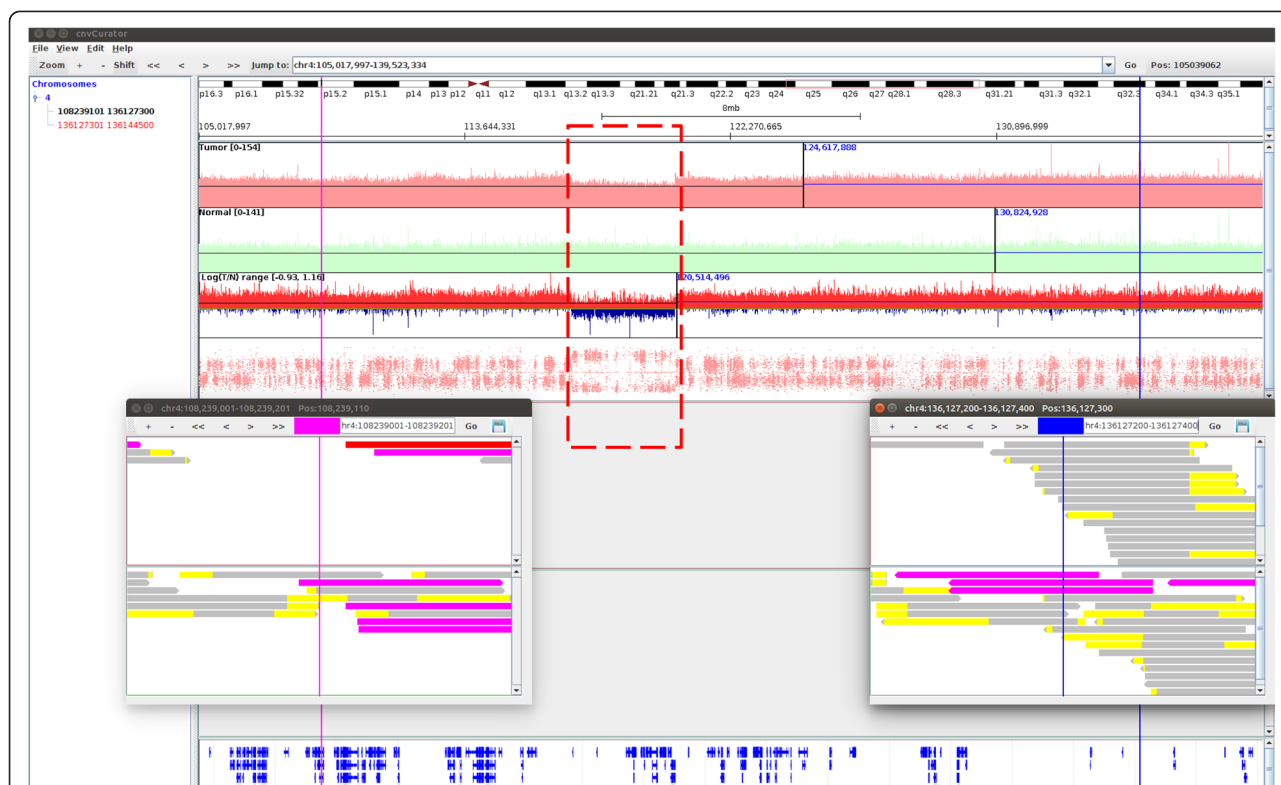
visualization and editing tool to facilitate manual inspection and curation of somatic CNV calling. Different from other general genomics viewers, cnvCurator incorporates multiple CNV-related information for concurrent, interactive display, as well as a number of practical features allowing user to examine and refine the CNV calls.

**Implementation**

cnvCurator is a Java utility designated to provide an interactive platform for conveniently reviewing and curating CNV predictions. There are two major components: 1) segment-central index and display of CNV calls for the purpose of manual review, and 2) editing the problematic CNV segments identified from manual review.

