PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Comprehensively benchmarking applications for detecting copy number variation

Le Zhang[1,2,3☯]*, **Wanyu Bai**[1☯], **Na Yuan**[4☯], **Zhenglin Du**[4]*

1 College of Computer Science, Sichuan University, Chengdu, China, **2** Medical Big Data Center, Sichuan University, Chengdu, China, **3** Zdmedical, Information polytron Technologies Inc. Chongqing, Chongqing, China, **4** BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, PR China

☯ These authors contributed equally to this work.
* zhangle06@scu.edu.cn (LZ); duzhl@big.ac.cn (ZD)

OPEN ACCESS

Citation: Zhang L, Bai W, Yuan N, Du Z (2019) Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol 15(5): e1007069. https://doi.org/10.1371/journal.pcbi.1007069

Editor: Ilya Ioshikhes, Ottawa University, CANADA

Received: January 23, 2019

Accepted: May 6, 2019

Published: May 28, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: LZ receives Chinese National Natural Science Foundation [61372138] (http://www.nsfc.gov.cn/) and National Science and Technology Major Project [2018ZX10201002] from Ministry of Science and Technology of the People´s Republic of China (http://www.most.gov.cn/eng/) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Motivation: Recently, copy number variation (CNV) has gained considerable interest as a type of genomic variation that plays an important role in complex phenotypes and disease susceptibility. Since a number of CNV detection methods have recently been developed, it is necessary to help investigators choose suitable methods for CNV detection depending on their objectives. For this reason, this study compared ten commonly used CNV detection applications, including CNVnator, ReadDepth, RDXplorer, LUMPY and Control-FREEC, benchmarking the applications by sensitivity, specificity and computational demands. Taking the DGV gold standard variants as a standard dataset, we evaluated the ten applications with real sequencing data at sequencing depths from 5X to 50X. Among the ten methods benchmarked, LUMPY performs the best for both high sensitivity and specificity at each sequencing depth. For the purpose of high specificity, Canvas is also a good choice. If high sensitivity is preferred, CNVnator and RDXplorer are better choices. Additionally, CNVnator and GROM-RD perform well for low-depth sequencing data. Our results provide a comprehensive performance evaluation for these selected CNV detection methods and facilitate future development and improvement in CNV prediction methods.

Author summary

As an important type of genomic structural variation, CNVs are associated with complex phenotypes because they change the number of copies of genes in cells, affecting coding sequences and playing an important role in the susceptibility or resistance to human diseases. To identify CNVs, several experimental methods have been developed, but their resolution is very low, and the detection of short CNVs presents a bottleneck. In recent years, the advancement of high-throughput sequencing techniques has made it possible to precisely detect CNVs, especially short ones. Many CNV detection applications were developed based on the availability of high-throughput sequencing data. Due to different CNV detection algorithms, the CNVs identified by different applications vary greatly. Therefore, it is necessary to help investigators choose suitable applications for CNV detection depending upon their objectives. For this reason, we not only compared ten commonly used CNV

Competing interests: No authors have competing interests. Le Zhang is a non-paid employee of Zdmedical, Information Polytron Technologies Inc.

detection applications but also benchmarked the applications by sensitivity, specificity and computational demands. Our results show that the sequencing depth can strongly affect CNV detection. Among the ten applications benchmarked, LUMPY performs best for both high sensitivity and specificity for each sequencing depth. We also give recommended applications for specific purposes, for example, CNVnator and RDXplorer for high sensitivity and CNVnator and GROM-RD for low-depth sequencing data.

This is a *PLOS Computational Biology* Benchmarking paper.

Introduction

Copy number variation (CNV) is a type of genomic structural variation that contains segmental duplications or deletions of a DNA fragment; the CNV size usually ranges from 1 kb to 3 Mb[1]. CNVs are found widely in individual human genomes, and they seldomly lead to genetic diseases[2]. CNVs can change the number of copies of a gene present in cells, thus affecting the coding sequences of genes, and they are associated with complex phenotypes [3]. CNVs also play an important role in the susceptibility or resistance to human diseases, such as cancer [4], Alzheimer disease [5], autism [6] and psoriasis [7].

Previously, researchers developed several experimental methods to explore CNVs, such as fluorescence in situ hybridization (FISH) and array comparative genomic hybridization (aCGH) [8], but the low resolution of these methods (approximately 5~10 Mbp for FISH and 10~25 kbp for aCGH) [9] presents a bottleneck for the detection of short CNVs [10]. In the last decade, Next Generation Sequencing (NGS) technology has enabled precise detection of CNVs, making it possible to identify small variants as short as 50 bp[11]. Many CNV detection algorithms were developed by NGS platforms.

The Read Depth (RD, or Read Count (RC))[12] and Pair-End Mapping (PEM, or Read Pair (RP))[13] algorithms are the most popular methods for CNV detection, and they use statistical models and clustering approaches for CNV detection[14], respectively. RD-based methods are good at detecting exact copy numbers, large insertions and CNVs in complex genomic region classes, whereas PEM-based methods can efficiently not only identify insertions and deletions but also locate mobile element insertions, inversions, and tandem duplications[14].

