

Comprehensive evaluation and guidance of structural variation

detection tools in chicken whole genome sequence data

Cheng Ma^{1,2,†}, Xian Shi^{1,3,†}, Xuzhen Li^{4,5,†}, Ya-Ping Zhang^{1,3,6,7,*}, Min-Sheng Peng^{1,3,7,*}

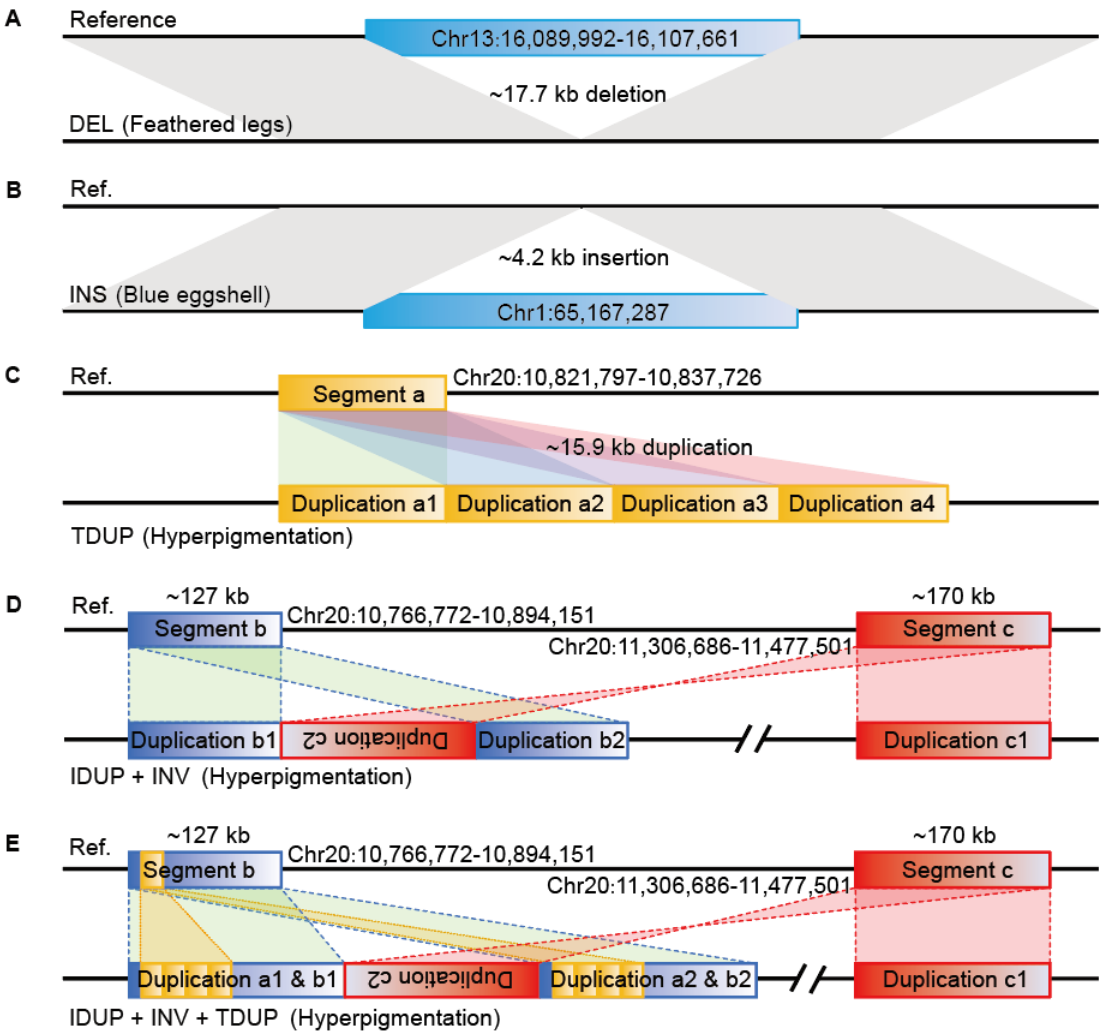
1. Key Laboratory of Genetic Evolution & Animal Models and Yunnan Key Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.
2. Department of Medical Biochemistry and Microbiology, Uppsala University, BMC, Uppsala SE-75123, Sweden.
3. University of Chinese Academy of Sciences, Beijing 100049, China.
4. State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan Agricultural University, Kunming 650201, China.
5. College of Biological Big Data, Yunnan Agriculture University, Kunming 650201, China.
6. State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan University, Kunming 650091, China.
7. KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.

† These authors contributed equally to this work.

* Correspondence: Tel: +86 871-68526519; Fax: +86 871-68526519; Email:

pengminsheng@mail.kiz.ac.cn or zhangyp@mail.kiz.ac.cn

23 **Supplemental Figures**

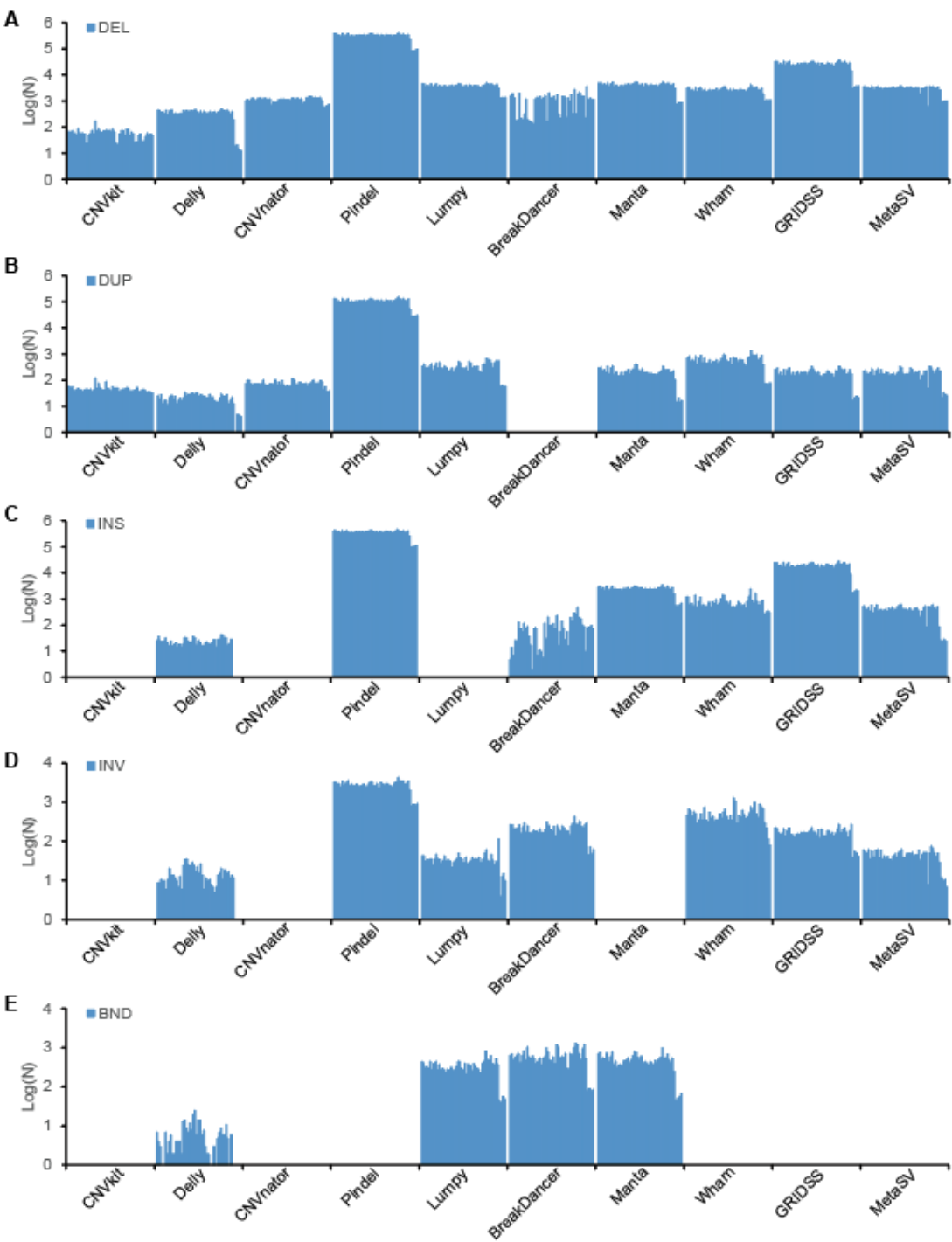


24

25 **Figure S1. Schematic diagram of SV types measured.** (A) An ~17.7 kb deletion is involved in
26 feathered legs in the chicken. (B) An ~4.2 kb insertion is reported to cause blue eggshell in chickens.
27 (C) An ~15.9 kb tandem duplication is associated with hyperpigmentation in chickens. (D)
28 interspersed duplications and inversions are reported to cause hyperpigmentation in chickens. (E)
29 The complex combination of tandem duplication (C) and interspersed duplication and inversion (D)
30 is involved in hyperpigmentation in chickens. Segment lengths are not proportional.

31

32



33

34 **Figure S2. The number of each type of SV detected by various tools.** Comparison of the number
35 of DELs (deletions) (A), DUPs (duplications) (B), INSs (insertions) (C), INVs (inversions) (D), and
36 BNDs (translocations) (E) detected by different tools. The number (N) in the axis of ordinates was
37 log transformed. Each bar represents a sample.