**Segment-central index and display of CNV calls for manual review**

The list of segments obtained from a given somatic CNV calling program will be indexed on the left of the main cnvCurator window. The segments can be generally



**Fig. 2** An example of missed CNV call. The shown region in the main window is base pairs 105017997-139523334 on chromosome 4. The two vertical lines (purple on the left, and blue on the right) in the main window delimit the boundary of a segment of copy number neutral (Chr4: 108239101-136127300) called by CONCERTING, with the breakpoint alignments displayed in the two separate pop-up windows. Upon manual review we found this segment contain a sub-segment of copy number loss (marked in rectangle with dashed blue line), as supported by three lines of evidences: 1) decreasing read depth in tumor over normal, as shown in the tracks of read depth and the track of logR ratio of tumor to normal; 2) the corresponding change of BAF pattern as shown in the BAF track; 3) the breakpoints of this sub-segment are in the exact position separating the soft-clipped part from the matched part of multiple split reads in the tumor bam, but not in the normal bam (shown in Fig. 3)

classified as copy gain, loss or neutral (i.e., diploid normal without copy number variations), depending on the program’s threshold for signal difference between tumor and matched normal. We list all segments on the left of the main window so that users can not only review a segment of copy number gain or loss, but also examine whether a segment of copy neutral call might be false negative, or contain sub-segment of copy number gain/loss. For easy navigation, we have added a feature to display the segment of copy loss as blue, copy gain as red, and copy neutral as black.

For a given segment of interest, cnvCurator supports concurrent visualization of diverse CNV-related data types on the right of the main window, which include 1) the ideogram of currently displayed chromosome, 2) the genomic coordinates, 3) the sequencing read depth information for tumor and matched normal samples, 4) the logR ratio derived from read depth data, 5) BAF at germline SNV positions [20], 6) details of sequencing read alignment for tumor and matched normal samples, and 7) annotations of genomic features. These tracks provide both direct and

indirect evidence from multiple angles to help the reviewer make decisions about the confidence of a given CNV call.

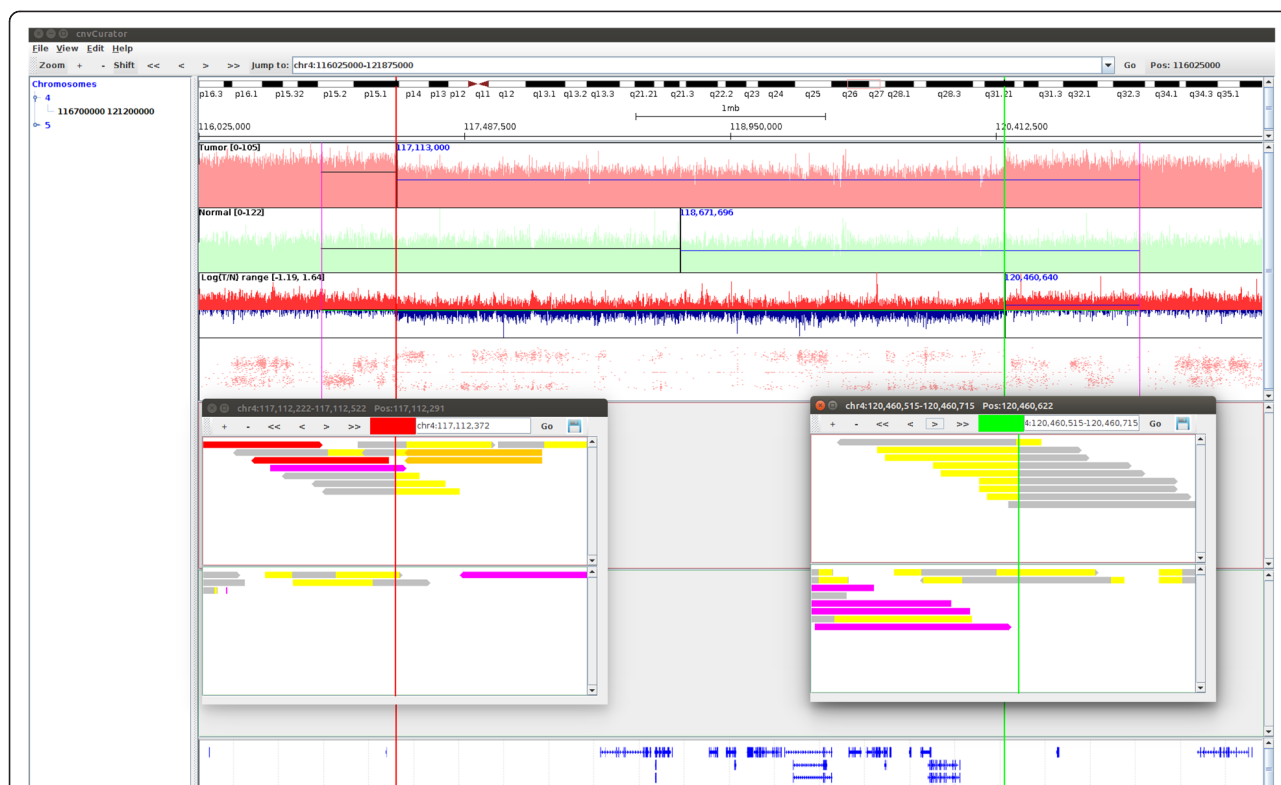
A number of navigating functions such as zoom in/out, moving around the genome, displaying detailed read alignment information, specifying alternative threshold for segment indexing, and changing of color scheme, are implemented in cnvCurator to assist CNV manual review and curation.