Many CNV detection methods have been developed based on the RD or PEM algorithms (Table 1). CNVnator is based on a statistical MSB model. It provides not only high sensitivity (86–96%) and genotyping accuracy (93–95%) but also a low false-discovery rate (3–20%)[15]. ReadDepth is based on a statistical CBS model, and it can interpret overdispersed data for better breakpoint estimation[16]. Control-FREEC is one of the most widely used RD-based CNV detection software programs, and it uses matched case-control samples or GC content to correct copy number[17]. CNVrd2 computes segmentation scores by integrating the linear regression algorithm[18] into a Bayesian normal mixed model; thus, it has the highest paralog ratio[19]. cn.MOPS decomposes variations in the depth of coverage across multiple samples into integer copy numbers and noise by means of its mixture components and Poisson distributions[20]. RDXplorer is based on the Event-Wise Testing (EWT) algorithm, which is a method based on significance testing, and the median size of detected CNVs is much longer than that using PEM methods[9]. Canvas is a favored tool for both somatic and germline CNV detection in large-scale sequencing studies, and it implements all steps of the variant calling

Table 1. CNV detection methods on WGS data.

Software	Methods	Algorithm detail	Input data	Publish	Latest update	Accessibility	URL	Programing Language	#Citations
#Canvas	RD	Expectation-maximization (EM) clustering	BAM	2011	2018/3	Y	https://github.com/Illumina/canvas	C#	29
#cn.MOPS	RD	Mixture Poisson model	BAM	2012	2018/10	Y	http://www.bioinf.jku.at/software/cnmops/cnmops.html	R	226
CNVeM	RD	Expectation-maximization (EM) algorithm	CSV	2013	NA	Y	https://omictools.com/cnvem-tool	C	14
CNVer	RP	Maximum-likelihood, Graphic flow	BAM	2010	2011/5	N	NA	C	158
#CNVnator	RD	Mean shift algorithm	BAM	2011	2016/11	Y	https://github.com/abyzovlab/CNVnator	C++	640
CNVrd2	RD	Expectation-maximization (EM) algorithm	BAM/SAM	2014	2015/11	Y	https://bioconductor.org/packages/release/bioc/html/CNVrd2.html	R	13
#Control-FREEC	RD	LASSO regression	BAM/SAM	2011	2018/8	Y	http://boevalab.com/FREEC/	C++	190
#GROM-RD	RD	Quantile normalization	BAM	2015	2017/5	Y	http://grigoriev.rutgers.edu/software/	C	7
#iCopyDAV	RD	DoC approaches	BAM	2018	2018/3	Y	https://github.com/vogetihrsh/icopydav	R,C++	1
JointSLM	RD	Population-based approach	SAM/BAM	2011	NA	N	NA	R	49
#LUMPY	RD, PEM	A probabilistic framework	BAM/CRAM	2014	2016/3	Y	https://github.com/arq5x/lumpy-sv	C++	157
mrCaNaVAR	RD	mrFAST	SAM	2009	2013/9	Y	http://mrcanavar.sourceforge.net/	C	685
#RDXplorer	RD	Event-wise testing algorithm	BAM	2009	2013/4	Y	https://sourceforge.net/projects/rdxplorer/	Python	496
#ReadDepth	RD	Circular binary segmentation algorithm	Bed Files	2011	2014/8	Y	https://github.com/chrisamiller/readDepth	R	150
#RSICNV	RD	Negative binomial transformations	BAM	2017	2017/7	Y	https://github.com/yhwu/rsicnv	C++	2

Note:

# indicates the software used in this study.

https://doi.org/10.1371/journal.pcbi.1007069.t001

workflow[21]. GROM-RD is a control-free CNV algorithm combining excessive coverage masking, GC bias mean and variance normalization[22]. iCopyDAV is a modular-framework based on DoC approaches[23]. RSICNV detects CNVs using the robust segment identification algorithm with negative binomial transformations[24]. LUMPY integrates the CNV detection methods of RD and PEM and allows for more sensitive CNV discovery[25].

Previous studies have surveyed CNV detection software with regards to specificity, sensitivity and computational demands, and they have evaluated their advantages and shortcomings. For example, Fatima et al. evaluate CNV detection software based on analysis of whole-exome sequencing (WES) data[26], and Junbo et al. evaluate six RD-based CNV detection software programs based on analysis of whole genome sequencing (WGS) data[27]. However, previous studies neither consider the impact of varied sequencing depth on the software performance nor use a standardized CNV dataset for evaluation based on analysis of real sequencing data. Our study not only adds several newer, untested software programs such as RSICNV, iCopyDAV and GROM-RD but also uses Database of Genomic Variants (DGV) as the gold standard

so that our test results are more extensive and reliable[28]. Here, we surveyed ten frequently used methods of CNV detection in WGS data (Table 1), including CNVnator, ReadDepth, RDXplorer, LUMPY and Control-FREEC, and evaluated not only the detected CNV number, length distribution and result coincidence between different CNV methods but also the accuracy, sensitivity and computational demand under the conditions of different sequencing depths. Our study also compares the advantages and shortcomings of such CNV detection methods, providing useful information for researchers to select a suitable method.

