BMC Genomics

RESEARCH

Open Access

Performance evaluation of structural variation detection using DNBSEQ whole-genome sequencing

Junhua Rao[1,2†], Huijuan Luo[2†], Dan An[1,2], Xinming Liang[1,2], Lihua Peng[2*] and Fang Chen[1,2*]

Abstract

Background DNBSEQ platforms have been widely used for variation detection, including single-nucleotide variants (SNVs) and short insertions and deletions (INDELs), which is comparable to Illumina. However, the performance and even characteristics of structural variations (SVs) detection using DNBSEQ platforms are still unclear.

Results In this study, we assessed the detection of SVs using 40 tools on eight DNBSEQ whole-genome sequencing (WGS) datasets and two Illumina WGS datasets of NA12878. Our findings confirmed that the performance of SVs detection using the same tool on DNBSEQ and Illumina datasets was highly consistent, with correlations greater than 0.80 on metrics of number, size, precision and sensitivity, respectively. Furthermore, we constructed a "DNBSEQ" SV set (4,785 SVs) from the DNBSEQ datasets and an "Illumina" SV set (6,797 SVs) from the Illumina datasets. We found that these two SV sets were highly consistent of SV sites and genomic characteristics, including repetitive regions, GC distribution, difficult-to-sequence regions, and gene features, indicating the robustness of our comparative analysis and highlights the value of both platforms in understanding the genomic context of SVs.

Conclusions Our study systematically analyzed and characterized germline SVs detected on WGS datasets sequenced from DNBSEQ platforms, providing a benchmark resource for further studies of SVs using DNBSEQ platforms.

Keywords Structural variation (SV), Whole-genome sequencing (WGS), DNBSEQ

Introduction

Structural variation (SV) is a general term for different types of genomic mutations with size large than 50 bp, including deletion (DEL), insertion (INS), duplication (DUP), inversion (INV) and translocation (TRA) [1]. DELs and DUPs are also classified as copy-number variations (CNVs) [1]. SVs differ from small variants, such as single-nucleotide variants (SNVs) and short insertions and deletions (INDELs), in size and formation mechanisms [2]. SVs significantly contribute to the diversity found within human populations and have a notable impact on human health and disease [3–7]. In recent years, the importance and mechanism of SVs in human populations and diseases has been further explored and confirmed through large-cohort studies [8, 9]. The 1000 Genomes Project (1KGP) analyzed SVs of 2,504 individuals and estimated that SVs were ~ 50-fold enriched for expression quantitative trait loci (eQTLs) compared with SNVs [6]. The Human Genome Structural Variation Consortium (HGSVC) identified 107,590 SVs with 278 SV hotspots in human genome and explored the contribution of SVs in population adaptive selection of humans

†Junhua Rao and Huijuan Luo contributed equally to this work.

*Correspondence:
Lihua Peng
penglihua@genomics.cn
Fang Chen
fangchen@mgi-tech.com
[1] MGI Tech, Shenzhen 518083, China
[2] BGI, Shenzhen 518083, China

[10]. Collins et al. constructed 163 genome-wide dosage sensitive segments of rare SVs across 54 disorders for human disease searches [11].

Currently, several large-cohort studies investigating human disease or human population described above are based on short-read whole-genome sequencing (WGS) technology [6, 8, 9, 12]. This is attributed to the development and utilization of massively parallel sequencing (MPS) platforms that generate short-read data, along with the advancements in analytical tools for detecting SVs. Dozens of tools that can use short-read (50–150 bp) WGS data sequenced on MPS platforms to detect SVs on a genome-wide scale [13]. Each SV tool is based on one of the following five algorithms: (1) read depth (RD), (2) read pair (RP), (3) split read (SR), (4) de novo assembly (AS) and (5) combination of approaches (CA). Therefore, the types of SVs detected by each tool may different. For instance, tools based on RD algorithm can only detected DELs and DUPs, such as CNVnator [14]. Some tools are specifically designed to detect only specific types of SVs, such as BASIL-ANISE [15] for INSs and Sprites [16] for DELs. Meanwhile, the sensitivities of SVs were reported to fluctuate in the range of 10%−70% depending on the size and type of SVs, while the false-positive rates were up to 89% [17]. Despite these limitations, SV detection on short-read data sequenced on MPS platforms is still a good approach for SV research due to variables such as cost, time, resolution and project scope [6, 9]. In short, the SV tools described above were mainly designed to detect SVs based on datasets from Illumina platforms, and the performance of SVs detected by these SV tools has been reported in many articles [13, 18, 19].