**Navigate between segments**

cnvCurator loads CNV segments into a panel docked at the left side of the application, and organizes the data into a tree structure with segments from the same chromosomes as leaves under the same node. By clicking any segment, the viewer switches to the corresponding genomic region with extra adjacent context displayed. The fraction of flanking regions can be specified by the user.

**Display breakpoint windows**

Real CNV boundaries are often associated with structural variations which can be detected by split reads or



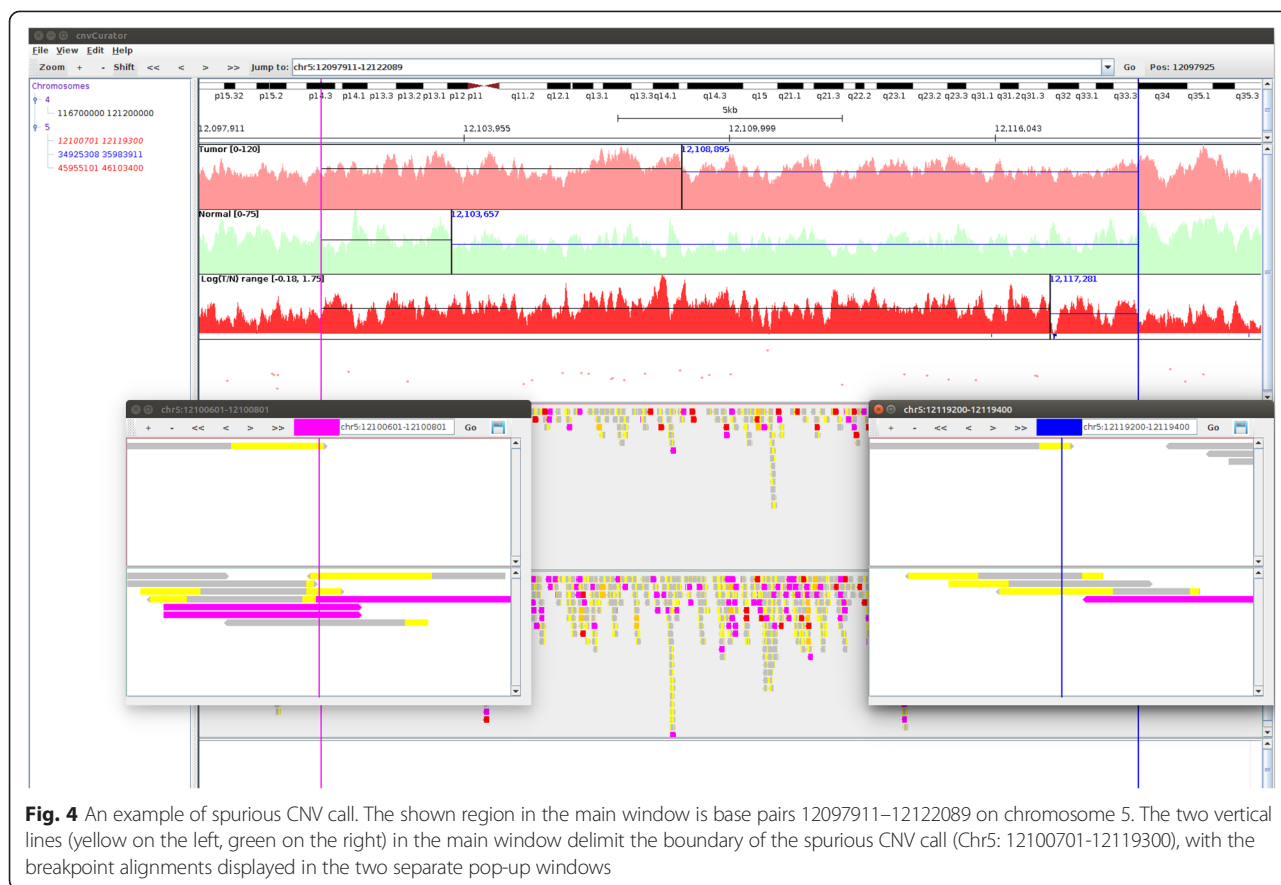
**Fig. 3** The breakpoint alignment of the missed CNV call. The shown region in the main window is base pairs 116025000-121875000 on chromosome 4. The two vertical lines (red on the left, and green on the right) in the main window delimit the boundary of the segment of copy number loss (Chr4: 117112372-120460622) missed by CONSERING as described in Fig. 2, with the breakpoint alignments displayed in the two separate pop-up windows