Materials and methods

Study data

The sequencing data (94x) of the individual NA12878 were downloaded from the website of the 1000 Genomes Project[29] as evaluation data to compare the performance of CNV detection methods using real sequencing data. The DGV Gold Standard Variants for NA12878 were download from the Database of Genomic Variants (DGV)[28], and a previously published SV benchmark of NA12878[30] was also fetched from the FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA12878/)[31].

Identification of CNVs in NA12878

After removing sequencing adapters and trimming consecutive low-quality bases from both the 5' and 3' end of the reads using an in-house Perl script, clean reads were subsampled by the sequencing depth of 5x, 20x, 10x, 30x, 40x and 50x using seqtk (https://github.com/lh3/seqtk) [32]. Then, the six datasets were mapped to the human reference genome (hg19) using BWA (V0.7.12) (http://bio-bwa.sourceforge.net/) [33] with default parameters. The Picard program (https://broadinstitute.github.io/picard/) [34] was used to sort mapping results to the BAM format. For CNV identification of NA12878, ten methods were used with default or recommended parameters, including CNVnator, ReadDepth, RDXplorer, LUMPY and Control-FREEC. The CNVs with lengths of more than 1 kb were kept as detected CNVs. The main parameters for each software program used are listed in S1 Table.

Performance evaluation criteria

In the two datasets of the DGV Gold Standard Variants and the SV benchmark, the CNVs longer than 1 kb were merged by location overlap of more than 50% and were taken as the standard CNV dataset for performance evaluation (S1 Table). The identified CNVs of each method were regarded as true positive results if there was more than 50% overlap on chromosome locations compared with the standard CNV dataset; otherwise, they were regarded as true negative CNVs. Then, the true positive rates (TPRs) and the false discovery rates (FDRs) were calculated and compared. The formulas to calculate TPR and FDR are shown in Table 2. For computing time estimation, each application was run five times, and the average running times were recorded for the related standard deviation computation. To compare the memory usage of the applications, each application was run five times, and the average memory sizes

Table 2. Formula to calculate TPR and FDR.

Measure	Formula	Illustration
TPR	$TPR = \frac{TP}{P}$	TP: the number of true positivis P:the number of positives in DGV
FDR	$FDR = \frac{FP}{TP+FP}$	TP: the number of true positives FP: the number of false positives

