A comprehensive benchmarking of WGS-based structural variant callers

Varuni Sarwal[1,2], Sebastian Niehus[3,4], Ram Ayyala[1], Sei Chang[1], Angela Lu[1], Nicholas Darci-Maher[1], Russell Littman[1], Emily Wesel[1], Jacqueline Castellanos[1], Rahul Chikka[1], Margaret G. Distler[5], Eleazar Eskin[1,6,7], Jonathan Flint[5†], Serghei Mangul[8,9*†]

[1] Department of Computer Science, University of California Los Angeles, 580 Portola Plaza, Los Angeles, CA 90095, USA

[2] Indian Institute of Technology Delhi, Hauz Khas, New Delhi, Delhi 110016, India

[3] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany

[4] Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany

[5] Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, 760 Westwood Plaza, Los Angeles, CA 90095, USA

[6] Department of Human Genetics, David Geffen School of Medicine at UCLA, 695 Charles E. Young Drive South, Box 708822, Los Angeles, CA, 90095, USA

[7] Department of Computational Medicine, David Geffen School of Medicine at UCLA, 73-235 CHS, Los Angeles, CA, 90095, USA

[8] Department of Clinical Pharmacy, School of Pharmacy, University of Southern California 1985 Zonal Avenue Los Angeles, CA 90089-9121

[9] Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, 90089, USA

* Corresponding author: serghei.mangul@gmail.com.

† These authors contributed equally to the work.

Abstract

Advances in whole genome sequencing promise to enable the accurate and comprehensive structural variant (SV) discovery. Dissecting SVs from whole genome sequencing (WGS) data presents a substantial number of challenges and a plethora of SV-detection methods have been developed. Currently, there is a paucity of evidence which investigators can use to select appropriate SV-detection tools. In this paper, we evaluated the performance of SV-detection tools using a comprehensive PCR-confirmed gold standard set of SVs. In contrast to the previous benchmarking studies, our gold standard dataset included a complete set of SVs allowing us to report both precision and sensitivity rates of SV-detection methods. Our study investigates the ability of the methods to detect deletions, thus providing an optimistic estimate of SV detection performance, as the SV-detection methods that fail to detect deletions are likely to miss more complex SVs. We found that SV-detection tools varied widely in their performance, with several methods providing a good balance between sensitivity and precision. Additionally, we have determined the SV callers best suited for low and ultra-low pass sequencing data.

Introduction

Structural variants (SVs) are genomic regions that contain an altered DNA sequence due to deletion, duplication, insertion, or inversion[1]. SVs are present in approximately 1.5% of the human genome[1,2], but this small subset of genetic variation has been implicated in the pathogenesis of psoriasis[3], Crohn's disease[4] and other autoimmune disorders[5], autism spectrum and other neurodevelopmental disorders[6–9], and schizophrenia[10–13]. Specialized computational methods—often referred to as SV callers—are capable of detecting structural variants directly from sequencing data. At present, the reliability, sensitivity, and precision of SV callers has not been systematically assessed. We benchmarked currently-available WGS-based SV callers in order to determine the efficacy of available tools and find methods with a good balance between sensitivity and precision.

Substantial differences exist in the number of identified variants in SV catalogs published during the past decade. The 1000 Genomes Project SV dataset identified over 68,000 SVs[14], a genome-wide survey of 769 Dutch individuals identified approximately 1.9 million structural variants[15], and a survey based on profiled whole genomes of 14,891 individuals across diverse global populations identified 498,257 SVs[16]. In addition, discrepancies in the number of SVs reported by these methods suggest that SV callers may fail to detect SVs and may report false positives (i.e., SVs that do not actually exist).

Lack of comprehensive benchmarking makes it impossible to adequately compare the performance of SV callers. In the absence of benchmarking, biomedical studies rely on the consensus of several SV callers[16,17]. In order to compare SV callers given the current lack of a comprehensive gold standard dataset, a recent study[18] used long read technologies to define a ground truth in order to evaluate a large number of currently available tools. However, a comprehensive gold standard dataset is still needed; long read technologies are often unable to cover the entire genome at sufficient resolution for consistent detection of SVs. In addition, current long read technologies are prone to producing high error rates, which confounds efforts to detect SVs at single-base pair resolution. In response to the pressing need for a comprehensive gold standard dataset, our paper presents a rigorous assessment of sensitivity and precision of SV-detection tools when applied to mouse data.

Results

Preparing the gold standard data and WGS data

Over the last decade, a plethora of SV-detection methods have been developed (Table 1 and Supplemental Table 1), but the relative performance of these tools is unknown[19–25]. In order to assess the precision and accuracy of currently available SV callers, we simplified the problem presented to the detectors by using a set of homozygous deletions present in inbred mouse chromosomes. Methods failing to detect deletions are likely unreliable for the more challenging task of identifying other SV categories (e.g., insertions, inversions, translocations). We manually curated the mouse deletions used in this benchmarking study, and we used targeted PCR amplification of the breakpoints and sequencing to resolve the ends of each deletion to the base pair[26]. We only used deletions since we could not confidently determine that other forms of SVs could be comprehensively detected with today's SV callers.

The set of deletions we used among seven inbred strains, called with reference to C57BL/6J, is illustrated in Figure 1a and listed in Supplemental Table 2[26]. We filtered out deletions shorter than 50 bp, as such genomic events that are usually detected by indel callers rather than SV callers. In total, we obtained 3,710 deletions with lengths ranging from 50 to 239,572 base pairs (Supplemental Figure 1 and Supplemental Table 2[26]). Almost half of the deletions were in the range of 100-500 bp. Almost 30% of deletions were larger than 1000 bp (Supplemental Figure 1). High coverage sequence data was used as an input to the SV callers in the form of aligned reads. Reads were mapped to the mouse genome (GRCm38 Mouse Build) using BWA with -a option. In total, we obtained 5.2 billion 2x100 bp paired end reads across seven mouse strains. The average depth of coverage was 50.75x (Supplemental Table 3). Details regarding the gold standard and raw data preparation and analysis are presented in the Supplementary Materials.

Choice of SV callers

For this benchmarking study, we selected methods capable of detecting SVs from aligned WGS reads. SV detection algorithms typically use information about coverage profile in addition to the alignment patterns of abnormal reads. We excluded tools that were designed to detect SVs in tumor-normal samples (e.g., Patchwork[86], COPS[87], rSW-seq[88], bic-seq[63], seqCBS[89]) and tools designed to detect only small (less than 50 bp in length) SVs (e.g., GATK[91], Platypus[92], Varscan[93]). Some tools were not suitable for inclusion in our dataset as they were unable to process aligned WGS data (e.g., Magnolya[27]). Other tools were designed solely for long reads (e.g., Sniffles[28]). The complete list of tools excluded from our analysis are provided in

Supplemental Table 4. In total, we identified 55 suitable SV methods capable of detecting deletions from WGS data (Table 1 and Supplemental Table 1).

Our benchmarking study produced an analysis of the results generated by 12 SV-detection tools (Table 1). We were able to internally install and run all tools except Biograph, which was run by the developers of the tool. The remaining 43 tools could not be installed and were not included in this study. Supplemental Table 4 presents detailed information about the issues that prevented us from installing these software tools. Commands to install the tools and details of the installation process are provided in the Supplementary Materials.

Comparing the performance of SV callers on mouse WGS data

We compared the performance of 12 SV callers in terms of inferring deletions. The number of deletions detected varied from 899 (indelMINER[29]) to 82,225 (GASV[38]). 50% of the methods reported fewer deletions than are known to be present in the sample (Figure 1b). We allowed deviation in the coordinates of the detected deletions and compared deviations to the coordinates of the true deletions. Even at relaxed stringency, the best method correctly detected the breakpoints of only 20% of known deletions in our curated dataset.

The majority of SV callers typically detect deletions whose coordinates differ from the correct positions by up to 100 bp . Figure 1c and 1d show the true positive (TP) and true negative (TN) rates for the SV callers at four different resolution values. It is notable that some tools with high TP rates also have decreased TN rates. For example, at the 100 bp threshold, the highest TP rate was achieved by CLEVER[30] followed by GRIDSS[39] and DELLY[31] (Figure 1c). However, for the same threshold, GRIDSS[39] and DELLY[31] underperform in the number of correctly detected non-deletions (TNs) compared to tools like LUMPY[94] (Figure 1d-f). The total number of false negative (FN) and false positive (FP) calls decreased with increase in threshold (Supplemental Figure 2). The FP rate for pindel Popdel[93] was more susceptible to changes in the threshold as compared to Pindel[39], GASV[38]. In general, the length distribution of detected deletions varied across tools and was substantially different from the distribution of true deletions across multiple SV detection methods (Figure 2 and Supplemental Table 2). Deletions detected by BreakDancer[32] were the closest to the true median deletion length, while five out of 12 SV callers overestimated deletion lengths (Figure 2).