discordant read pairs [21–24]. Therefore, the presence of such reads can be used as supporting information to distinguish real CNVs from false calls caused by uneven sequencing read depth. In addition to the traditional alignment tracks which allow zoom-in/out visualization, cnvCurator provides advanced options to display such signatures in a user-friendly way. For example, once a CNV segment is selected, two pop-up windows for the two breakpoints ( $\pm 100$  bps by default) of the segment will be automatically displayed. Users can zoom-in/out the breakpoint pop-up windows, and simultaneously display multiple pop-up windows in adjacent regions to examine alternative breakpoints.

**Edit the problematic CNV segments identified from manual review**

An important and unique feature of cnvCurator is the ability to dynamically refine the results during the manual review process and generate a set of manually curated CNV calls. Mis-segmentation is a common problem for CNV detection methods, which will greatly affect downstream interpretation and application of CNV calls. It could be the missing of true breakpoints, which could incorrectly merge two distinct segments into one or miss the genuine

segments. It could also be introducing false breakpoints, which could incorrectly separate one segment into two parts, create segments with incorrect boundary, or create entirely false segments. Once the segmentation errors are spotted during the manual review, it would be handy to be able to fix them and have the segment list updated. cnvCurator provides several functionalities to assist the manual curation procedure, such as removing a false/spurious call, adding a missed/genuine segment, and correcting the breakpoint (s) of a segment with incorrect boundary. These can be achieved through merging two adjacent segments and/or splitting a segment. Merging adjacent segments is less complex, while splitting a segment requires knowing the exact location of the new breakpoint. Users can use cnvCurator to narrow down potential breakpoint location in the current window by identifying the position which minimizes the variations of logR ratio within the two segments. When there are multiple break points in the current window (segment), the user can recursively apply the splitting function within each sub-window (segment) to generate multiple new segments. We will implement the function of simultaneously splitting multiple breakpoints in the future version once we can find a solid statistical model for this problem.



The user can adjust the range of the current window and examine supporting features (e.g., reads alignment signature) to refine the searching. The segment texts on the left of the main window will be automatically updated once the segment edition (i.e., through splitting and/or merging the segments) is performed, and we have provided an “undo” function to these operations in the software. The updated segment list can be saved to a file and reloaded in the segmentation panel. The detailed tutorial to perform CNV editing during manual review process can be found in the project website. It should be noted that the process of merging two adjacent segments and/or splitting a segment with cnvCurator is not automated, and these editing functions are provided to assist in the manual review process.