https://doi.org/10.1371/journal.pcbi.1007069.t002

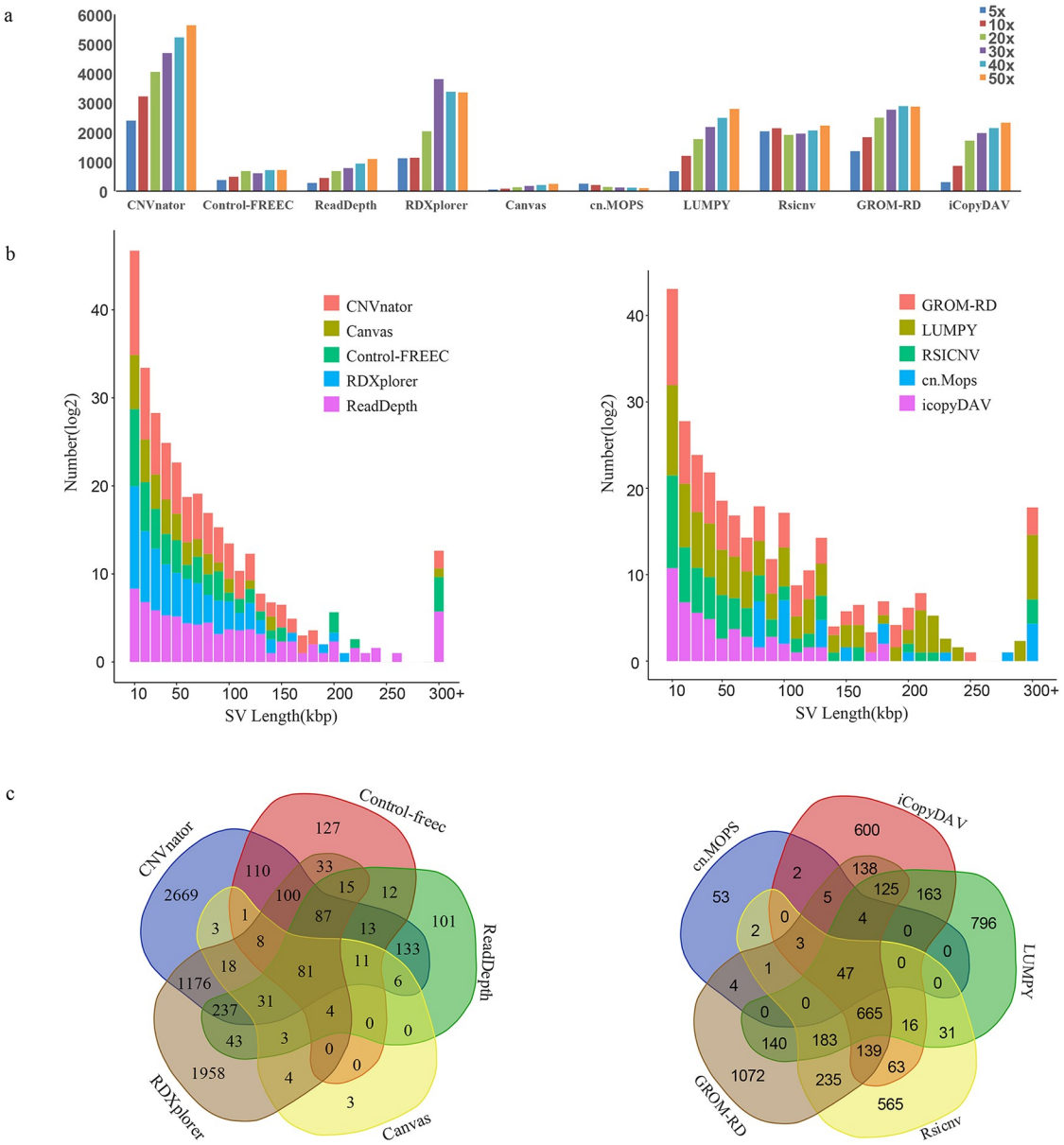

Fig 1. Statistics of the detected CNVs. (a) Detected CNV number. (b) Distribution of CNV size. (c) The Venn diagram of CNV detection methods.

https://doi.org/10.1371/journal.pcbi.1007069.g001

were recorded for the related standard deviation[35–38] computation. The process used for performance evaluation is shown in S1 Fig.

Results

Comparison of identified CNVs

With sequencing data with depths from 5X to 50X, ten methods were used to identify CNVs in NA12878 (shown in Table 1), and the tested CNVs were listed in the supplementary files(S1–S11 Files). As shown in Fig 1a, due to different CNV detection algorithms, the numbers of detected CNVs varied greatly. CNVnator and RDXplorer identified the most CNVs, whereas Canvas and cn.MOPS identified the fewest. In most cases, the number of CNVs identified

were positively correlated with the sequencing depth. However, RDXplorer detected the most SVs at 30X depth, probably because the method was tested and optimized at a 30X sequencing depth[9].

On the other hand, each software program tended to detect CNVs of different sizes, ranging from less than 1 kb to several hundred kbp. As shown in Fig 1b, most methods identified many small CNVs shorter than 10 kb, whereas LUMPY and ReadDepth predicted more CNVs longer than 200 kb.

The detected CNVs for each method at a 30X sequencing depth were also compared in Fig 1c. Generally, CNVs identified by more than one method are more specific than those called by only one method[39]. As shown in Fig 1c, 98.27% of CNVs identified by Canvas were also identified by four other methods; the program with the next highest level of consistency with other methods was ReadDepth (87.00%), whereas CNVnator and RDXplorer identified the most CNVs that were only called in a single method.

Sensitivity and specificity of CNV prediction

As shown in Fig 2a, the TPR curves of the ten methods were plotted at six sequencing depths from 5X to 50X. At a low sequencing depth of 5X, the TPR of LUMPY reached 0.432, followed