Increasing the resolution threshold increases the number of deletions detected by the SV callers (Figure 1c). Several methods detected all deletions in the sample at 10,000 bp resolution but with precision close to zero (Figure 3b). We used the harmonic mean between precision and sensitivity (F-score) rates to determine the method with the best balance between sensitivity and precision. Several methods (e.g., LUMPY[94], BreakDancer[32], CLEVER[30], BioGraph) offered the highest F-score for deletion detection—consistently between 100-10,000 bp resolution across all the mouse strains (Supplemental Figure 3). For a resolution of 10 bp, the method with the best performance for all the samples was LUMPY[94] (Supplemental Figure 3). The method with the best precision for a threshold of 100-1,000 bp was PopDel[93], but the sensitivity rate of PopDel[93] did not exceed 50% (Figure 3a, Supplemental Figure 4 and 5).

While specificity is often used o compare the deletions detected by tools, use specificity to provide insights on the tools' ability to predict diploid regions of the genome. Methods that produced a higher F-score tend to also have significantly higher specificity rates with Spearman's correlations greater than 0.75 (Figure 3d and Supplemental Figure 3 and 6) and are the most balanced in precision and sensitivity; few methods skewed towards just one of the metrics (Figure 3e and Supplemental Figure 7.) Specificity rate was generally lower for the majority of the methods when compared with sensitivity rate. Methods with a high precision tend to also have significantly higher specificity rates, with Spearman's correlation greater than 0.8 (p-value<0.0005) (Supplemental Figure 8). Several tools, such as PopDel[93] and LUMPY[94], were able to balance precision and specificity, with rates exceeding 70% for each metric (Figure 3f). LUMPY[94] and CLEVER[30] were the only methods able to successfully balance precision and sensitivity, with rates above 50% for each metric (Figure 3e and Supplemental Figure 7). CLEVER[30] was able to achieve the highest sensitivity rate at the majority of thresholds (Figure 3a and Supplemental Figure 5). The most precise method we observed was PopDel[93], with rates exceeding 80% for thresholds 1000 bp onwards, but the sensitivity of this method was two times lower than the majority of other tools.

We examined whether the SV callers included in this study maintained similar SV detection accuracy across the different mouse strains. We compared results from each tool when applied to the sample with the highest and lowest rates of sensitivity and specificity. Among the tools with a sensitivity rate above 10%, LUMPY[94] maintained the most consistent sensitivity rate across samples, with the highest rate of 60% when applied to both C3H_HeJ and CBA_J strains. The lowest sensitivity rate achieved by LUMPY[94] was 58% for A_J and DBA_2J strains. Several tools, such as PopDel[93], LUMPY[94], and DELLY[31], maintained a consistent specificity across mouse strains (Figure 3d and Supplemental Figure 6).

We have also compared CPU time and the maximum amount of RAM used by each of the tools. Across all of the tools, GASV[38] required the highest amount of computational resources. PopDel[93] has the lowest computational resources required to run the analysis. MiStrVar[40] required the longest amount of time to perform the analysis. Breakdancer[32] was the fastest tool. We have also compared the computational resources and speed of SV callers based on datasets with full coverage and those with ultra-low coverage (Supplemental Figure 9).

Performance of SV-detection tools on low and ultra-low coverage data

We assessed the performance of SV callers at different coverage depths generated by down-sampling the original WGS data. The simulated coverage ranged from 32x to 0.1x, and ten subsamples were generated for each coverage range. For each method, the number of correctly detected deletions generally decreased as the coverage depth decreased (Supplemental Figure 10). Some of the methods were able to call deletions from ultra low coverage (<=0.5x) data. While tools like PopDel[93] reached a precision of 75%, the overall sensitivity, specificity, and F-score values were less than 8% for all tools. None of the methods were able to detect deletions from 0.1x coverage.

As suggested by other studies[89], most tools reached a maximum precision and specificity at an intermediate coverage (Figure 4b-c). Both the sensitivity rate and the F-score improved as the coverage increased (Figure 4a,d). Overall, DELLY[31] showed the highest F-score for coverage below 4x (Figure 4d). For coverage between 8 and 32x, LUMPY[94] showed the best performance. LUMPY[94] was the only tool to attain precision above 90% for coverages 1x to 4x. However, a decreased sensitivity in coverages below 4x led to a decreased F-score when compared to DELLY[31]. Precision in results from DELLY[31] for ultra-low coverage data was above 90% when the threshold was set at 1000 bp, but changing the threshold had no effect on LUMPY[94] (Supplemental Figure 11). Sensitivity rates were the most stable across the 7 different strains (Supplemental Figure 5). Specificity showed the highest variability among strains compared to other measures (Supplemental Figure 6)**.** Precision shows the second highest variability across the strains, with the most stable results provided by Pindel[39] and indelMINER[29] (Supplemental Figure 4).

Length of deletions impacts the performance of the SV callers

We separately assessed the effect of deletion length on the accuracy of detection for four categories of deletions (Figure 5). The performance of the SV callers was significantly affected by deletion length. For example, for deletions shorter than 100 bp, precision, specificity, and F-score values were typically below 40% regardless of the tool (Figure 5b,c,d and Supplemental Figures 12, 14, 15), while sensitivity values were above 50% for several tools (Figure 5a) (Supplemental Figure 13). For deletions longer than 100 bp, the best performing tool in terms of sensitivity and precision significantly varied depending on the deletion length (Figure 5a,b). CLEVER[30] provided a sensitivity of above 60% for deletions less than 500 bp, however DELLY[31] provided the highest sensitivity for deletions longer than 500 bp (Figure 5a and

Supplemental Figures 17, 21, 25) . LUMPY[94] delivered the best precision for deletion lengths from 50-500 bp, and CLEVER[30] performed well for longer deletion lengths (Figure 5b and Supplemental Figure 14, 18, 22, 26). indelMINER[29] provided the high precision rate of detection of deletions in the range of 100 bp-500 bp and when longer than 1000 bp, but the precision of detecting deletion in the 500 bp-1000 bp range was lower than that of other tools (Figure 5b). In general, LUMPY[94] was the only method able to deliver an F-score above 30% across all categories (Figure 5d and Supplemental Figure 15, 19, 23, 27). Specificity was low for all the tools across all the categories, except for deletions with lengths higher than 1,000bp (Figure 5c and Supplemental Figure 12, 16, 20, 24).

Discussion

In this paper, we performed a systematic benchmarking of algorithms to identify structural variants (SVs) from whole-genome sequencing data. In contrast to methods which are used to identify single nucleotide polymorphisms and have coalesced around a small number of approaches, there is currently no consensus on the best way to detect SVs in mammalian genomes. Indeed, we were able to find 56 different methods, each claiming relatively high specificity and sensitivity rates in the original publication. Upon applying the tools to our curated datasets, many did not perform as reported in the original publication. This discrepancy may be because molecular data were not used in the analyses performed for the original publication. Instead, authors often solely derive conclusions from simulated data that may fail to capture the full complexity of real sequencing data[33].

In comparison to previous benchmarking efforts based on simulated data [19,21,25,34,35], we obtained and employed a set of molecularly defined deletions for which breakpoints are known at base pair resolution. Other benchmarking studies have employed long-read-based gold standard datasets with approximate coordinates of deletions[18]. Long read technologies are often unable to cover the entire genome at sufficient resolution for precise SV characterization. In addition, long-read technologies carry a high error rate that limits their ability to detect SVs at single-base pair resolution. Our benchmarking method, using a gold standard set of molecular-defined deletions, overcomes the limitations of simulated data and incomplete characterization. Thus, our benchmarking study represents a robust assessment of the performance of currently-available SV detection methods when applied to a representative data set.

When installing the majority of SV callers, we noticed significant difficulties due to inadequate software implementation and technical factors[36]. Deprecated dependencies and segmentation faults were the most common reasons preventing successful tool installation[37]. The majority of the tools have a consensus on the output format to be used (Table S6) , but the requirements for the format varied among tools. Lack of documentation about format requirements may further limit the use of SV callers.

We identified a series of factors that determined the performance of SV-caller methods. The most important factors were the size of deletions and the coverage of WGS data. For example, BreakDancer[32] only detected deletions larger than 100 bp. Some tools achieved excellent sensitivity, with the caveat that their precision was close to zero. For example, Pindel[32] achieved the highest sensitivity rate among all the tools, with a precision rate of less than 0.1%. Other tools (e.g., PopDel[93]) employ a more conservative SV detection approach, resulting in higher precision at the cost of decreased sensitivity for smaller deletion events. Few tools were able to maintain a good balance between precision and sensitivity. For example, CLEVER[30], LUMPY[94], BreakDancer[32], and BioGraph maintained both precision and sensitivity rates above 40%. In addition to differences in the accuracy of SV detection, we observed significant differences in run times and required computational resources (Supplemental Figure 9).