38

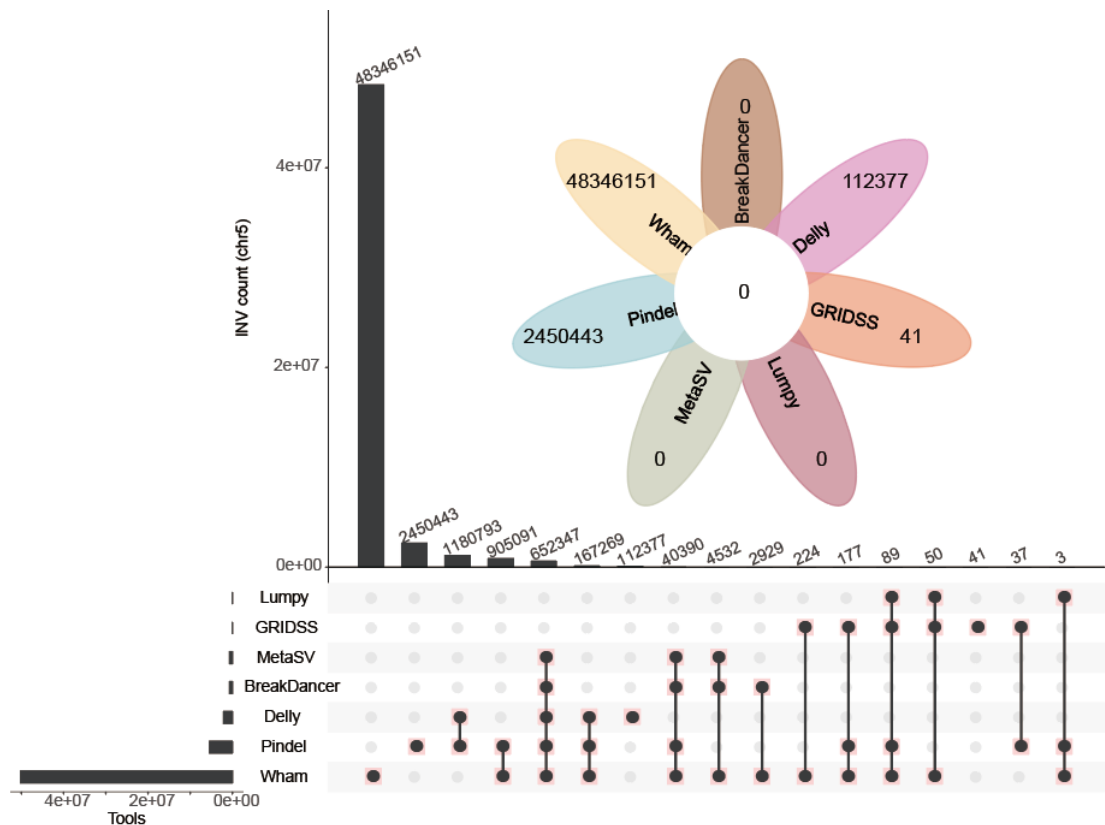


Figure S4. Comparison of the SVs detected by different tools. The Venn diagram shows the distribution of INVs (inversions) detected by different tools. Chromosome 5 of the selected sample (sample33) was tested, and $50 \text{ bp} < \text{SVs} < 1 \text{ Mb}$ were measured.

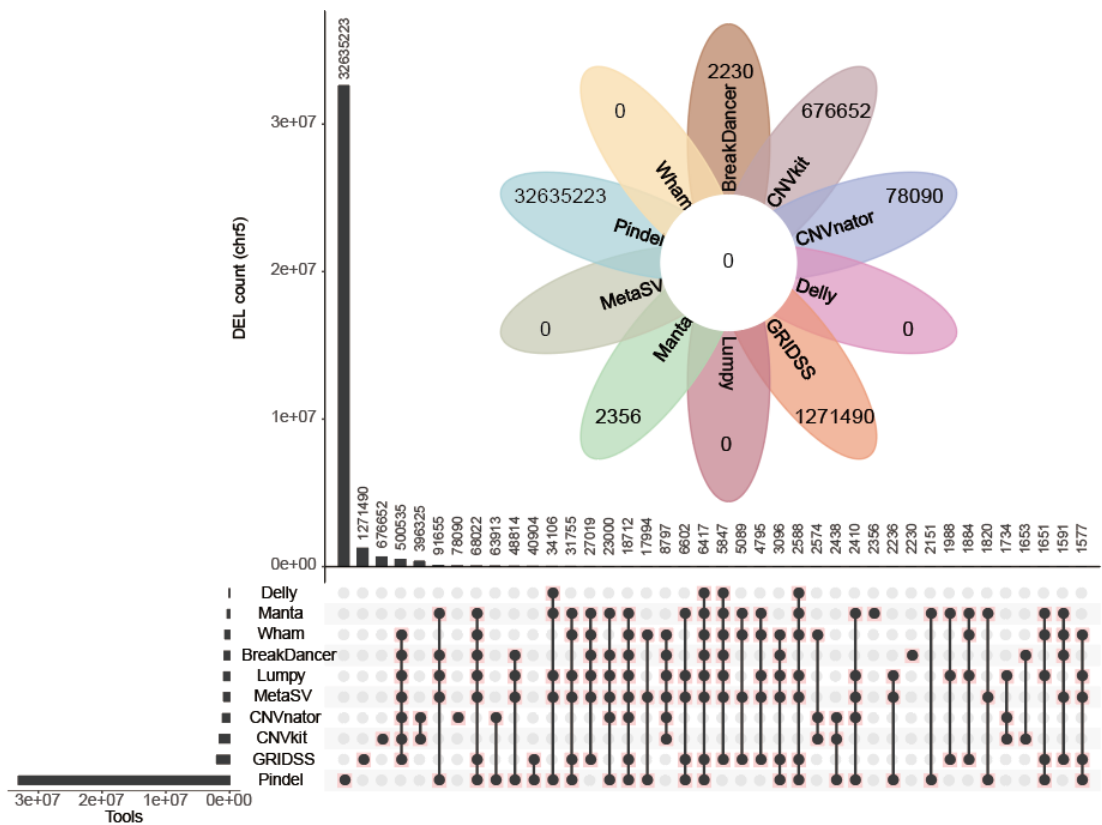


Figure S6. Comparison of the SVs detected by different tools. The Venn diagram shows the distribution of DELs (deletions) detected by different tools. Chromosome 5 of the selected sample (sample33) was tested, and 50 bp < SVs < 1 Mb were measured.

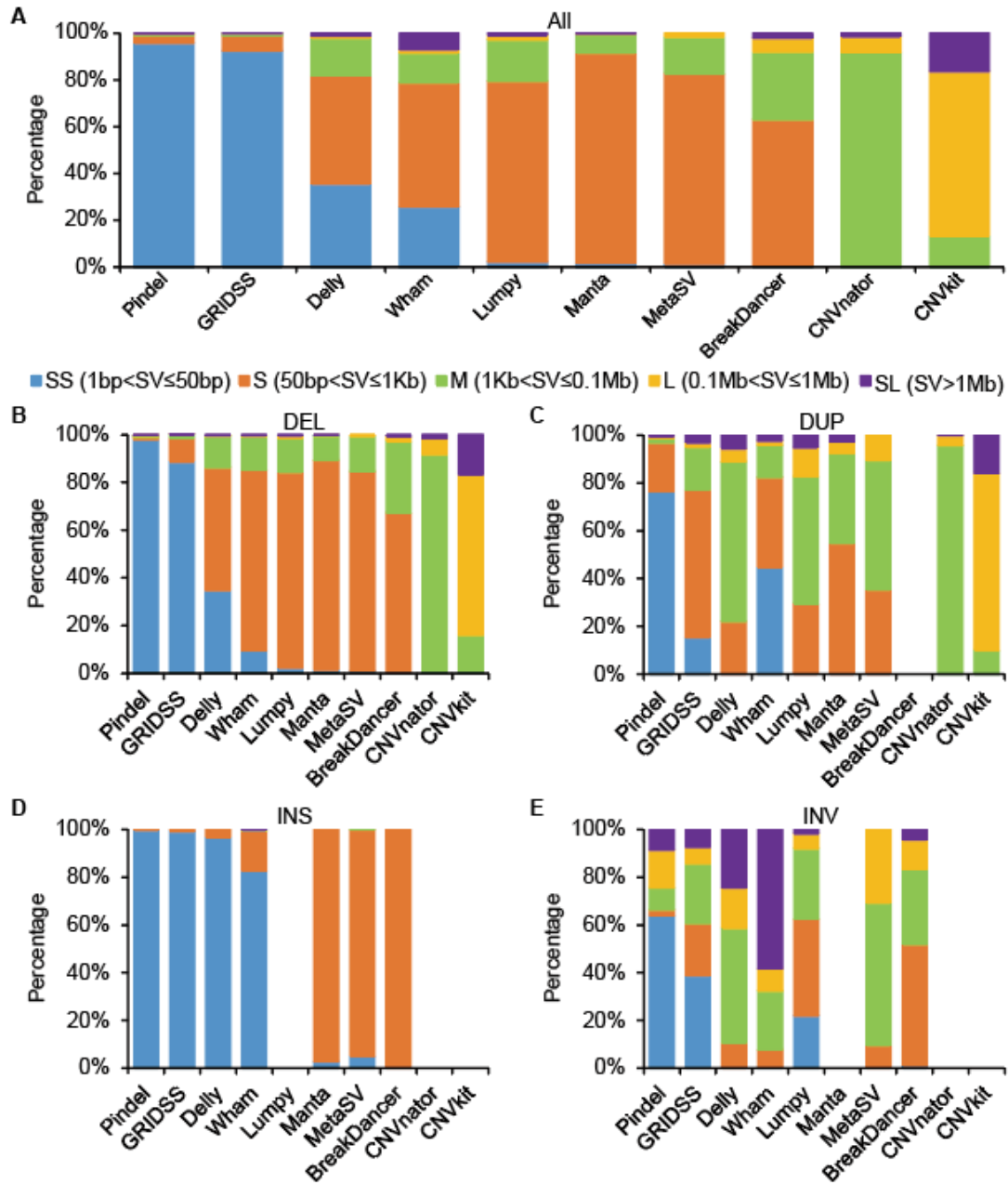


Figure S7. The proportion of different sizes and types of SVs. (A) The proportion of different sizes of all SVs detected by various tools. (B) The proportion of different sizes of DELs detected by various tools. (C) The proportion of different sizes of DUPs detected by various tools. (D) The proportion of different sizes of INSs detected by various tools. (E) The proportion of different sizes of INVns detected by various tools.