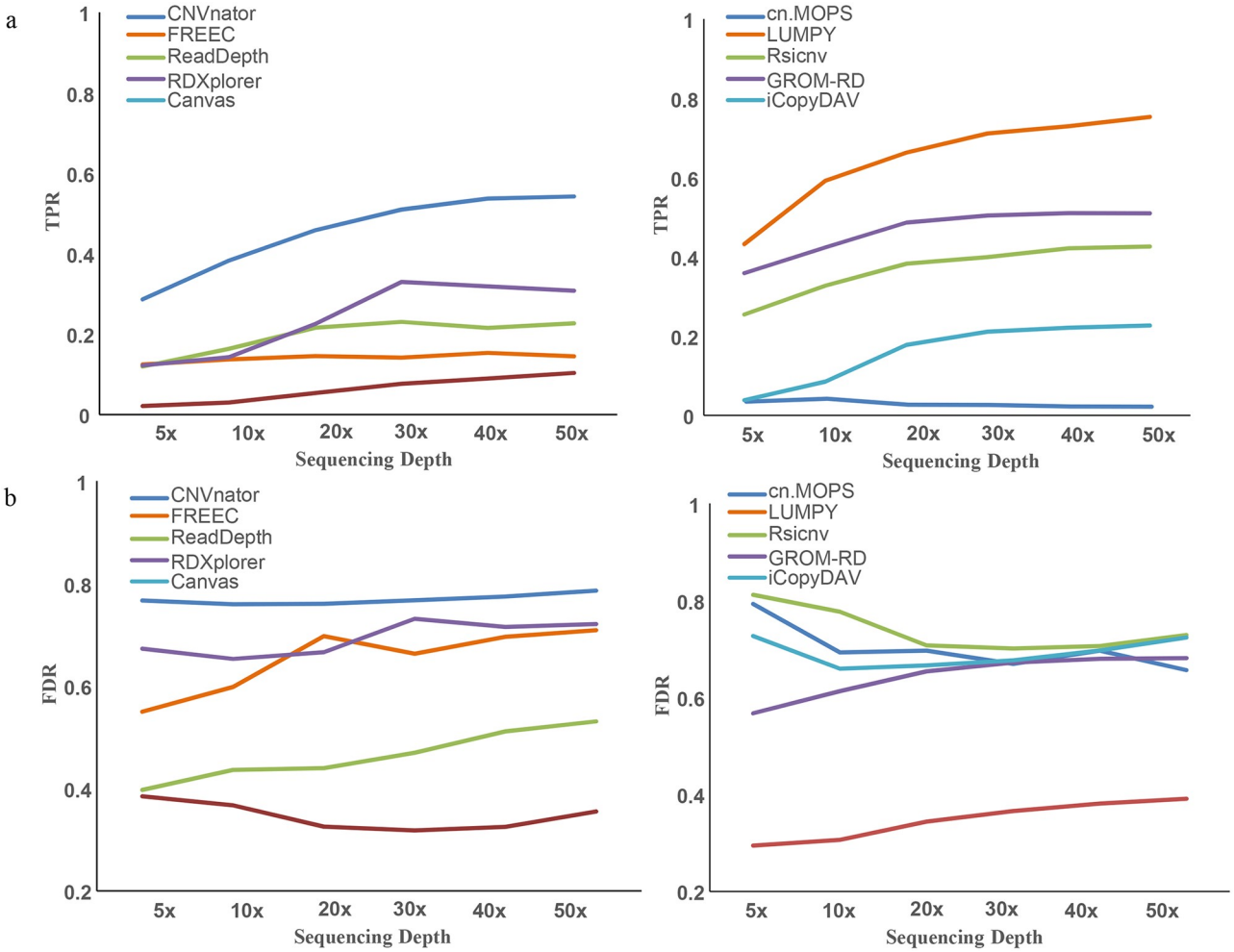

Fig 2. The evaluation of sensitivity and specificity of CNV detection methods. (a) TPR curves of the ten applications at sequencing depths from 5X to 50X. (b) FDR curves of the ten applications at sequencing depths from 5X to 50X.

https://doi.org/10.1371/journal.pcbi.1007069.g002

by CNVnator (0.370) and GROM-RD (0.359), which was much greater than other methods (0.021 to 0.254), implying that these three methods have greater sensitivity at low sequencing depth. At high sequencing depths of 30X and 50X, CNVnator also showed the highest TPR of 0.725 and 0.800, followed by LUMPY (0.711, 0.753) and RDXplorer (0.678, 0.621), implying higher sensitivity than other methods. Overall, at each sequencing depth from low to high, CNVnator and LUMPY had the best performance with respect to the sensitivity of CNV detection.

At increasing sequencing depths, the trends of the TPR curves were different from one another. For CNVnator, LUMPY and ReadDepth, the range with varying TPR was much wider (Fig 2a), and the TPR curve visibly increased, which indicates that the sensitivity of CNV detection is positively correlated with the sequencing depth. The TPR curve of RDXplorer also significantly increased with sequencing depth from 5X to 30X but reached a plateau at a 30X depth. This may result from the algorithm design as mentioned above.

Considering the sensitivity of detecting CNVs and sequencing costs, a sequencing depth of 30X provides the best value for CNV detection, as is indicated by the trends in the TPR curves (Fig 2a). However, the TPR curves were independent from sequencing depth for FREEC, cn.MOPS and Canvas (Fig 2a). With regards to the specificity of CNV detection methods, the FDR curves of Canvas and LUMPY were lower than the others, implying that the specificities of these two methods are better than those of the other methods, i.e., they predicted the least false positive results (Fig 2b). The FDR value of iCopyDAV reached a peak value at a 30X depth (0.878), followed by CNVnator (0.767) and RDXplorer (0.731), but these three methods also predicted the most CNVs (Fig 2b).

Computational demands. The computational demands of these methods with respect to computing time and memory usage are shown in Fig 3. Computing times of these ten