We envision that future SV-caller methods should enable detection of deletions with precise coordinates. The inability of current methods to precisely detect breakpoints was coupled with the issue of 61.5% tools underestimating the true size of SVs. A limitation of our benchmarking study is that our gold standard used inbred homozygous mouse genomes, which potentially poses as an easier target for assessment when compared to heterozygous human genomes. Additionally, the human genomes, for which most SV callers were designed, contain a higher number of repetitive regions than does the mouse, posing an additional challenge which is not reflected in mouse-based gold standard datasets.

Data availability

VCF file with true deletions from gold standard, and the output VCF's produced by the tools, the gold standard VCF's, the analysis scripts, and figures are available at https://github.com/Mangul-Lab-USC/benchmarking-sv-callers-paper/

Code availability

Source code to compare SV detection methods and to produce the figures contained within this text is open source, free to use under the MIT license, and available at https://github.com/Mangul-Lab-USC/benchmarking-sv-callers-paper/

Acknowledgments

We thank Dr. Lana Martin for the helpful discussions and comments on the manuscript. We thank our colleagues Niranjan Shekar, Lisa Herta, and Adam English at Spiral Genetics for developing and running their tool, BioGraph software (http://www.spiralgenetics.com/biograph-engine) on mouse WGS data. We thank the authors

of the tools surveyed in this work—Cenk Sahinalp, Ryan Layer, Ira Hall, Tony Papenfuss, Gerton Lunter, Michael Schatz, Alexander Schoenhuth, Ken Chen, Aakrosh Ratan, and Tobias Rausch—for providing helpful feedback and verifying information related to their tool.

Author contributions

V.S. created scripts for running and evaluating the software tools. E.W., J.C., M.G.D., N.D.-M., R.A., R.C., R.L., S.C., and V.S. contributed to installing, running, and evaluating software tools. S.N. applied GRIDSS, LUMPY, Manta, and PopDel[93] to the mouse data and discussed evaluation metrics. R.A. curated data and prepared mouse data for evaluation with the software tools. A.L., R.A., and V.S. generated the figures; R.A. and V.S. generating the tables. E.E., J.F., M.G.D., S.M., S.N., and V.S. wrote, reviewed, and edited the manuscript. J.F. and S.M. led the project.

Funding

S.N. is funded by the BMBF Grant #031L0180 from the German Federal Ministry for Education and Research.

Competing interests

The authors declare that they have no competing interests.

References Cited

1. Feuk, L. *et al.* Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. *Am. J. Hum. Genet.* **79**, 965–972 (2006).
2. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11**, R52 (2010).
3. Hollox, E. J., Barber, J. C. K., Brookes, A. J. & Armour, J. A. L. Defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1. *Genome Res.* **18**, 1686–1697 (2008).
4. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
5. Fanciulli, M. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39**, 721–723 (2007).

6. Girirajan, S. & Eichler, E. E. De novo CNVs in bipolar disorder: recurrent themes or new directions? *Neuron* vol. 72 885–887 (2011).

7. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).

8. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).

9. Elia, J. *et al.* Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nat. Genet.* **44**, 78–84 (2011).

10. Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142–153 (2012).

11. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).

12. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).

13. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).

14. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

15. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).

16. Collins, R. L. *et al.* An open resource of structural variation for medical and population genetics. *bioRxiv* 578674 (2019) doi:10.1101/578674.

17. Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).

18. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).

19. Alkodsi, A., Louhimo, R. & Hautaniemi, S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief. Bioinform.* **16**, 242–254 (2015).

20. Pabinger, S., Rödiger, S., Kriegner, A., Vierlinger, K. & Weinhäusel, A. A survey of tools for the analysis of quantitative PCR (qPCR) data. *Biomolecular Detection and Quantification* **1**, 23–33 (2014).

21. Duan, J., Zhang, J.-G., Deng, H.-W. & Wang, Y.-P. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* **8**, e59128 (2013).

22. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).

23. Legault, M.-A., Girard, S., Lemieux Perreault, L.-P., Rouleau, G. A. & Dubé, M.-P. Comparison of sequencing based CNV discovery methods using monozygotic twin quartets. *PLoS One* **10**, e0122287 (2015).

24. Hasan, M. S., Wu, X. & Zhang, L. Performance evaluation of indel calling tools using real short-read data. *Hum. Genomics* **9**, 20 (2015).

25. Neuman, J. A., Isakov, O. & Shomron, N. Analysis of insertion-deletion from deep-

sequencing data: software evaluation for optimal detection. *Brief. Bioinform.* **14**, 46–55 (2013).

26. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
27. Nijkamp, J. F. *et al.* De novo detection of copy number variation by co-assembly. *Bioinformatics* **28**, 3195–3202 (2012).
28. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
29. Ratan, A., Olson, T. L., Loughran, T. P., Jr & Miller, W. Identification of indels in next-generation sequencing data. *BMC Bioinformatics* **16**, 42 (2015).
30. Marschall, T. *et al.* CLEVER: clique-enumerating variant finder. *Bioinformatics* **28**, 2875–2882 (2012).
31. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
32. Fan, X., Abbott, T. E., Larson, D. & Chen, K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. in *Current Protocols in Bioinformatics* 15.6.1–15.6.11 (2014).
33. Mangul, S. *et al.* Systematic benchmarking of omics computational tools. *Nat. Commun.* **10**, 1393 (2019).
34. Guan, P. & Sung, W.-K. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* **102**, 36–49 (2016).
35. Cameron, D. L., Di Stefano, L. & Papenfuss, A. T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat.*

Commun. **10**, 3240 (2019).

36. Mangul, S., Martin, L. S., Eskin, E. & Blekhman, R. Improving the usability and archival stability of bioinformatics software. *Genome Biol.* **20**, 47 (2019).

37. Mangul, S. *et al.* Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLoS Biol.* **17**, e3000333 (2019).

38. Sindi, S., Helman, E., Bashir, A. & Raphael, B. J. A geometric approach for classification and comparison of structural variants. *Bioinformatics* **25**, i222–30 (2009).

39. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

40. Lin, Y.-Y. *et al.* Computational identification of micro-structural variations and their proteogenomic consequences in cancer. *Bioinformatics* **34**, 1672–1681 (2018).

41. Qi, J. & Zhao, F. inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/21715388.

42. Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. Detecting copy number variation with mated short reads. *Genome Res.* **20**, 1613–1622 (2010).

43. Jiang, Y., Wang, Y. & Brudno, M. PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **28**, 2576–2583 (2012).

44. Karakoc, E. *et al.* Detection of structural variants and indels within exome data. *Nat. Methods* **9**, 176–178 (2011).

45. Hajirasouliha, I. *et al.* Detection and characterization of novel sequence insertions using

paired-end next-generation sequencing. *Bioinformatics* **26**, 1277–1283 (2010).

46. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919 (2013).

47. Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* **6**, 473–474 (2009).

48. Korbel, J. O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10**, R23 (2009).

49. Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M. & Boeva, V. SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability. *Bioinformatics* **32**, 984–992 (2016).

50. Mimori, T. *et al.* iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. *BMC Syst. Biol.* **7**, 1–8 (2013).

51. Vikas Bansal, O. L. A probabilistic method for the detection and genotyping of small indels from population-scale sequence data. *Bioinformatics* **27**, 2047 (2011).

52. Albers, C. A. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res.* **21**, 961–973 (2011).

53. Bartenhagen, C. & Dugas, M. Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. *Brief. Bioinform.* **17**, 51–62 (2016).

54. Dan, S. *et al.* Prenatal detection of aneuploidy and imbalanced chromosomal arrangements by massively parallel sequencing. *PLoS One* **7**, e27835 (2012).

55. Suzuki, S., Yasuda, T., Shiraishi, Y., Miyano, S. & Nagasaki, M. ClipCrop: a tool for

detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* **12**, S7 (2011).

56. Emde, A.-K. *et al.* Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics* **28**, 619–627 (2012).
57. Chen, K. *et al.* TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Res.* **24**, 310 (2014).
58. Simpson, J. T., McIntyre, R. E., Adams, D. J. & Durbin, R. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics* **26**, 565 (2010).
59. Magi, A., Benelli, M., Yoon, S., Roviello, F. & Torricelli, F. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res.* **39**, e65 (2011).
60. Ivakhno S, E. al. CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/20966003.
61. Sindi, S. S., Önal, S., Peng, L. C., Wu, H.-T. & Raphael, B. J. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.* **13**, R22 (2012).
62. Wong, K., Keane, T. M., Stalker, J. & Adams, D. J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **11**, R128 (2010).
63. Xi, R., Lee, S., Xia, Y., Kim, T.-M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* **44**, 6274–6286 (2016).

64. Chiara, M., Pesole, G. & Horner, D. S. SVM2: an improved paired-end-based tool for the detection of small genomic structural variations using high-throughput single-genome resequencing data. *Nucleic Acids Res.* **40**, e145 (2012).