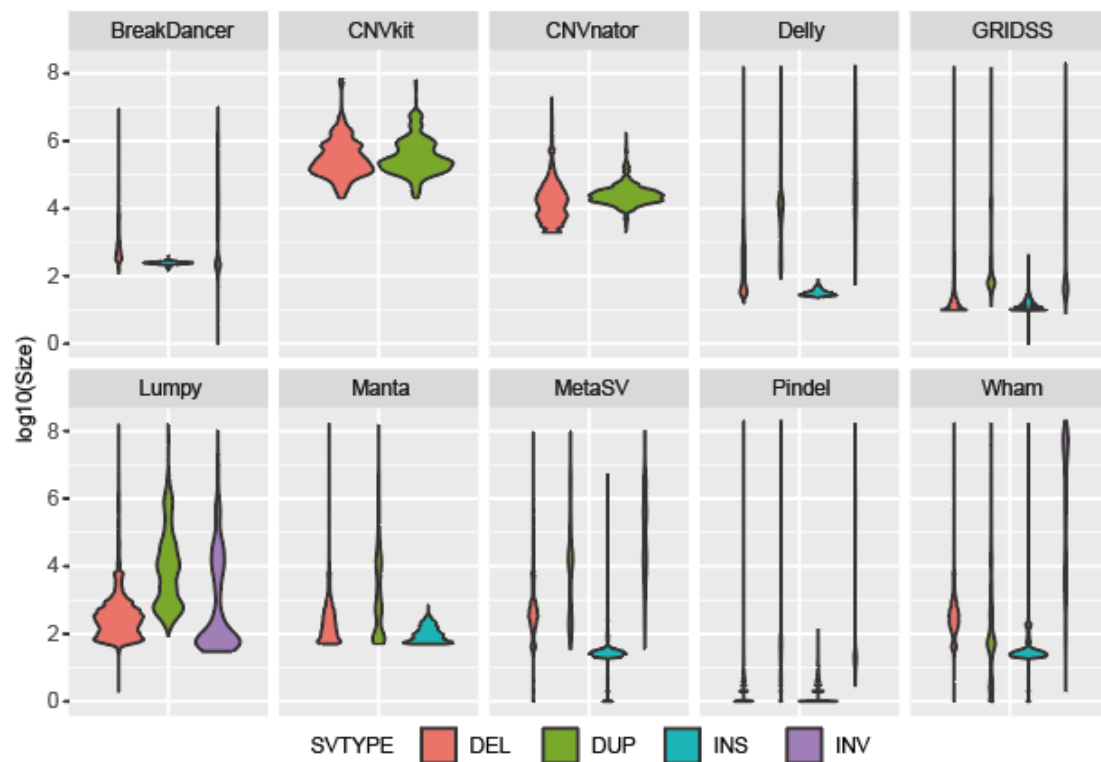


Figure S8. The size distribution of different types of SVs detected by various tools. The sizes of SVs in the axis of ordinates were log transformed. Different colors indicate different types of SVs. The shape of the violin diagram shows the distribution of SVs of different sizes.

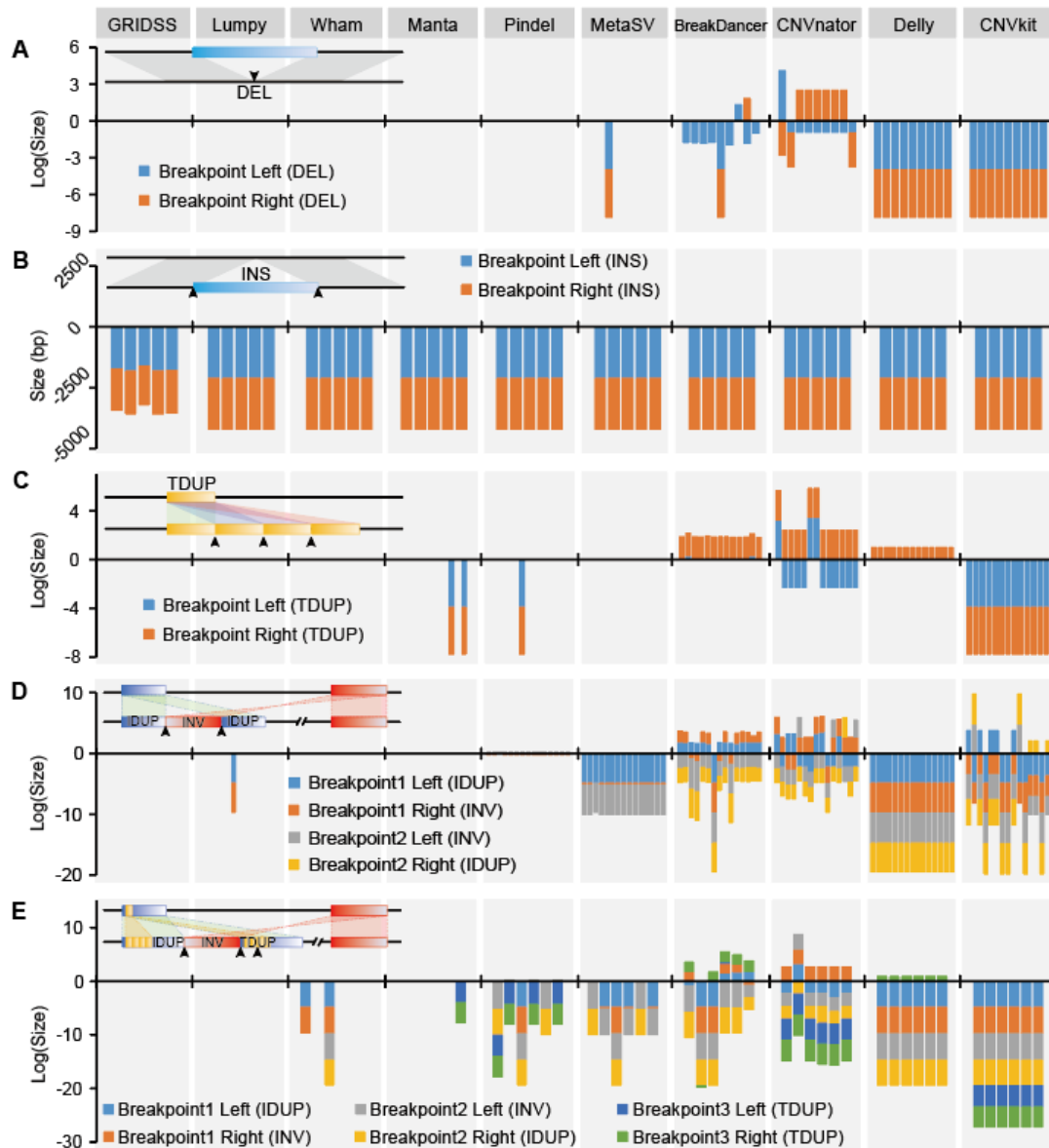


Figure S9. The breakpoint detection ability of various tools detects different SV types. (A) Breakpoint detection on DEL. (B) Breakpoint detection on INS. (C) Breakpoint detection on TDUP. (D) Breakpoint detection on IDUP and INV. (E) Breakpoint detection on IDUP+INV+TDUP. The bar plotted above the zero line (+) indicates that the detected SV is longer than the true size of the target SV type, while the bar plotted below the zero line (-) indicates that the detected SV is shorter than the true size of the target SV type. The schematic diagram on the top right represents the detected SV type, and the black arrow indicates the breakpoint.

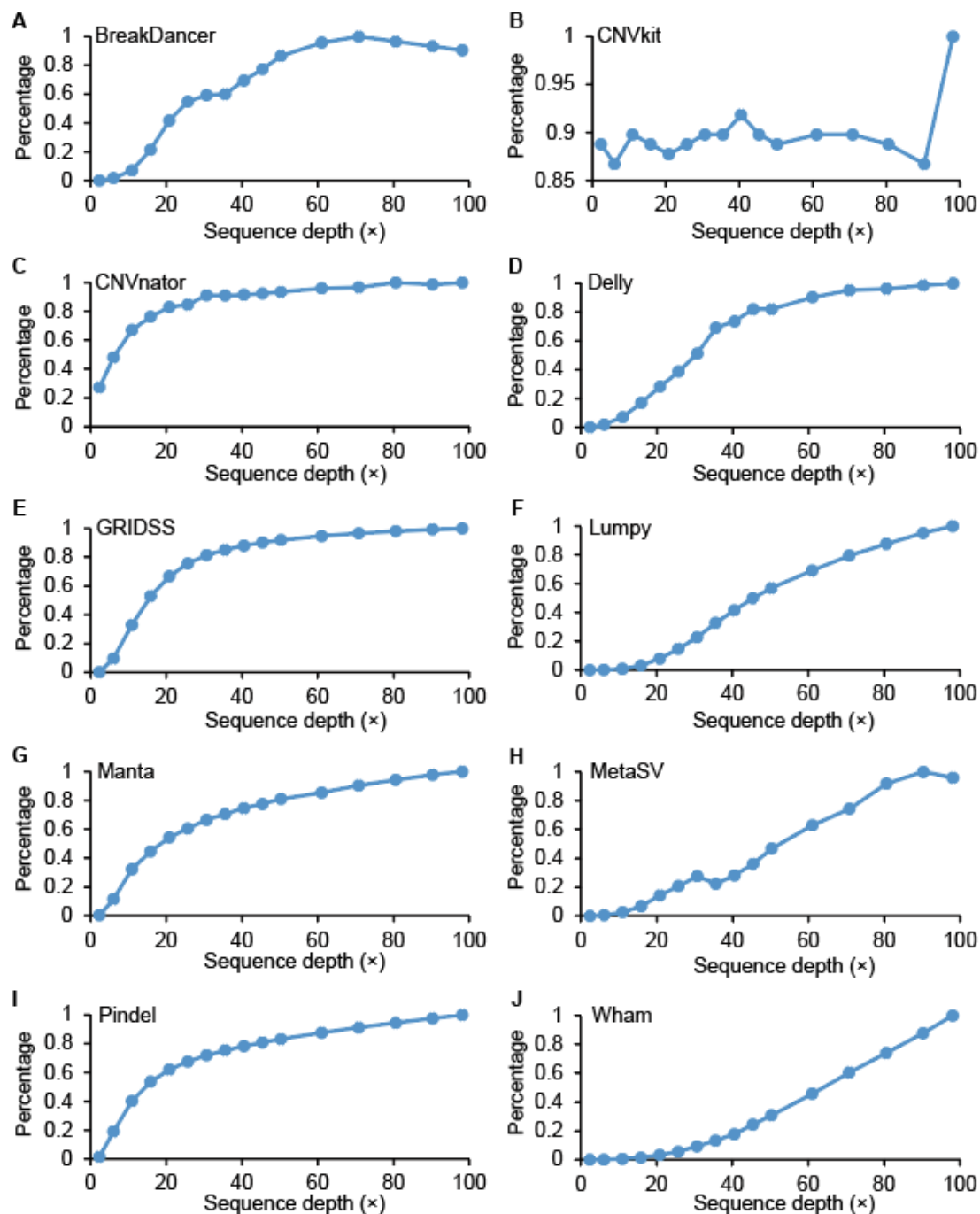


Figure S10. The detection rate on different data sizes of various tools. (A) BreakDancer; (B) CNVkit; (C) CNVnator; (D) Delly; (E) GRIDSS; (F) Lumpy; (G) Manta; (H) MetaSV; (I) Pindel; (J) Wham.