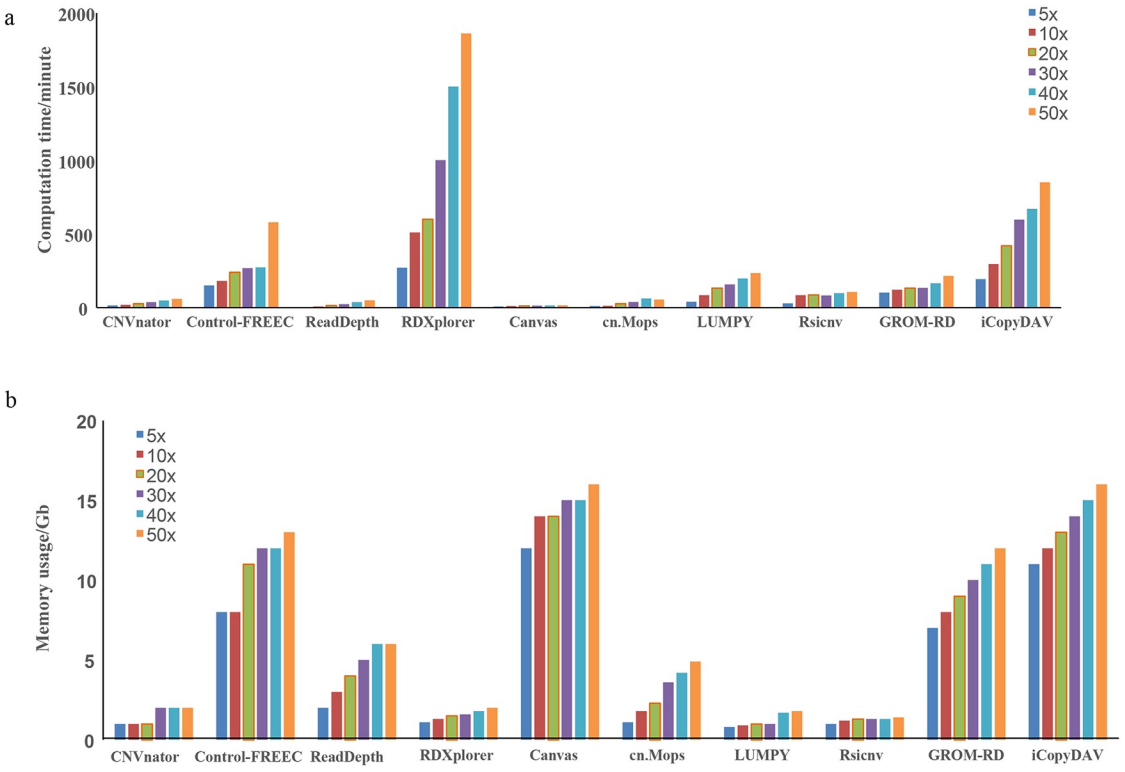

Fig 3. The computational demands of the ten methods. (a) Computation time as a function of sequencing depth from 5X to 50X. (b) Memory usage as a function of sequencing depth from 5X to 50X.

https://doi.org/10.1371/journal.pcbi.1007069.g003

applications increased with the increment of sequencing depth (Fig 3a). In particular, RDXplorer had the highest cost, followed by iCopyDAV and FREEC, with the other methods being comparable with low runtime costs. As shown in Fig 3b, the memory usage rates of these ten methods were positively related to sequencing depth. CNVnator, RDXplorer, RSICNV and LUMPY used the least amount of memory, while iCopyDAV, Canvas and FREEC needed more memory to run.

Discussion

This study surveyed the performance of ten CNV detection applications with regards to sensitivity, specificity and computational demands over a range of sequencing depths.

We found that most CNVs detected by Canvas and ReadDepth could be explored by other methods, but CNVnator and RDXplorer identified many specific CNVs (Fig 1c). Of all the CNV detection methods, LUMPY showed the best performance in terms of both sensitivity and specificity, probably because LUMPY integrates two different algorithms of PEM and RD for CNV prediction[25], and the PEM algorithm can provide better mapping accuracy on highly repetitive genomic regions than RD-based methods in some cases.

Since TPR values for most methods were below 0.8 and the FDR values for most methods were above 0.3 (Fig 2), we believe that the sensitivity and specificity for CNV detection are not likely to be improved in the future.

Limiting the CNV detection algorithms studied, our results are consistent with a previous report[39]. For all the ten methods, including RD-based algorithms, the read depth distribution is affected by the following three major causes. First, the GC-content in genomes leads to PCR bias during the construction of sequencing libraries, and the genome regions with ultrahigh or ultralow GC-contents are difficult to sequence, so the read depths on these regions are uneven. Second, because the genome sequencing was performed using short reads and it is difficult to correctly map short reads to genome regions with highly repetitive sequences, false positive CNV results arise in most studies. Lastly, the valuation results for cn.MOPS fall short of expectations. Since the cn.MOPS method was designed for input data from multisamples, the sensitivity and specificity are both very low when inputting single samples.

The high FDR of CNV detection was also likely caused by the imperfectness of the standard CNV dataset. We also conducted the evaluation with another set of gold standard CNVs used in a previous study[40], but the evaluation results were similar. A possible explanation is that it is difficult to identify all the CNVs on real experimental data, in spite of the fact that many platforms were used to confirm the detected CNVs on DGV Gold Variations. Therefore, the standard CNV dataset may not comprise all the true CNVs in NA12878, and it may include some incorrect CNVs. For example, of all the CNVs in the standard CNV data set, 623 CNVs were not detected by any of the ten methods; these are most likely false positive detection results.

The benchmarking above was based on single subsampling on each sequencing depth. To avoid subsampling bias, we evaluated the effect of subsampling on CNV prediction using multiple random subsampling. As shown in S2 Fig, we calculated TPR and FDR using five times subsampling for each CNV program on 30X depth (S2a & S2b Fig), which is a typical depth for whole genome resequencing studies, and also subsampled five times on each depth for one program LUMPY (S2c & S2d Fig). Most CNV prediction results of multiple subsampling are steady and the trends of TPR and FDR curves of each program were consistent with previous benchmarking conclusions (Fig 2a & 2b).