65. Bellerophon. http://cbc.case.edu/Bellerophon/.

66. Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–1112 (2014).

67. Xiao, F. *et al*. modSaRa: a computationally efficient R package for CNV identification. *Bioinformatics* **33**, 2384–2385 (2017).

68. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. *PLoS One* **6**, e16327 (2011).

69. Budczies, J. *et al.* Ioncopy: an R Shiny app to call copy number alterations in targeted NGS data. *BMC Bioinformatics* **19**, 157 (2018).

70. Website. https://github.com/xyc0813/SVmine.

71. Noll, A. C. *et al.* Clinical detection of deletion structural variants in whole-genome sequences. *NPJ Genom Med* **1**, 16026 (2016).

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

73. Lindberg, M. R., Hall, I. M. & Quinlan, A. R. Population-based structural variation discovery with Hydra-Multi. *Bioinformatics* **31**, 1286–1289 (2015).

74. Klambauer, G. *et al.* cn.MOPS: mixture of Poissons for discovering copy number variations

in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 (2012).

75. Zhao, H. & Zhao, F. BreakSeek: a breakpoint-based algorithm for full spectral range INDEL detection. *Nucleic Acids Res.* **43**, 6701–6713 (2015).
76. Li, H. FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **31**, 3694–3696 (2015).
77. Lam, H. Y. K. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28**, 47–55 (2010).
78. Toolkit for automated and rapid discovery of structural variants. *Methods* **129**, 3–7 (2017).
79. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011).
80. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
81. Stancu, M. C. *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **8**, 1–13 (2017).
82. English, A. C., Salerno, W. J. & Reid, J. G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* **15**, 1–7 (2014).
83. Hormozdiari, F. *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**, i350 (2010).
84. Kim, M., Farnoud, F. & Milenkovic, O. HyDRA: gene prioritization via hybrid distance-score rank aggregation. *Bioinformatics* **31**, 1034–1043 (2015).
85. Schröder, J. *et al.* Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics* **30**, 1064 (2014).

86. Mayrhofer, M., DiLorenzo, S. & Isaksson, A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biol.* **14**, 1–10 (2013).

87. Krishnan, N. M., Gaur, P., Chaudhary, R., Rao, A. A. & Panda, B. COPS: A Sensitive and Accurate Tool for Detecting Somatic Copy Number Alterations Using Short-Read Sequence Data from Paired Samples. *PLoS One* **7**, (2012).

88. Kim, T.-M., Luquette, L. J., Xi, R. & Park, P. J. rSW-seq: Algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics* **11**, 1–13 (2010).

89. Chen, H., Bell, J. M., Zavala, N. A., Ji, H. P. & Zhang, N. R. Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res.* **43**, e23 (2015).

90. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

91. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

92. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

93. Roskosch, Sebastian, et al. "PopDel identifies medium-size deletions jointly in tens of thousands of genomes." *bioRxiv* (2019): 740225.

94. Layer, Ryan M., et al. "LUMPY: a probabilistic framework for structural variant discovery." *Genome biology* 15.6 (2014): R84.

95. Yoon, Seungtai, et al. "Sensitive and accurate detection of copy number variants using read depth of coverage." *Genome research* 19.9 (2009): 1586-1592.

96. Sedlazeck, Fritz J., et al. "Accurate detection of complex structural variations using single-molecule sequencing." Nature methods 15.6 (2018): 461-468.

97. Kosugi, Shunichi, et al. "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing." Genome biology 20.1 (2019): 117.

Tables

Software tool	Version	Underlying algorithm	Published year	Tool's webpage	Bioconda version	Format
GASV[38]	1.4	RP	2009	http://compbio.cs.brown.edu/projects/GASV/	No	Custom
Pindel[39]	0.2.5b9	RP+SR SR (2015, 0.2.5b9)	2009	http://gmt.genome.wustl.edu/packages/pindel/	Yes	Custom
RDXplorer[95]	3.2	RD	2009	http://RDXplorer.sourceforge.net/	No	Custom
CLEVER[30]	2.4	RP	2012	https://bitbucket.org/tobiasmarschall/CLEVER-toolkit/wiki/Home	Yes	Custom
DELLY[31]	0.8.2	RP+SR	2012	https://github.com/DELLYtools/DELLY	Yes	Custom

BreakDancer[32]	1.3.6	RP	2014	https://github.com/genome/BreakDancer	Yes	Custom
indelMINER[29]	N/A	RP+SR	2015	https://github.com/aakrosh/indelMINER	No	VCF
GRIDSS[39]	2.5.1	RP+SR	2017	https://github.com/PapenfussLab/GRIDSS	Yes	VCF
MiStrVar[40]	N/A	N/A	2017	https://bitbucket.org/compbio/MiStrVar	No	VCF
LUMPY[94]	0.2.4	RP, SR, RD	2018	https://github.com/brentp/smoove	Yes	VCF
PopDel[93]	1.1.3	RP	2019	https://github.com/kehrlab/PopDel	Yes	VCF
BioGraph*	5.0.1	N/A	2020	http://www.spiralgenetics.com/biograph-engine	No	VCF

Table 1. Overview of SV-detection methods included in this study. Surveyed SV-detection methods sorted by their year of publication from 2009 to 2018 are listed along with their underlying algorithm: Read-depth (RC), Read-Pair Algorithms (RP), Split-Read Approaches (SR), Discordant Pairs (DP), or a combination of algorithms. We documented the version of the software tool used in the study (‘Version’), the year the software tool was published (‘Published year’), the webpage where each SV-detection method is hosted (‘Tool’s webpage’), and whether or not the Bioconda package of the software was available (‘Bioconda version’). Asterisk (*) denotes that the method was proprietary.

Figures

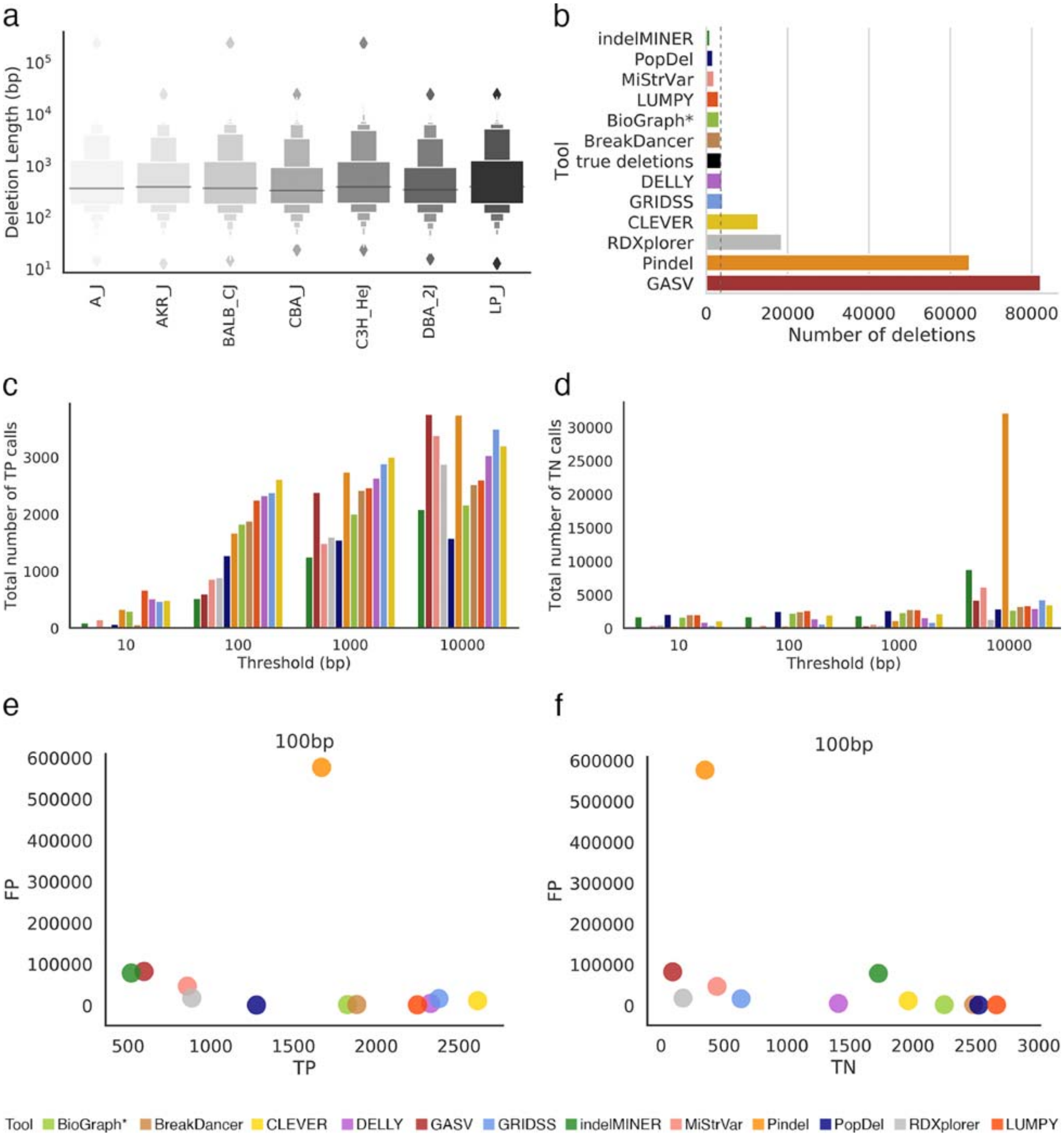


Figure 1. Comparison of inferred deletions across SV callers on mouse data. (a) Length distribution of molecularly-confirmed deletions from chromosome 19 across seven strains of mice. (b) Number of molecularly-confirmed deletions ('true deletions' black color) and number of deletions detected by SV callers. (c) Barplot depicting the total number of true positive calls across all error thresholds for each SV caller. (d) Barplot depicting the total number of true