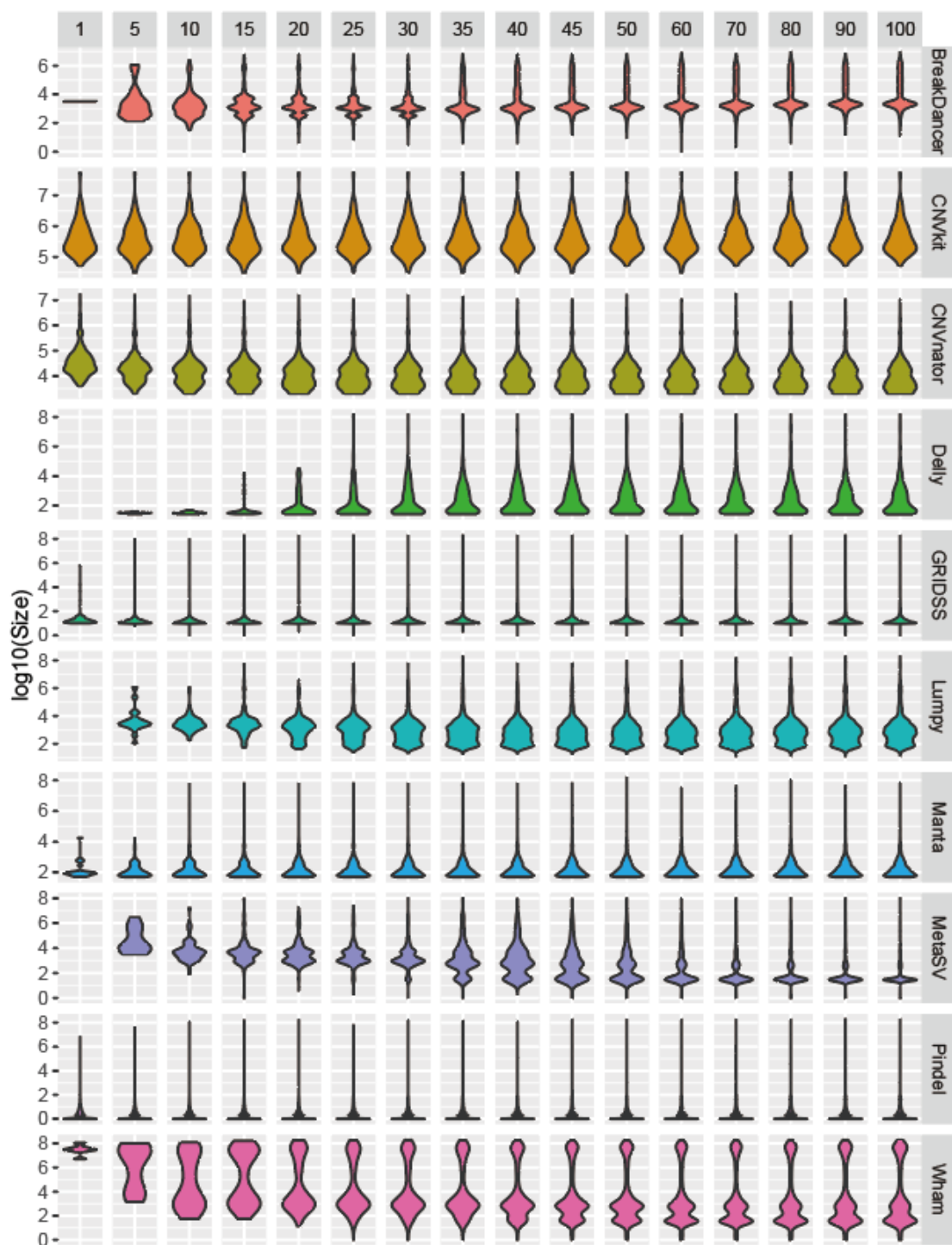


Figure S11. The size distribution of detected SVs on different data sizes. The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different tools.

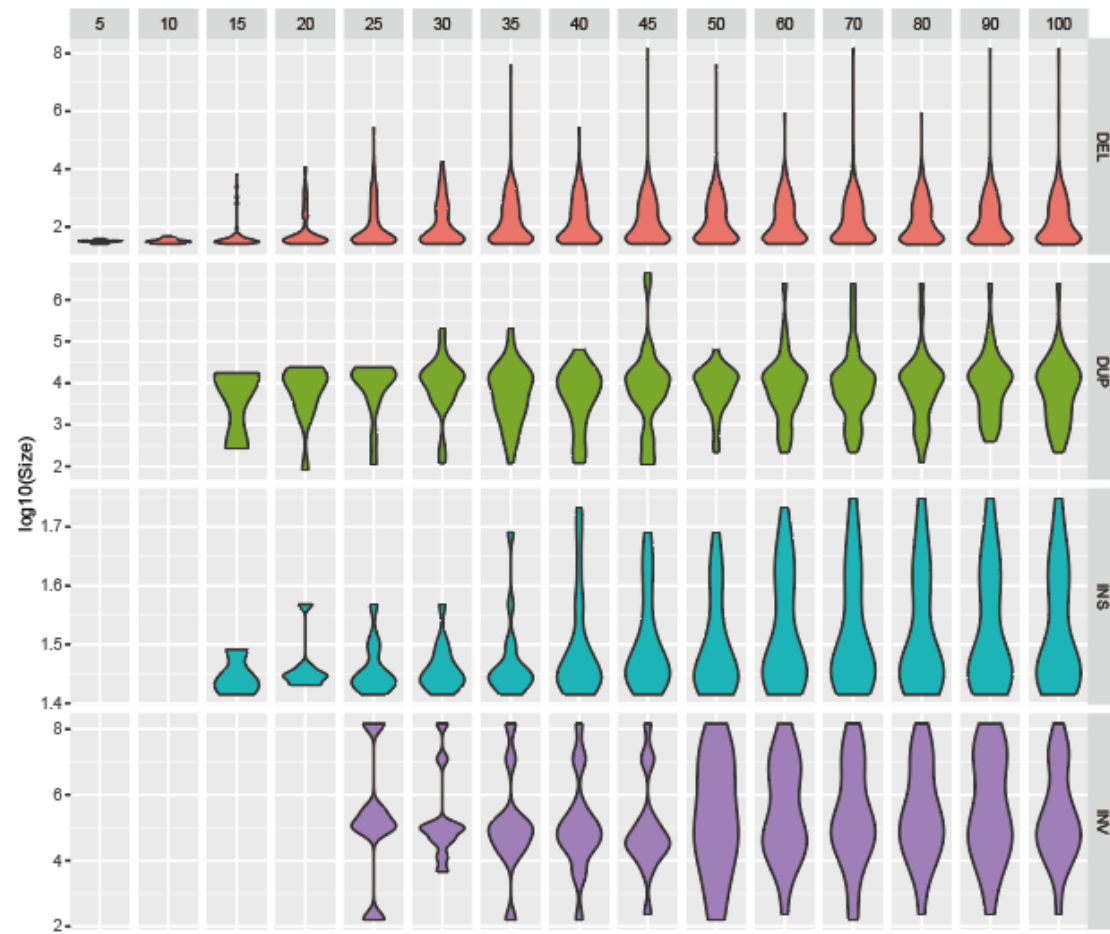


Figure S12. The size distribution of detected SVs on different data sizes detected by Delly. The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

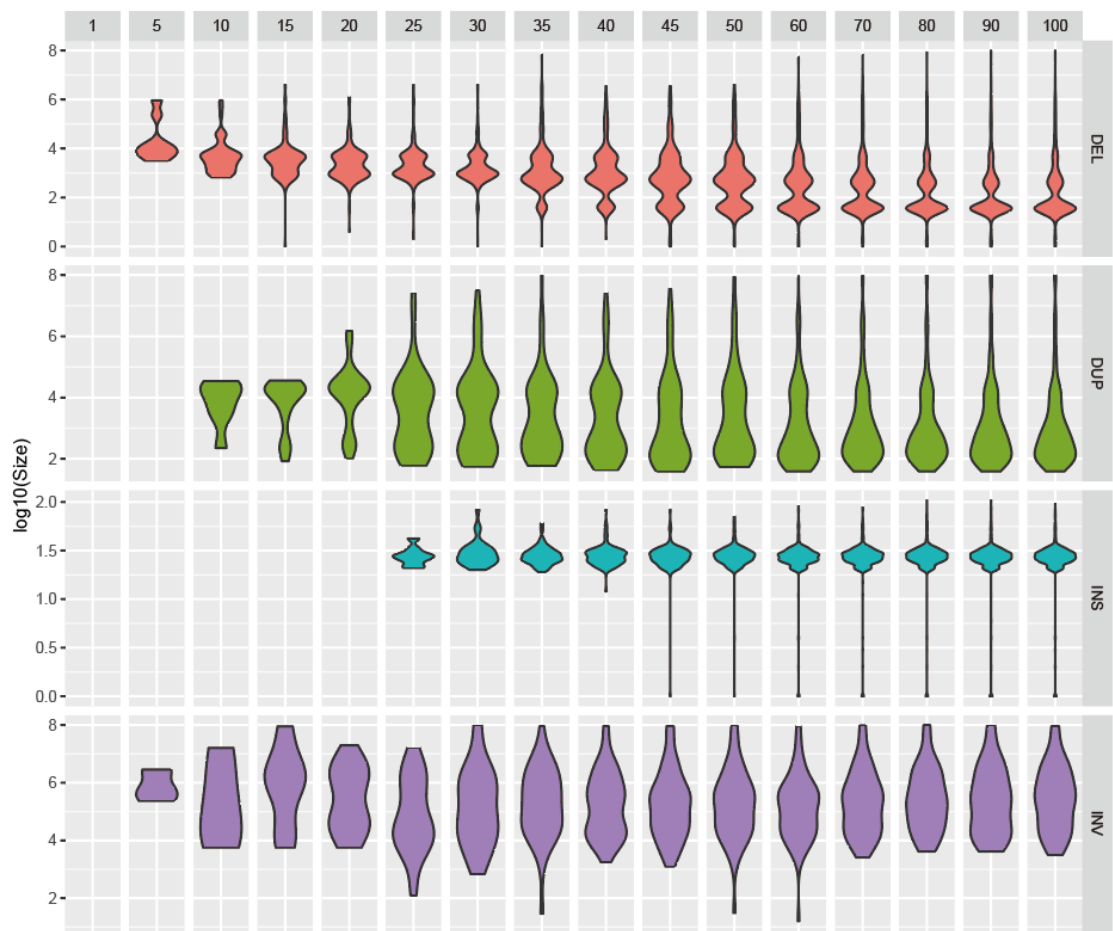


Figure S13. The size distribution of detected SVs on different data sizes detected by MetaSV.
The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