**Results**

The inputs for cnvCurator include 1) the CNV calling results in segmented data file format [25] from a chosen somatic CNV caller 2) the read depth file in bigWig format for tumor and matched normal sample, separately, 3) the BAF file in bigWig format for tumor sample, and 4) the read alignment file in Binary Alignment/Map (BAM) format [26] for tumor and matched normal sample,

separately. The detailed format instruction for each required input file is provided in the project website.

We used CONSERGING [27] with default setting to make somatic CNV calls from a dataset of whole-genome sequencing of breast cancer specimens procured in the Roswell Park Cancer Institute (RPCI) (Wei et al., in preparation), and cnvCurator to review the calling results. All study participants provided consent for using their data and specimen for research purposes and the study was approved by Institutional Review Boards at RPCI. The cnvCurator outputs of a CNV call supported by multiple evidences, a missed CNV call, a spurious CNV call, and a CNV call with spurious breakpoint are shown as below. A mini BAM file containing all these exemplar regions is available in the cnvCurator package as demo.

**An example of CNV call supported by multiple evidences**

As shown in Fig. 1, this segment of copy number loss is supported by three lines of evidences: 1) decreasing read depth in tumor over normal, as shown in the tracks of read depth and the track of logR ratio of tumor to normal; 2) the corresponding change of B allele frequency (BAF) pattern as shown in the BAF track; 3) the breakpoints are in the exact position separating the soft-clipped part from

the matched part of multiple split reads in the tumor bam, but not in the normal bam.

**An example of missed CNV call**

The segment shown in Fig. 2 was called as copy number neutral by CONSERVING. However, upon manual review we found this segment contain a sub-segment of copy number loss, as supported by three lines of evidences: 1) decreasing read depth in tumor over normal, as shown in the tracks of read depth and the track of logR ratio of tumor to normal; 2) the corresponding change of BAF pattern as shown in the BAF track; 3) as shown in Fig. 3, the breakpoints of this sub-segment are in the exact position separating the soft-clipped part from the matched part of multiple split reads in the tumor bam, but not in the normal bam.

**An example of spurious CNV call**

As shown in Fig. 4, this CNV call is spurious as there is a lack of supporting evidences: 1) there is no clear change of read depth in tumor over normal, as shown in the tracks of read depth and the track of logR ratio of tumor to

normal; 2) there is no clear change of BAF pattern as shown in the BAF track; 3) the breakpoint positions are not supported by either split reads or discordant read pairs.

**An example of spurious breakpoint and alternative one**

As shown in Fig. 5, compared with the spurious breakpoint, the alternative breakpoint is supported by two lines of evidences: 1) the alternative breakpoint is more adjacent to the transition point of read depth changes in tumor over normal, as shown in the tracks of read depth and the track of logR ratio of tumor to normal; 2) the alternative breakpoint is in the exact position separating the soft-clipped part from the matched part of multiple split reads in the tumor bam, and is surrounded by multiple discordant read pairs.

**Discussion**

There are several notable differences between cnvCurator and existing genomics viewers, which are summarized as follows: First, as segmentation is a central part of somatic CNV calling methods, the index and display of CNV



**Fig. 5** An example of spurious breakpoint and alternative one. The shown region in the main window is base pairs 45948669–45966575 on chromosome 5. The vertical left line (45955101, in purple) in the main window delimits the spurious breakpoint call (Chr5: 45955101), and the vertical right line (45955840, in red) is the alternative breakpoint identified through manual review. The read alignments for the spurious breakpoint in the original call and the alternative breakpoint identified through manual review are displayed in the two separate pop-up windows

calling results in *cnvCurator* is segment based; Second, as boundary prediction is critical for calling CNV (especially focal CNV), *cnvCurator* provides automatic pop-up windows to examine the detailed breakpoint information; Third, as manual review is often required to curate the somatic variant calling from tumor genome sequencing data, we have implemented several editing functions in *cnvCurator* to assist in the process of CNV curations. Once the potential segmentation errors are spotted during the manual review, they can be modified through segment splitting/merging functions.