negative calls across all error thresholds for each SV caller. (e) Scatter plot depicting number of correctly detected deletions (true positives - 'TP') by number of incorrectly detected deletions (false positives - 'FP') at the 100 bp threshold. Deletion is considered to be correctly predicted if the distance of right and left coordinates are within the given threshold from the coordinates of true deletion. (f) Scatter plot depicting number of correctly detected non-deletions (true negatives - 'TN') by number of incorrectly detected deletions (false positives - 'FP') at the 100 bp threshold. An SV caller was considered to detect a given non-deletion if no deletions were reported in a given region.

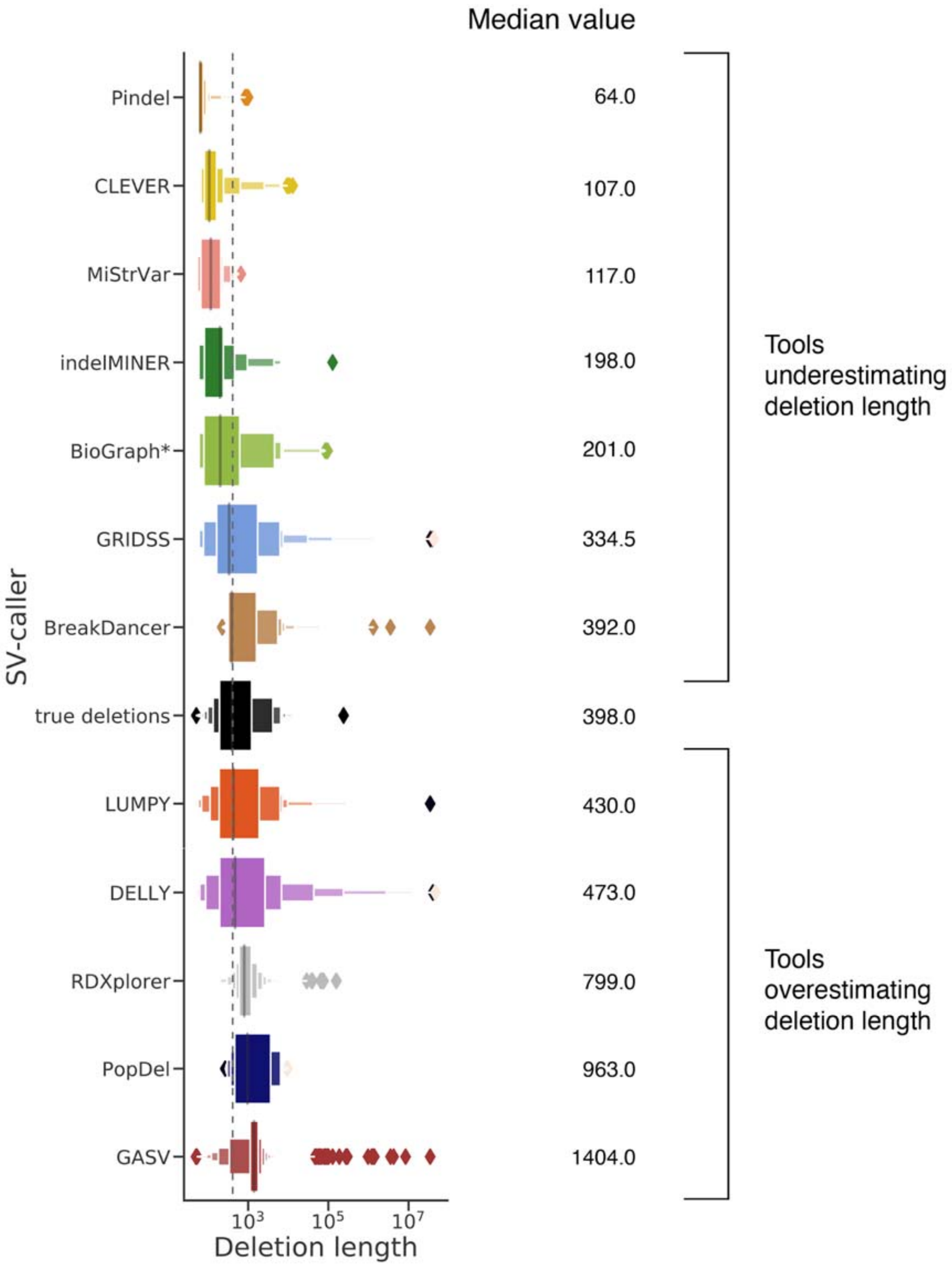


Figure 2. Length distribution of deletions detected by each SV caller. True deletions indicated in black. Tools were sorted in increasing order based on their median deletion length. The vertical dashed line corresponds to the median value of true deletions

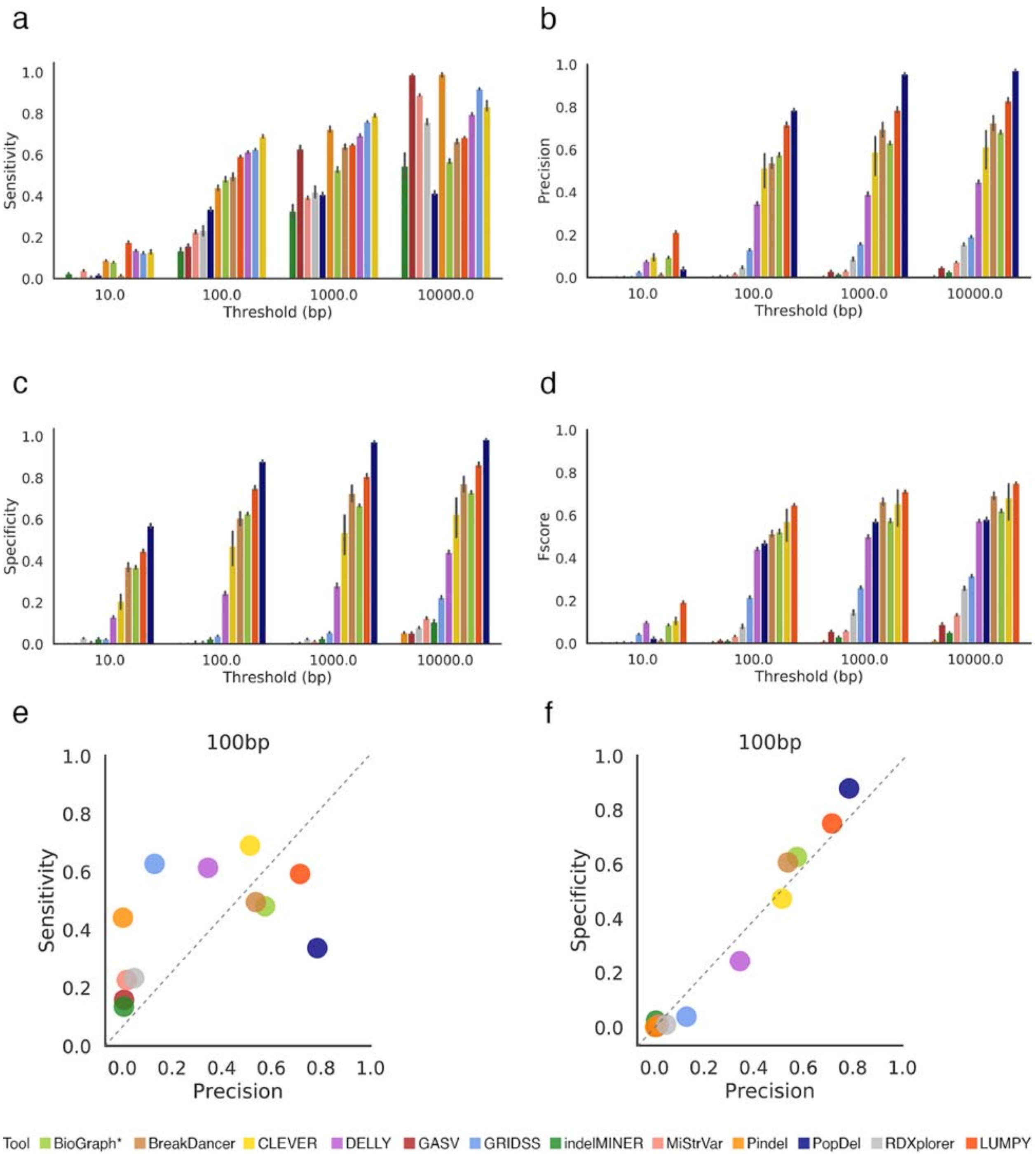


Figure 3. Comparing the performance of SV callers based on whole genome (WGS) data across seven inbred mouse strains. A deletion is considered to be correctly predicted if the distance of right and left coordinates are within the threshold τ from the coordinates of a true deletion. (a) Sensitivity of SV callers at different thresholds. (b) Precision of SV callers at different thresholds. (c) Specificity of SV callers at different thresholds. (d) F-score of SV callers