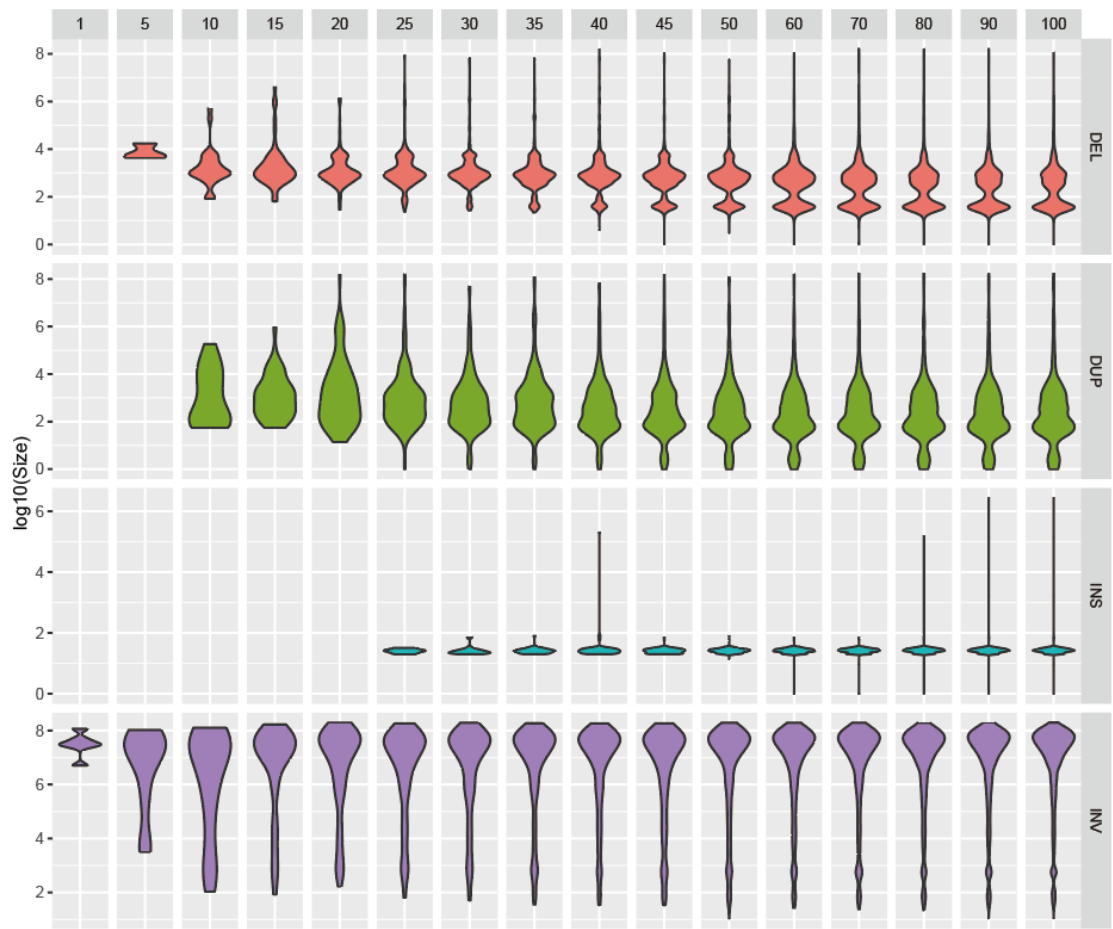


Figure S14. The size distribution of detected SVs on different data sizes detected by Wham. The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

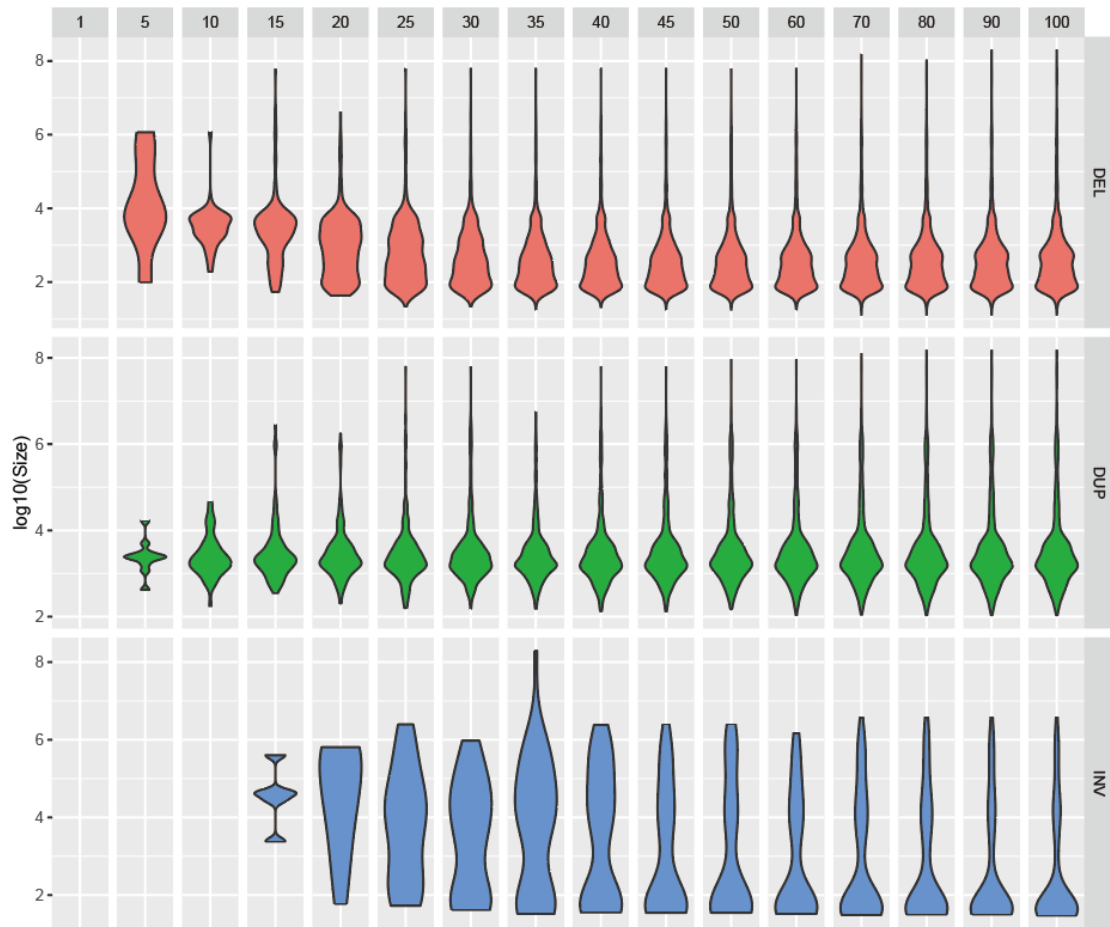


Figure S15. The size distribution of detected SVs on different data sizes detected by Lumpy.
 The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