The aim of this survey is to help researchers choose appropriate CNV detection methods according to their specific purposes and the features of their data. We suggest that (1) when low FDR is preferable, LUMPY and Canvas are better choices (Fig 2); (2) when high sensitivity

is preferable, LUMPY, CNVnator and RDXplorer are better choices (Fig 2); and (3) if the speed/computation demand is the first priority, CNVnator and ReadDepth should be considered (Fig 3).

In this study, we employed the default or recommended parameters of each application for performance comparison. We plan to compare the best performance for each application by fine tuning the parameters and to include more recently published CNV applications in the future. Considering the limitations of sequencing data comprised of short reads, we are also preparing to evaluate CNV detection methods using long sequencing reads, such as PacBio or Oxford Nanopore, which may further improve the CNV prediction performance with regards to sensitivity and specificity.

Supporting information

S1 Fig. The evaluation workflow.
(TIF)

S2 Fig. Evaluation of sensitivity and specificity of CNV detection methods using five times subsampling. (a) TPR of the ten application at 30X depth using five times subsampling. (b) FDR of the ten application at 30X depth using five times subsampling. (c) TPR of Lumpy from 5X to 50X depth using five times subsampling at each depth. (d) FDR of Lumpy from 5X to 50X depth using five times subsampling at each depth.
(TIF)

S1 Table. The detailed information concerning the tested software.
(DOCX)

S1 File. Standard CNVs for NA12878.
(XLSX)

S2 File. Detected CNVs using Canvas.
(XLSX)

S3 File. Detected CNVs using cn.MOPS.
(XLSX)

S4 File. Detected CNVs using CNVnator.
(XLSX)

S5 File. Detected CNVs using iCopyDAV.
(XLSX)

S6 File. Detected CNVs using GROM-RD.
(XLSX)

S7 File. Detected CNVs using Rsicnv.
(XLSX)

S8 File. Detected CNVs using Control-FREEC.
(XLSX)

S9 File. Detected CNVs using RDXplorer.
(XLSX)

S10 File. Detected CNVs using ReadDepth.
(XLSX)

S11 File. Detected CNVs using LUMPY.
(XLSX)

Author Contributions

Data curation: Wanyu Bai, Na Yuan, Zhenglin Du.

Software: Na Yuan.

Writing – original draft: Le Zhang, Wanyu Bai, Zhenglin Du.

Writing – review & editing: Wanyu Bai, Zhenglin Du.

References

1. Carson AR, Feuk L, Mohammed M, Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. Human Genomics 2: 403–414. https://doi.org/10.1186/1479-7364-2-6-403 PMID: 16848978
2. Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy Number Variation in Human Health, Disease, and Evolution. Annurevgenomics Humgenet 10: 451–481.
3. Handsaker RE, Vanessa VD, Berman JR, Giulio G, Seva K, et al. (2015) Large multiallelic copy number variations in humans. Nature Genetics 47: 296–303. https://doi.org/10.1038/ng.3200 PMID: 25621458
4. Pagter MSD, Kloosterman WP (2015) The Diverse Effects of Complex Chromosome Rearrangements and Chromothripsis in Cancer Development. Recent Results Cancer Res 200: 165–193. https://doi.org/10.1007/978-3-319-20291-4_8 PMID: 26376877
5. Anne RL, Didier H, Gregory R, Nathalie LM, Annie L, et al. (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nature Genetics 38: 24–26. https://doi.org/10.1038/ng1718 PMID: 16369530
6. Jonathan S, Lakshmi B,., Dheeraj M, Jennifer T, Christa LM, et al. (2007) Strong association of de novo copy number mutations with autism. Science.
7. Rafael DC, Eva RM, Zeeuwen PLJM, Jason R, Wilson L, et al. (2009) Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nature Genetics 41: 211–215. https://doi.org/10.1038/ng.313 PMID: 19169253
8. Buysse K, Chiaie BD, Coster RV, Loeys B, Paepe AD, et al. (2009) Challenges for CNV interpretation in clinical molecular karyotyping: Lessons learned from a 1001 sample experience. European Journal of Medical Genetics 52: 398–403. https://doi.org/10.1016/j.ejmg.2009.09.002 PMID: 19765681
9. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research 19: 1586–1592. https://doi.org/10.1101/gr.092981.109 PMID: 19657104
10. Junbo D, Ji-Gang Z, Hong-Wen D, Yu-Ping W (2013) Comparative studies of copy number variation detection methods for next-generation sequencing technologies. Plos One 8: e59128. https://doi.org/10.1371/journal.pone.0059128 PMID: 23527109
11. Altshuler DL, Durbin R, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. https://doi.org/10.1038/nature09534 PMID: 20981092
12. Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nature Genetics 39: S16. https://doi.org/10.1038/ng2028 PMID: 17597776
13. Korbel JO, Alexander Eckehart U, Affourtit JP, Brian G, Fabian G, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420–426. https://doi.org/10.1126/science.1149504 PMID: 17901297
14. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. Bmc Bioinformatics 14: S1.
15. Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21: 974. https://doi.org/10.1101/gr.114876.110 PMID: 21324876
16. Miller CA, Oliver H, Cristian C, Aleksandar M (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. Plos One 6: e16327. https://doi.org/10.1371/journal.pone.0016327 PMID: 21305028