at different thresholds. (e) Scatter plot depicting the Precision (x-axis) and Sensitivity (y-axis) for 100 bp threshold. (f) Scatter plot depicting the Precision(x-axis) and Specificity (y-axis) 100 bp threshold. Results for other thresholds are presented in Supplemental Figure 6.

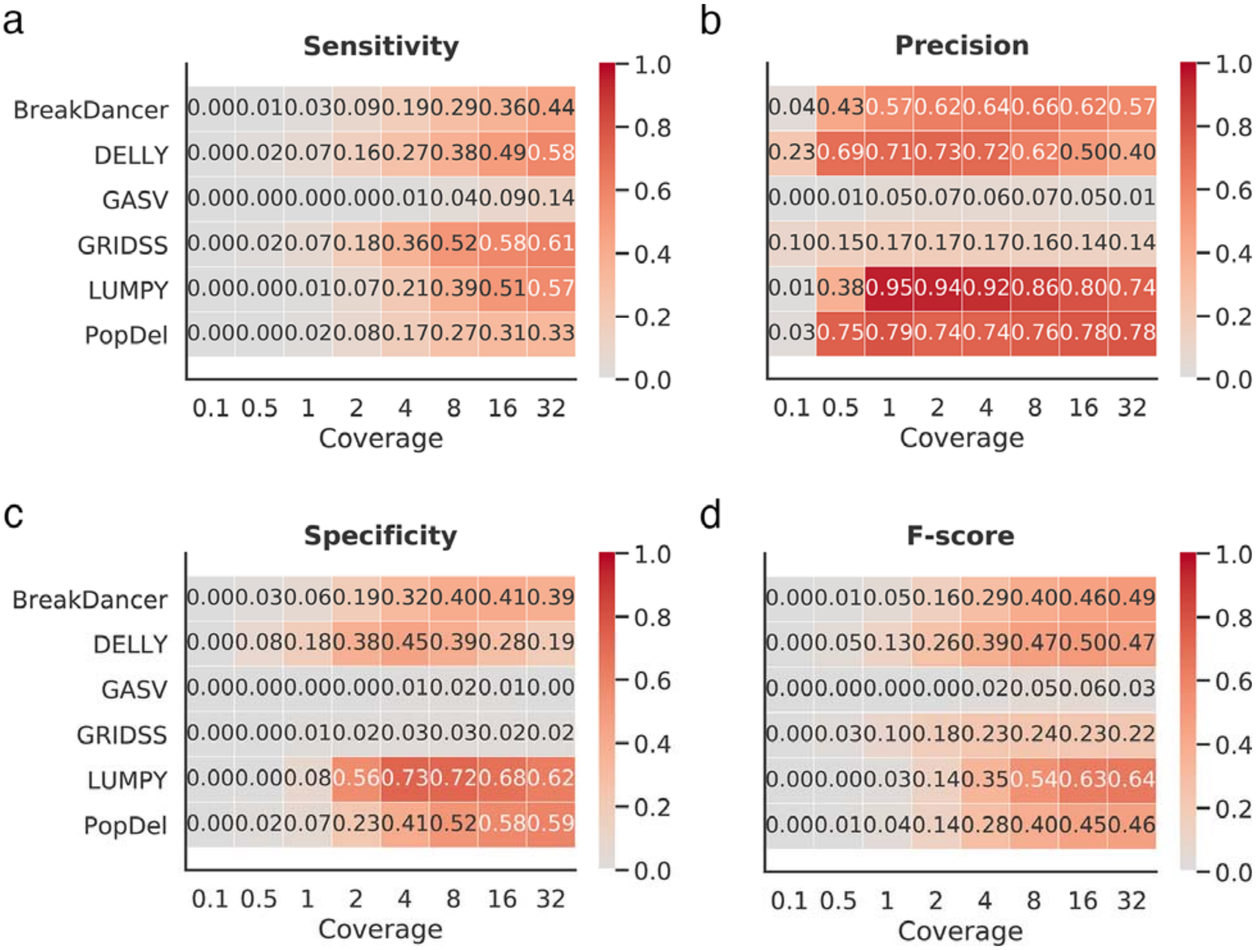


Figure 4. Performance of SV-detection tools on low and ultra-low coverage data. (a) Heatmap depicting the sensitivity based on 100 bp threshold across various levels of coverage. (b) Heatmap depicting the precision based on 100 bp threshold across various levels of coverage. (c) Heatmap depicting the specificity based on 100 bp threshold across various levels of coverage. (d) Heatmap depicting the F-score based on 100 bp threshold across various levels of coverage.