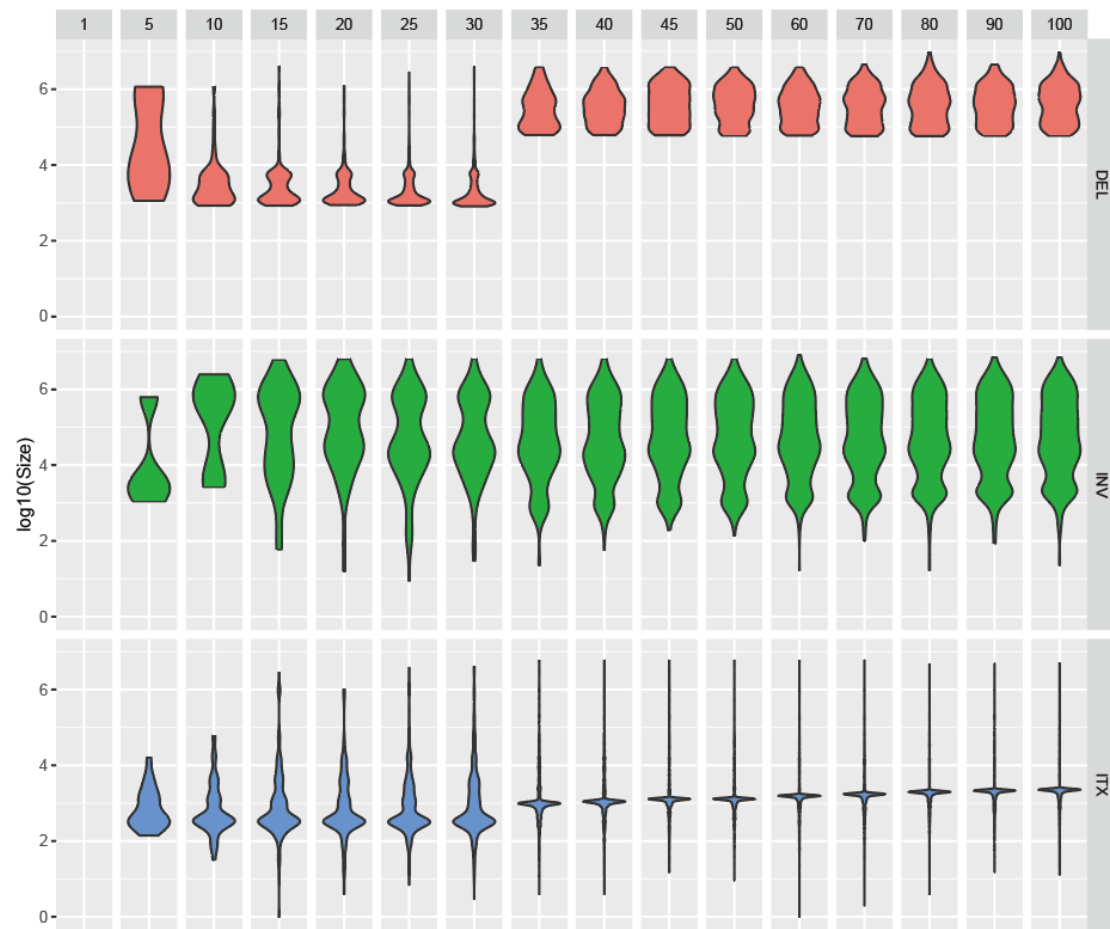


Figure S16. The size distribution of detected SVs on different data sizes detected by BreakDancer. The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

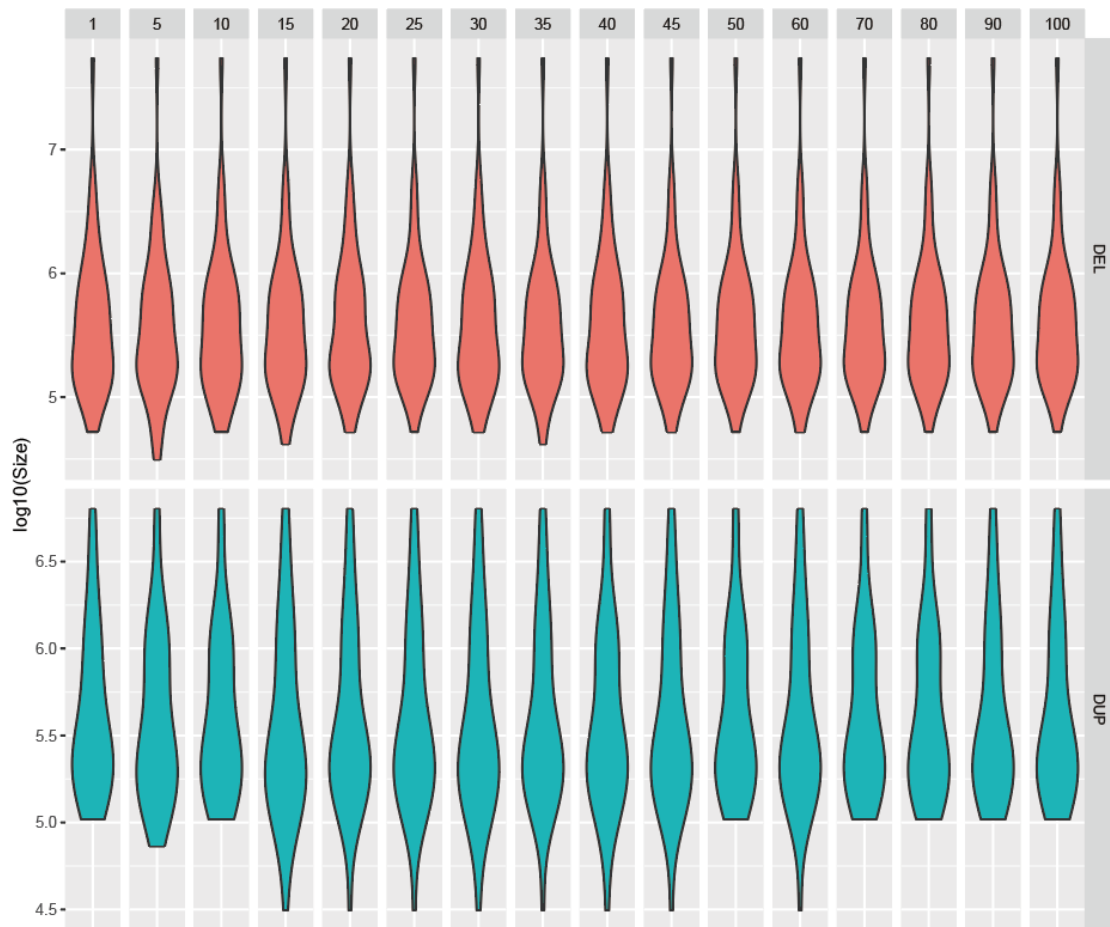


Figure S17. The size distribution of detected SVs on different data sizes detected by CNVkit. The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

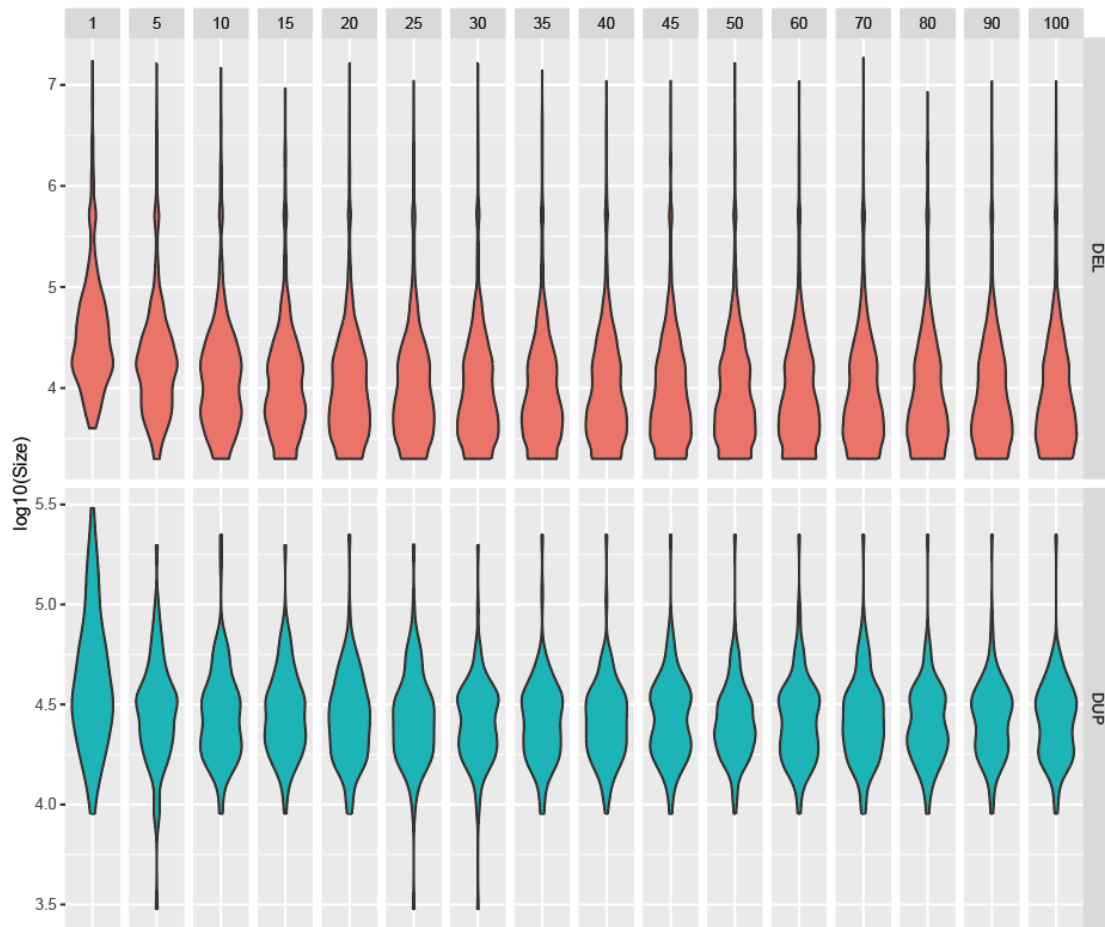


Figure S18. The size distribution of detected SVs on different data sizes detected by CNVnator.
 The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