## Conclusions

Due to the special characteristics of tumor samples and the extraordinary complexity of tumor genomes, accurate detection of somatic CNVs is still a great challenge for the community. Here we describe a user-friendly visualization and editing tool specifically designed to facilitate manual review and curation of CNV segments generated by a chosen somatic CNV caller. The viewer provides important and practical utilities to identify problematic CNV segments through manual review, and refine the segmentation results so that user can generate curated results for further analysis.

## Availability and requirements

**Project name:** *cnvCurator*

**Project home page:** <http://www.acsu.buffalo.edu/~lm69/cnvCurator>

**Operating system(s):** Windows, Unix-like (Linux, Mac OSX)

**Programming language:** Java

**Other requirements:** Java (TM) 6.0 or higher

**License:** GNU LGPL

**Any restrictions to use by non-academics:** None

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

JW and SL conceived of the project. LM, LW, JW and SL designed the software. LM and MQ implemented the software. BL and QH tested the software. LM, BL, JW and SL drafted the manuscript. MQ and LW assisted in preparing the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

We would like to thank the reviewers and associate editor for their very helpful and constructive comments that allow us to substantially improve the manuscript. This work was supported by an award from the Roswell Park Alliance Foundation. The RPCI Bioinformatics Shared Resource is CCSG Shared Resources, supported by P30 CA016056.

## Author details

<sup>1</sup>Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY 14263, USA. <sup>2</sup>Department of Mathematics, University at Buffalo, Buffalo, NY 14260, USA. <sup>3</sup>College of Engineering, University of Michigan, Ann Arbor, MI 48109, USA.

Received: 19 March 2015 Accepted: 8 October 2015

Published online: 15 October 2015

## References

- Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet.* 2009;10(8):551–64.
- Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet.* 2006;7(2):85–97.
- Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet.* 2011;12(5):363–76.
- Shlien A, Malkin D. Copy number variations and cancer. *Genome Med.* 2009;1(6):62.
- Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. EPIGENETICS AND GENETICS A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer.* 2010;10(1):59–64.
- Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet.* 2010;11(10):685–96.
- Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget.* 2013;4(11):1868–81.
- Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics.* 2013;14 Suppl 11:S1.
- Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat.* 2014;35(7):899–907.
- Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotech.* 2011;29(6):512–20.
- Duan J, Zhang JG, Deng HW, Wang YP. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One.* 2013;8(3):e59128.
- Alkodsai A, Louhimo R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief Bioinform.* 2015;16(2):242–54.
- Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet.* 2012;13(3):189–203.
- Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer.* 2012;12(5):323–34.
- Shackleton M, Quintana E, Fearon ER, Morrison SJ. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell.* 2009;138(5):822–9.
- Andersson AK, Ma J, Wang J, Chen X, Gedman AL, Dang J, et al. The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat Genet.* 2015;47(4):330–7.
- Chen X, Bahrami A, Pappo A, Easton J, Dalton J, Hedlund E, et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Reports.* 2014;7(1):104–12.
- Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet.* 2013;45(3):242–52.
- Robinson JT, Thorvaldsdottir H, Winkler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24–6.
- Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, 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(11):1665–74.
- Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell.* 2013;153(4):919–29.
- Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods.* 2011;8(8):652–4.
- Liu B, Conroy JM, Morrison CD, Odunsi AO, Qin M, Wei L, et al. Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. *Oncotarget.* 2015;6(8):5477–89.
- Qin M, Liu B, Conroy JM, Morrison CD, Hu Q, Cheng Y, et al. SCNVSIM: somatic copy number variation and structure variation simulator. *BMC Bioinformatics.* 2015;16:66.

25. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557–72.
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
27. Chen X, Gupta P, Wang J, Nakitandwe J, Roberts K, Dalton JD, et al. CONSERTING: integrating copy-number analysis with structural-variation detection. *Nat Methods*. 2015;12(6):527–30.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