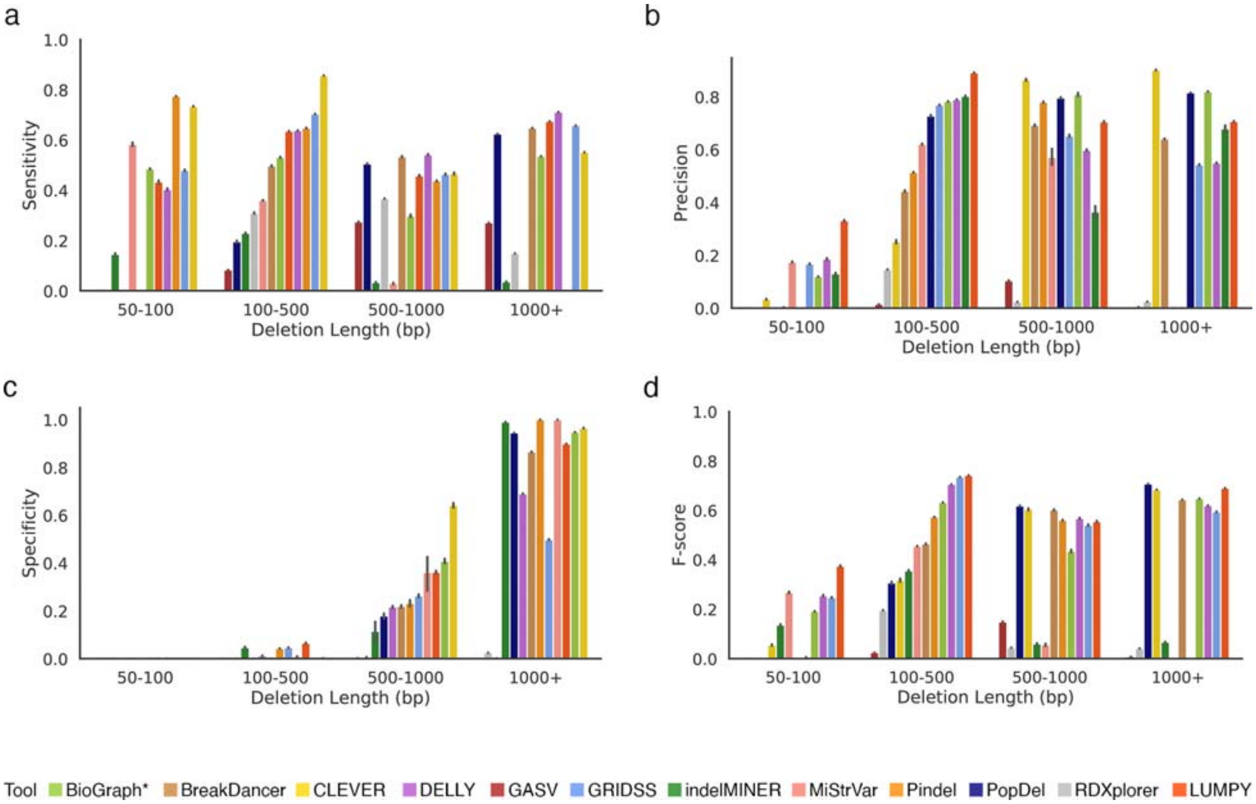
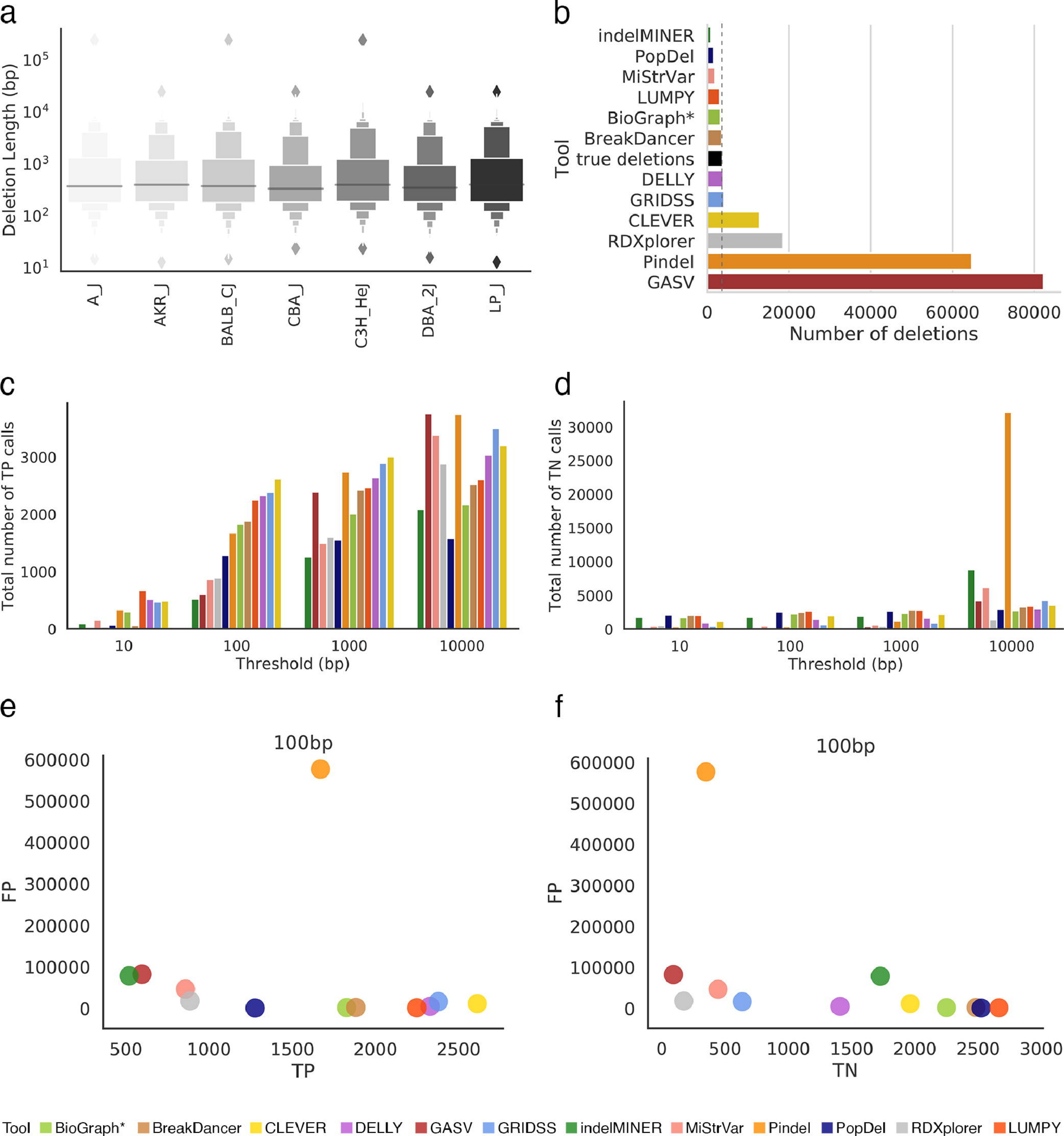
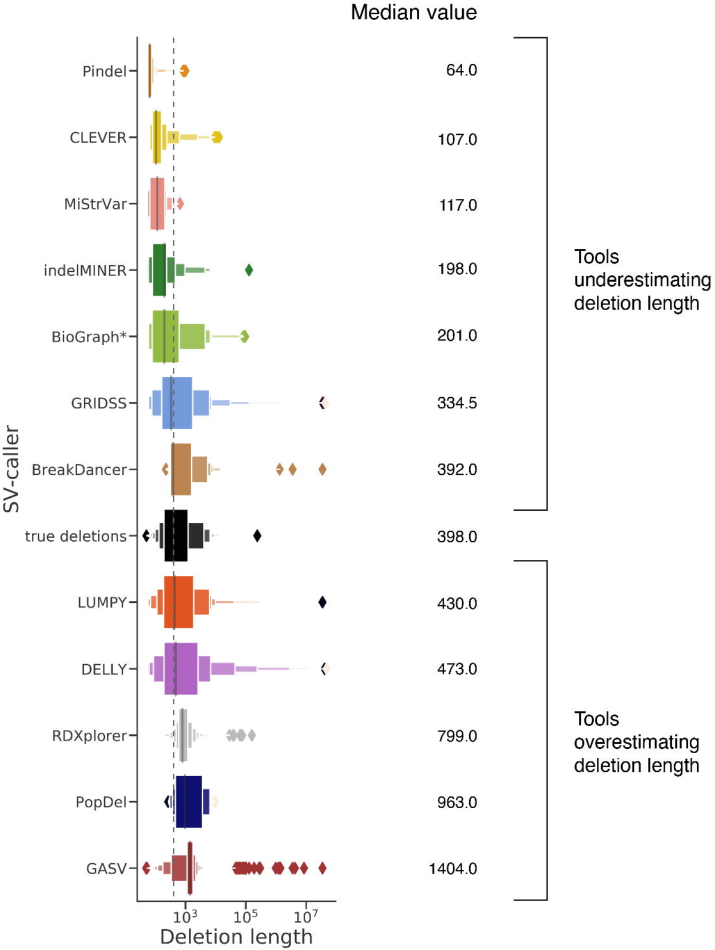
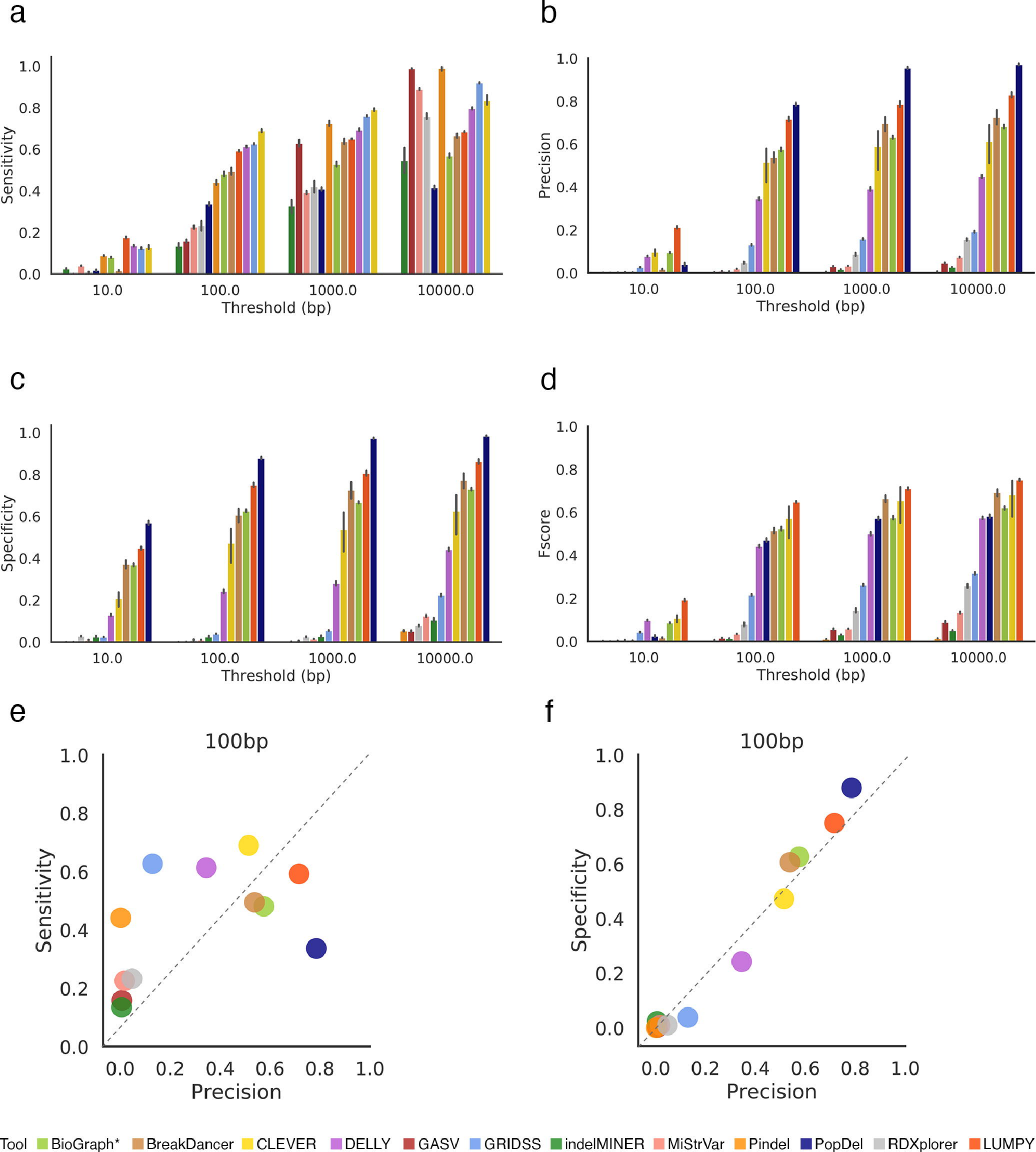
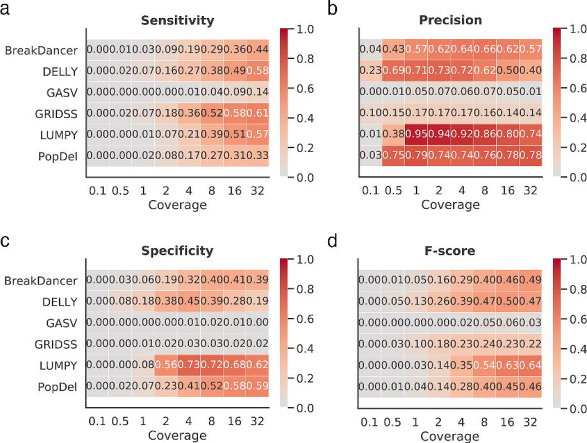