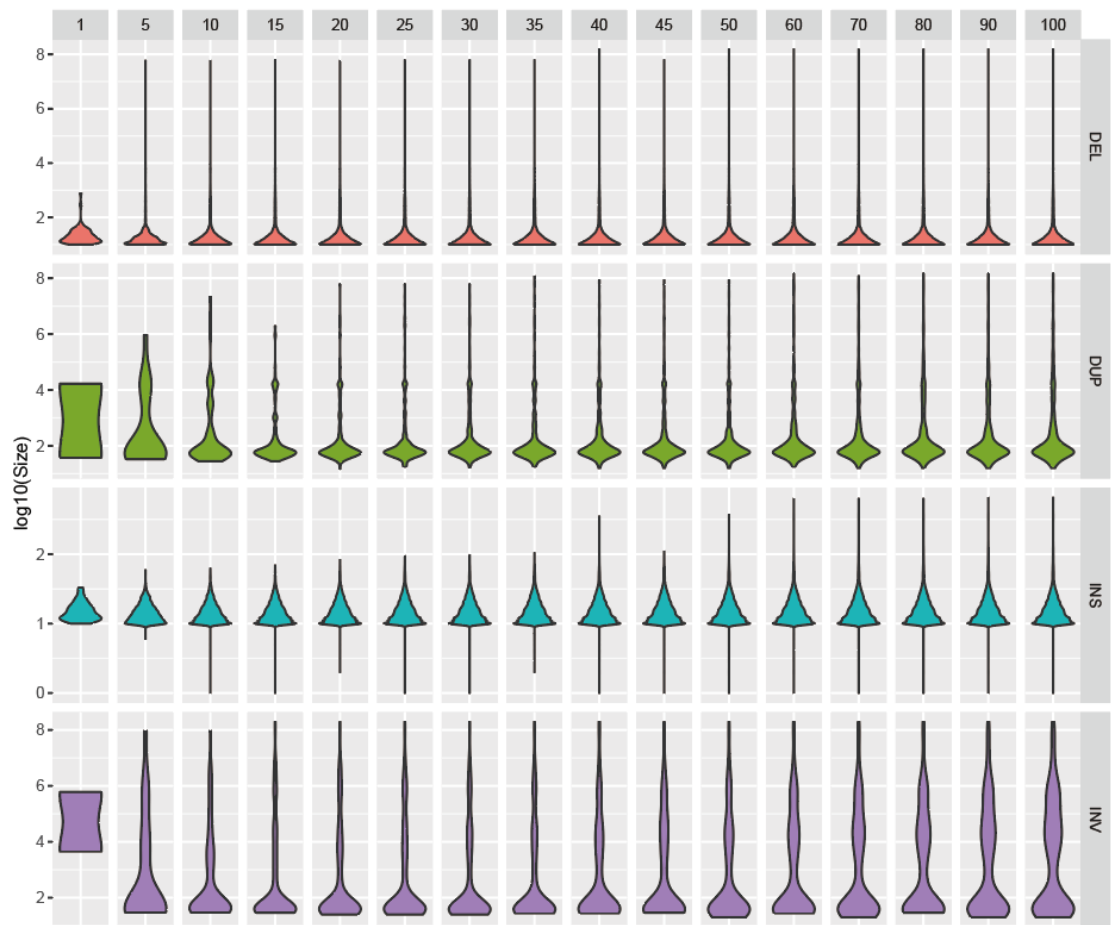


Figure S19. The size distribution of detected SVs on different data sizes detected by GRIDSS. The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

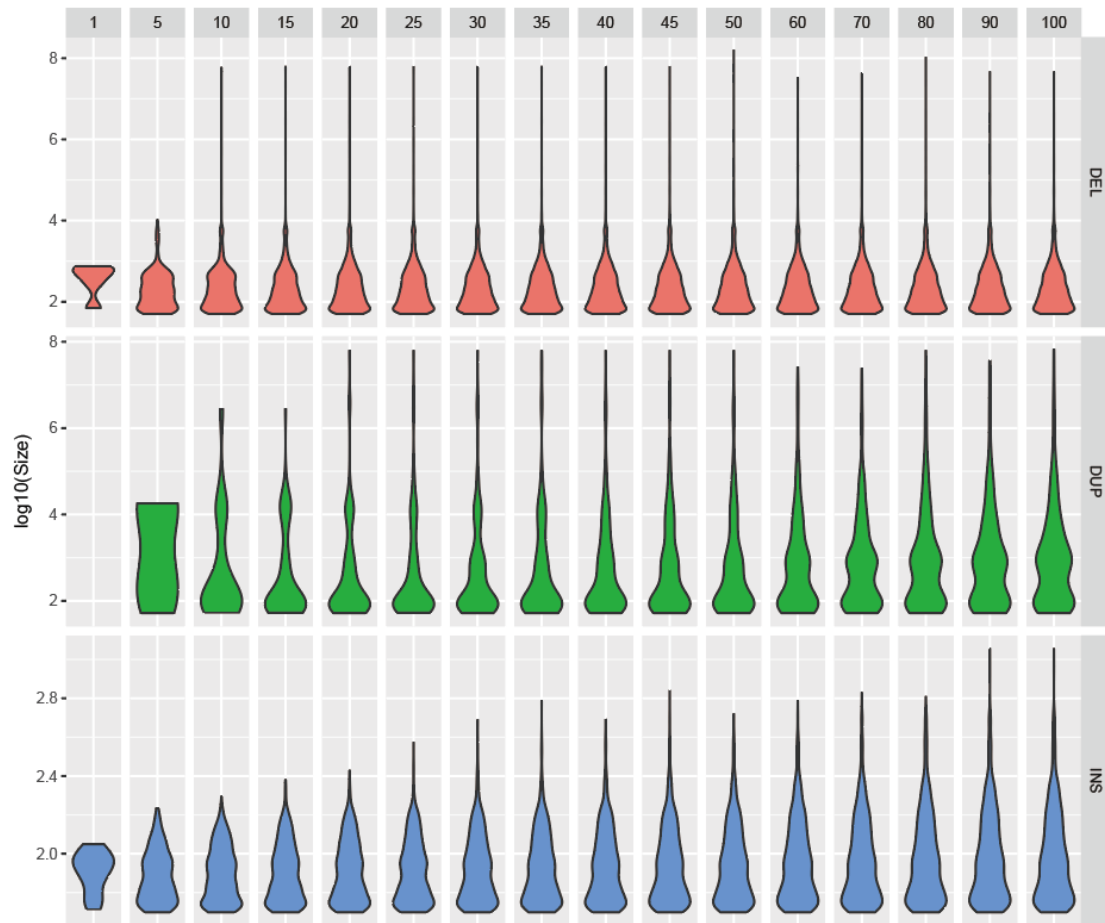


Figure S20. The size distribution of detected SVs on different data sizes detected by Manta.
 The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

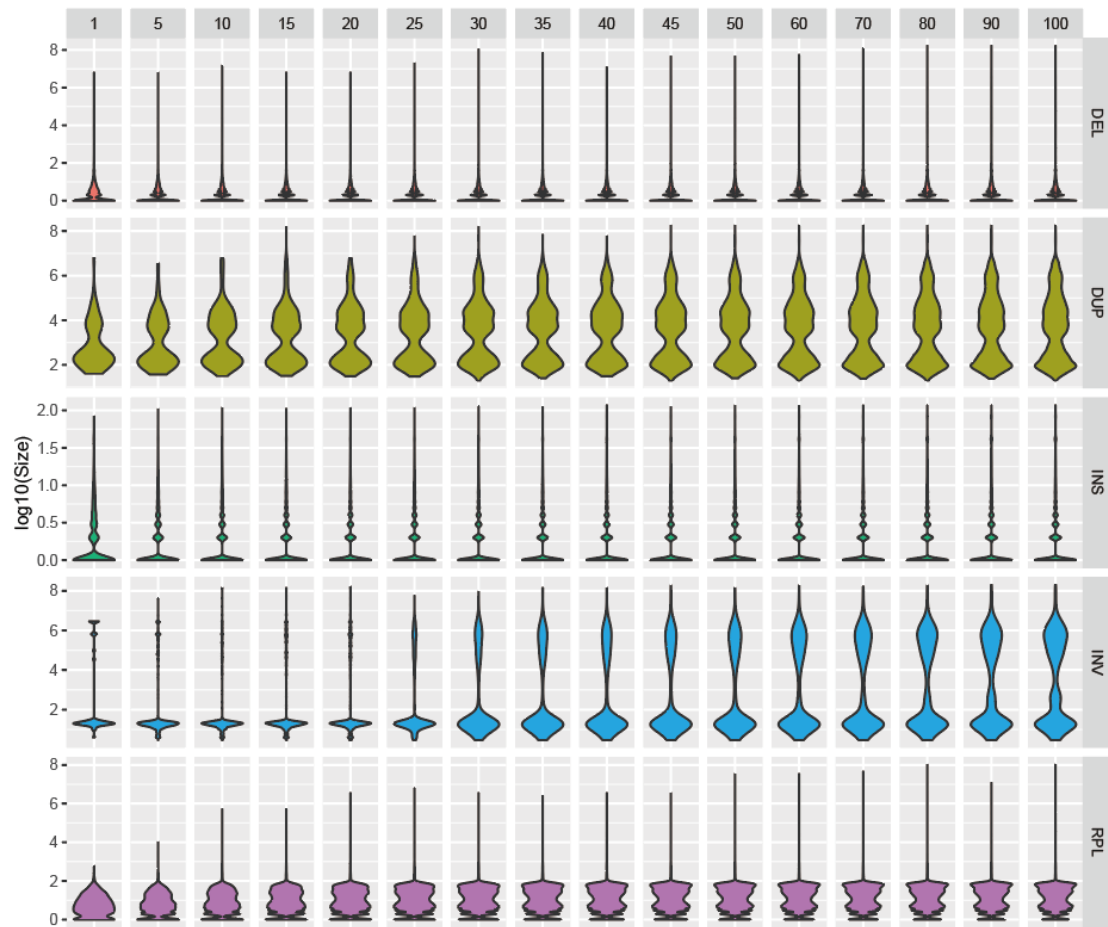


Figure S21. The size distribution of detected SVs on different data sizes detected by Pindel.
 The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