17. Valentina B, Tatiana P, Kevin B, Pierre C, Julie C, et al. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28: 423–425. https://doi.org/10.1093/bioinformatics/btr670 PMID: 22155870

18. Zhang L, Zheng CQ, Li T, Xing L, Zeng H, et al. (2017) Building Up a Robust Risk Mathematical Platform to Predict Colorectal Cancer. Complexity 2017: 14.

19. Nguyen HT, Merriman TR, Black MA (2014) The CNVrd2 package: measurement of copy number at complex loci using high-throughput sequencing data. Frontiers in Genetics 5: 248. https://doi.org/10.3389/fgene.2014.00248 PMID: 25136349

20. Klambauer G, Schwarzbauer K, Mayr A, Clevert D, Mitterecker A, et al. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Research 40.

21. Roller E, Ivakhno S, Lee S, Royce T, Tanner S (2016) Canvas: versatile and scalable detection of copy number variants. Bioinformatics 32: 2375–2377. https://doi.org/10.1093/bioinformatics/btw163 PMID: 27153601

22. Smith SD, Kawash JK, Grigoriev A (2015) GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. Peerj 3: e836. https://doi.org/10.7717/peerj.836 PMID: 25802807

23. Dharanipragada P, Vogeti S, Parekh N (2018) iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization. Plos One 13: e0195334. https://doi.org/10.1371/journal.pone.0195334 PMID: 29621297

24. Saran V, Jessie XJ, Yinghua W, Hongzhe L (2014) Parametric modeling of whole-genome sequencing data for CNV identification. Biostatistics 15: 427–441. https://doi.org/10.1093/biostatistics/kxt060 PMID: 24478395

25. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biology 15: R84. https://doi.org/10.1186/gb-2014-15-6-r84 PMID: 24970577

26. Zakani A,., Saghari M,., Eftekhari M,., Fard-Esfahani A,., Fallahi B,., et al. (2011) Evaluation of radioiodine therapy in differentiated thyroid cancer subjects with elevated serum thyroglobulin and negative whole body scan using 131I with emphasize on the thallium scintigraphy in these subgroups. European Review for Medical & Pharmacological Sciences 15: 1215.

27. Guo Y, Sheng Q, Samuels DC, Lehmann B, Bauer JA, et al. (2013) Comparative Study of Exome Copy Number Variation Estimation Tools Using Array Comparative Genomic Hybridization as Control. BioMed Research International,2013,(2013-11-4) 2013: 417–422.

28. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42: D986–992. https://doi.org/10.1093/nar/gkt958 PMID: 24174537

29. Adam A, Lisa D B, Richard M D, Erik P G, Hyun Min K, et al. (2015) A global reference for human genetic variation. Nature 526: 68–74. https://doi.org/10.1038/nature15393 PMID: 26432245

30. Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, et al. (2016) svclassify: a method to establish benchmark structural variant calls. Bmc Genomics 17: 64. https://doi.org/10.1186/s12864-016-2366-2 PMID: 26772178

31. IGSR (2008) 1000 Genome Project Data.

32. Github (2016) Seqtk.

33. sourceforge (2010) Burrows-Wheeler Aligner.

34. Github (2019) Picard.

35. Zhang L, Liu Y, Wang M, Wu Z, Li N, et al. (2017) EZH2-, CHD4-, and IDH-linked epigenetic perturbation and its association with survival in glioma patients. J Mol Cell Biol 9: 477–488. https://doi.org/10.1093/jmcb/mjx056 PMID: 29272522

36. Zhang L, Qiao M, Gao H, Hu B, Tan H, et al. (2016) Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation. Nanoscale 8: 14877–14887. https://doi.org/10.1039/c6nr01637e PMID: 27460959

37. Zhang L, Xiao M, Zhou J, Yu J (2018) Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA). Bioinformatics 34: 3624–3630. https://doi.org/10.1093/bioinformatics/bty392 PMID: 29762634

38. Zhang L, Zhang S (2017) Using game theory to investigate the epigenetic control mechanisms of embryo development: Comment on: "Epigenetic game theory: How to compute the epigenetic control of maternal-to-zygotic transition" by Qian Wang et al. Phys Life Rev 20: 140–142. https://doi.org/10.1016/j.plrev.2017.01.007 PMID: 28109753

39. Renjie T, Yadong W, Kleinstein SE, Yongzhuang L, Xiaolin Z, et al. (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. Human Mutation 35: 899–907. https://doi.org/10.1002/humu.22537 PMID: 24599517

40. Haraksingh RR, Abyzov A, Urban AE (2017) Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. Bmc Genomics 18: 321. https://doi.org/10.1186/s12864-017-3658-x PMID: 28438122