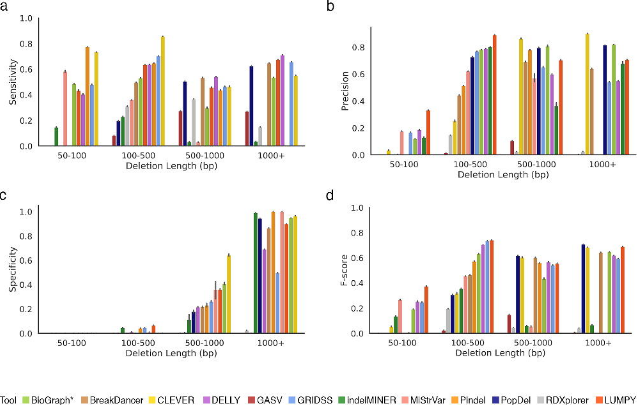
Figure 5. Comparing the performance of SV callers across various deletions lengths. (a) Sensitivity of SV callers at 100 bp thresholds across deletion length categories. (b) Precision of SV callers at 100 bp thresholds across deletion length categories. (c) Specificity of SV callers at 100 bp thresholds across deletion length categories. (d) F-score of SV callers at 100 bp thresholds across deletion length categories.


a
Deletion Length (bp)
10^1
10^2
10^3
10^4
10^5
A_J
AKR_J
BALB_CJ
CBA_J
C3H_HeJ
DBA_2J
LP_J
b
Tool
indelMINER
PopDel
MiStrVar
LUMPY
BioGraph*
BreakDancer
true deletions
DELLY
GRIDSS
CLEVER
RDXplorer
Pindel
GASV
0
20000
40000
60000
80000
Number of deletions
c
Total number of TP calls
0
1000
2000
3000
10
100
1000
10000
Threshold (bp)
d
Total number of TN calls
0
5000
10000
15000
20000
25000
30000
10
100
1000
10000
Threshold (bp)
e
100bp
FP
0
100000
200000
300000
400000
500000
600000
500
1000
1500
2000
2500
TP
f
100bp
FP
0
100000
200000
300000
400000
500000
600000
0
500
1000
1500
2000
2500
3000
TN
Tool
BioGraph*
BreakDancer
CLEVER
DELLY
GASV
GRIDSS
indelMINER
MiStrVar
Pindel
PopDel
RDXplorer
LUMPY


Median value
Pindel
64.0
CLEVER
107.0
MiStrVar
117.0
indelMINER
198.0
BioGraph*
201.0
GRIDSS
334.5
BreakDancer
392.0
true deletions
398.0
LUMPY
430.0
DELLY
473.0
RDXplorer
799.0
PopDel
963.0
GASV
1404.0
Tools underestimating deletion length
Tools overestimating deletion length
SV-caller
10^3
10^5
10^7
Deletion length


a
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
10.0
100.0
1000.0
10000.0
Threshold (bp)
b
Precision
0.0
0.2
0.4
0.6
0.8
1.0
10.0
100.0
1000.0
10000.0
Threshold (bp)
c
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
10.0
100.0
1000.0
10000.0
Threshold (bp)
d
Fscore
0.0
0.2
0.4
0.6
0.8
1.0
10.0
100.0
1000.0
10000.0
Threshold (bp)
e
100bp
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Precision
0.0
0.2
0.4
0.6
0.8
1.0
f
100bp
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Tool
BioGraph*
BreakDancer
CLEVER
DELLY
GASV
GRIDSS
indelMINER
MiStrVar
Pindel
PopDel
RDXplorer
LUMPY


a
Sensitivity
BreakDancer 0.00 0.01 0.03 0.09 0.19 0.29 0.36 0.44
DELLY 0.00 0.02 0.07 0.16 0.27 0.38 0.49 0.58
GASV 0.00 0.00 0.00 0.00 0.01 0.04 0.09 0.14
GRIDSS 0.00 0.02 0.07 0.18 0.36 0.52 0.58 0.61
LUMPY 0.00 0.00 0.01 0.07 0.21 0.39 0.51 0.57
PopDel 0.00 0.00 0.02 0.08 0.17 0.27 0.31 0.33
0.1 0.5 1 2 4 8 16 32
Coverage
1.0 0.8 0.6 0.4 0.2 0.0
b
Precision
0.04 0.43 0.57 0.62 0.64 0.66 0.62 0.57
0.23 0.69 0.71 0.73 0.72 0.62 0.50 0.40
0.00 0.01 0.05 0.07 0.06 0.07 0.05 0.01
0.10 0.15 0.17 0.17 0.17 0.16 0.14 0.14
0.01 0.38 0.95 0.94 0.92 0.86 0.80 0.74
0.03 0.75 0.79 0.74 0.74 0.76 0.78 0.78
0.1 0.5 1 2 4 8 16 32
Coverage
1.0 0.8 0.6 0.4 0.2 0.0
c
Specificity
BreakDancer 0.00 0.03 0.06 0.19 0.32 0.40 0.41 0.39
DELLY 0.00 0.08 0.18 0.38 0.45 0.39 0.28 0.19
GASV 0.00 0.00 0.00 0.00 0.01 0.02 0.01 0.00
GRIDSS 0.00 0.00 0.01 0.02 0.03 0.03 0.02 0.02
LUMPY 0.00 0.00 0.08 0.56 0.73 0.72 0.68 0.62
PopDel 0.00 0.02 0.07 0.23 0.41 0.52 0.58 0.59
0.1 0.5 1 2 4 8 16 32
Coverage
1.0 0.8 0.6 0.4 0.2 0.0
d
F-score
0.00 0.01 0.05 0.16 0.29 0.40 0.46 0.49
0.00 0.05 0.13 0.26 0.39 0.47 0.50 0.47
0.00 0.00 0.00 0.00 0.02 0.05 0.06 0.03
0.00 0.03 0.10 0.18 0.23 0.24 0.23 0.22
0.00 0.00 0.03 0.14 0.35 0.54 0.63 0.64
0.00 0.01 0.04 0.14 0.28 0.40 0.45 0.46
0.1 0.5 1 2 4 8 16 32
Coverage
1.0 0.8 0.6 0.4 0.2 0.0


a
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
50-100
100-500
500-1000
1000+
Deletion Length (bp)
b
Precision
0.8
0.6
0.4
0.2
0.0
50-100
100-500
500-1000
1000+
Deletion Length (bp)
c
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
50-100
100-500
500-1000
1000+
Deletion Length (bp)
d
F-score
1.0
0.8
0.6
0.4
0.2
0.0
50-100
100-500
500-1000
1000+
Deletion Length (bp)
Tool
BioGraph*
BreakDancer
CLEVER
DELLY
GASV
GRIDSS
indelMINER
MiStrVar
Pindel
PopDel
RDXplorer
LUMPY