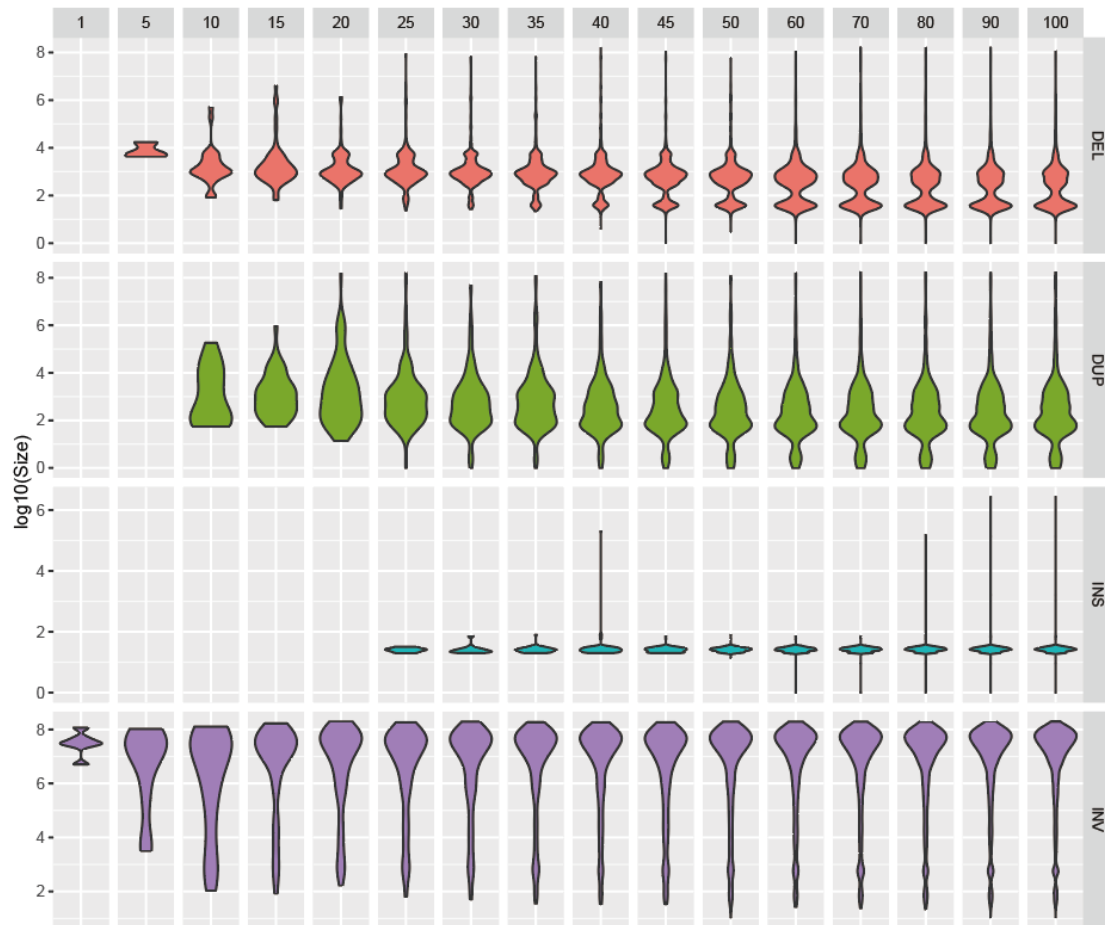


Figure S22. The size distribution of detected SVs on different data sizes detected by Wham.
 The numbers on the left axis of the ordinates represent the sizes of SVs detected on different data sizes by various tools, and the number was log transformed. The number in the abscissa axis represents different sequence depths. Different colors indicate different types of SV.

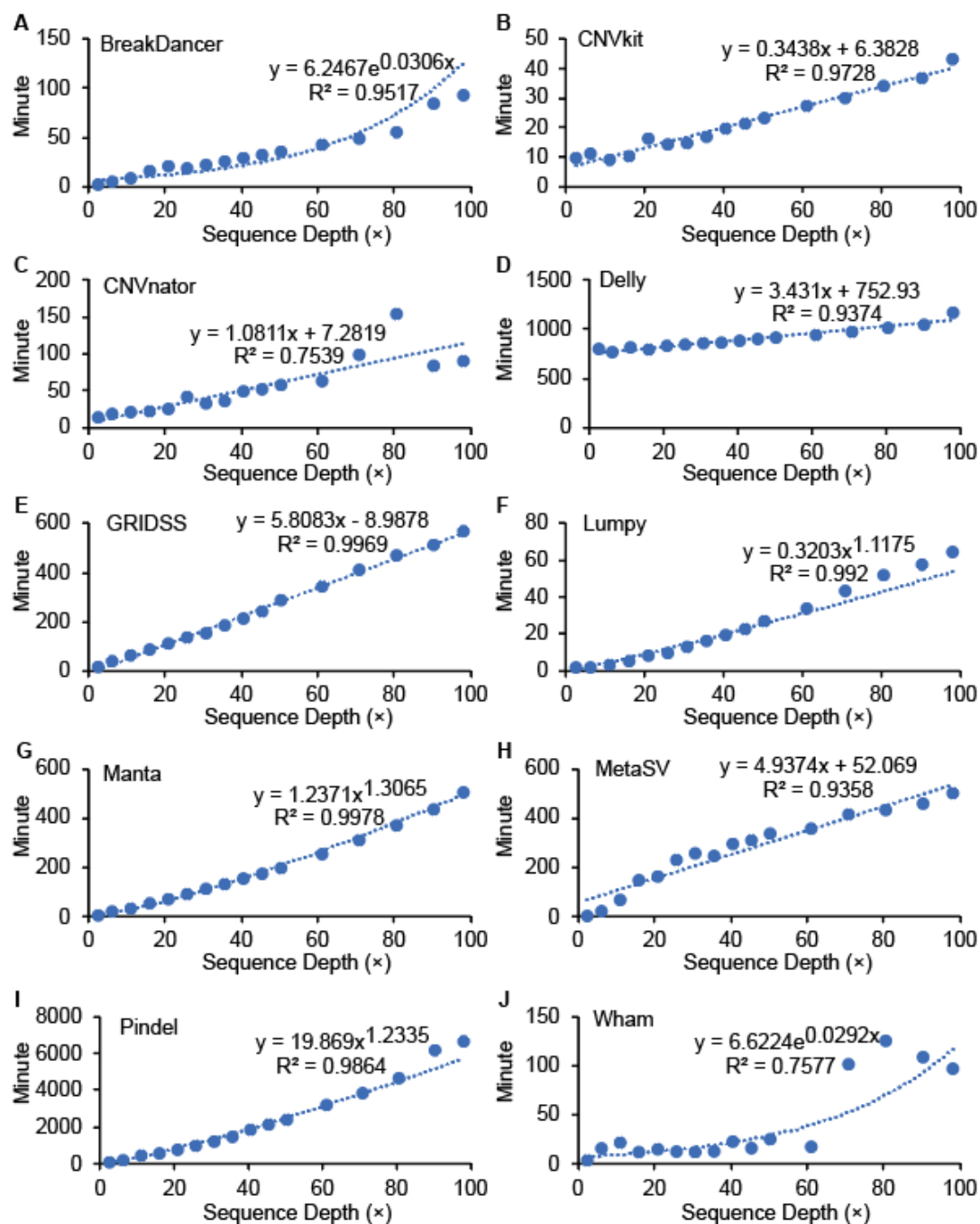


Figure S23. The CPU running time on different data sizes of various tools. (A) BreakDancer; (B) CNVkit; (C) CNVnator; (D) Delly; (E) GRIDSS; (F) Lumpy; (G) Manta; (H) MetaSV; (I) Pindel; (J) Wham.

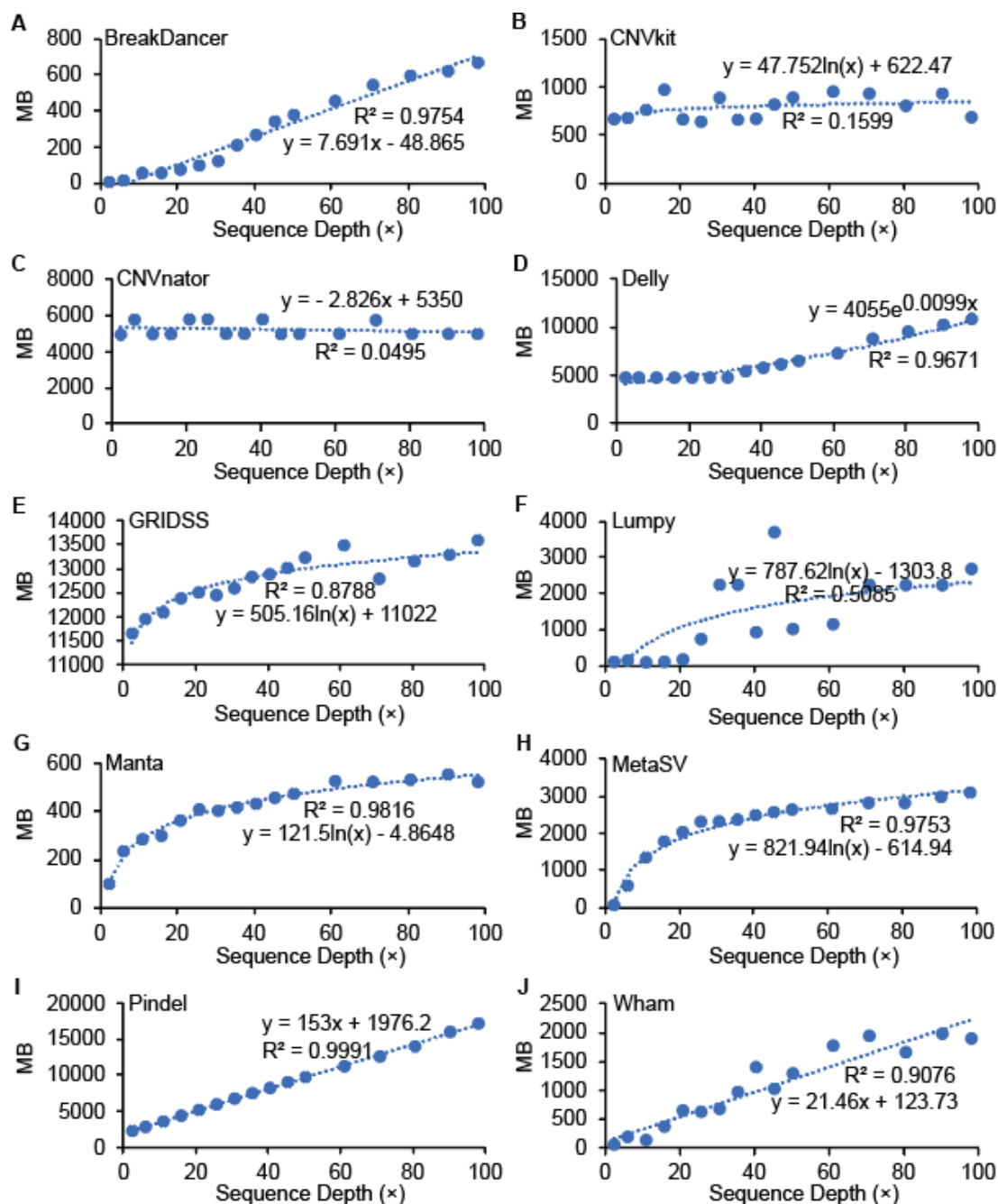


Figure S24. The maximum memory consumption on different data sizes of various tools. (A) BreakDancer; (B) CNVkit; (C) CNVnator; (D) Delly; (E) GRIDSS; (F) Lumpy; (G) Manta; (H) MetaSV; (I) Pindel; (J) Wham.