As is known to all, Illumina platforms, such as HiSeq 2500 and NovaSeq6000, are the main MPS platforms widely used in research of SVs [12]. For example, leveraging ~ 30X WGS data generated by the NovaSeq6000 system, researchers have broadened the spectrum of genomic variants, including the SV catalog, for the 1KGP [6, 9]. Since 2015, the DNBSEQ sequencing platforms, based on the technologies of DNA nanoballs (DNBs) and combinatorial Probe-Anchor Synthesis (cPAS), has been widely utilized in genomic researches for its high sequencing accuracy, low duplication rates, and reduced index hopping [20]. To date, the DNBSEQ platforms have been used to carry out many important genomic studies about plant, animal, human health and disease [21–27]. For example, recently, Jin et al. utilized the DNBSEQ platform to enhance our understanding of diseases and phenotypic variations during pregnancy in Asian populations [28]. Since the DNBSEQ and Illumina platforms are both widely used MPS platforms, researchers are concerned about the consistency and interchangeability of genomic variant detection performance between these two platforms. Of which, the performance of SNVs, INDELs and CNVs based on DNBSEQ platforms has been studied and verified to be consistent with those based on Illumina platforms [29–31]. Specially, in our prior research, we employed five different algorithms to evaluate the CNV detection capabilities of data derived from the DNBSEQ and Illumina platforms [29]. Our findings indicated that the CNVs identified by both platforms were similar in terms of size, number, sensitivity and precision. Notably, the DNBSEQ platform demonstrated a superior performance in detecting smaller CNVs. However, the comprehensive characteristics of SVs, especially INSs and INVs, identified by DNBSEQ platform remained elusive. To address this, our study embarked on an extensive SV detection analysis, applying 40 different software tools to WGS data generated by the DNBSEQ and Illumina platforms for the first time. We meticulously examined their sequencing and genomic attributes to gain a deeper understanding of these variants.

Results

Similar SV detection performance between DNBSEQ and Illumina platforms

In our study, we introduced ten WGS datasets of the germline sample NA12878 from public databases [29] for SV detection, analyzing each dataset with 40 different tools that encompass all five algorithm types as described (Fig. 1, Additional file 2: Table S1, see Supplementary "Methods" for details). Eight datasets were sequenced on two DNBSEQ platforms (BGISEQ-500 and DNBSEQ-G400) with an average depth of 31.43X, and two datasets were sequenced on two Illumina platforms (HiSeq2500 and NovaSeq6000) with an average depth of 30.61X (Additional file 2: Table S2). Finally, based on DNBSEQ platforms, we detected an average of 2,838 DELs using 32 tools, 1,490 DUPs using 21 tools, 1,117 INSs using 22 tools, 422 INVs using 16 tools, and 2,793 TRAs using eight tools across all eight datasets. These results were very similar to those obtained on the Illumina platforms, including an average of 2,676 DELs, 1,664 DUPs, 737 INSs, 239 INVs, and 2,878 TRAs (Fig. 2, Additional file 1: Fig. S1 and Additional file 2: Table S3).

We proceeded to assess the precision and sensitivity of SVs detected by various tools on the DNBSEQ and Illumina platforms. To facilitate a direct comparison with the findings reported by Shunichi et al. in 2019, we adopted their methodologies and benchmark of the NA12878 sample [13]. In this context, we calculated the precision and sensitivity for DELs, DUPs, INSs, and INVs. TRAs were excluded from the evaluation analysis due to the lack of TRA benchmark of NA12878 and the fact that TRAs are always false positive [32]. The average precision and sensitivity of DELs detected on Illumina

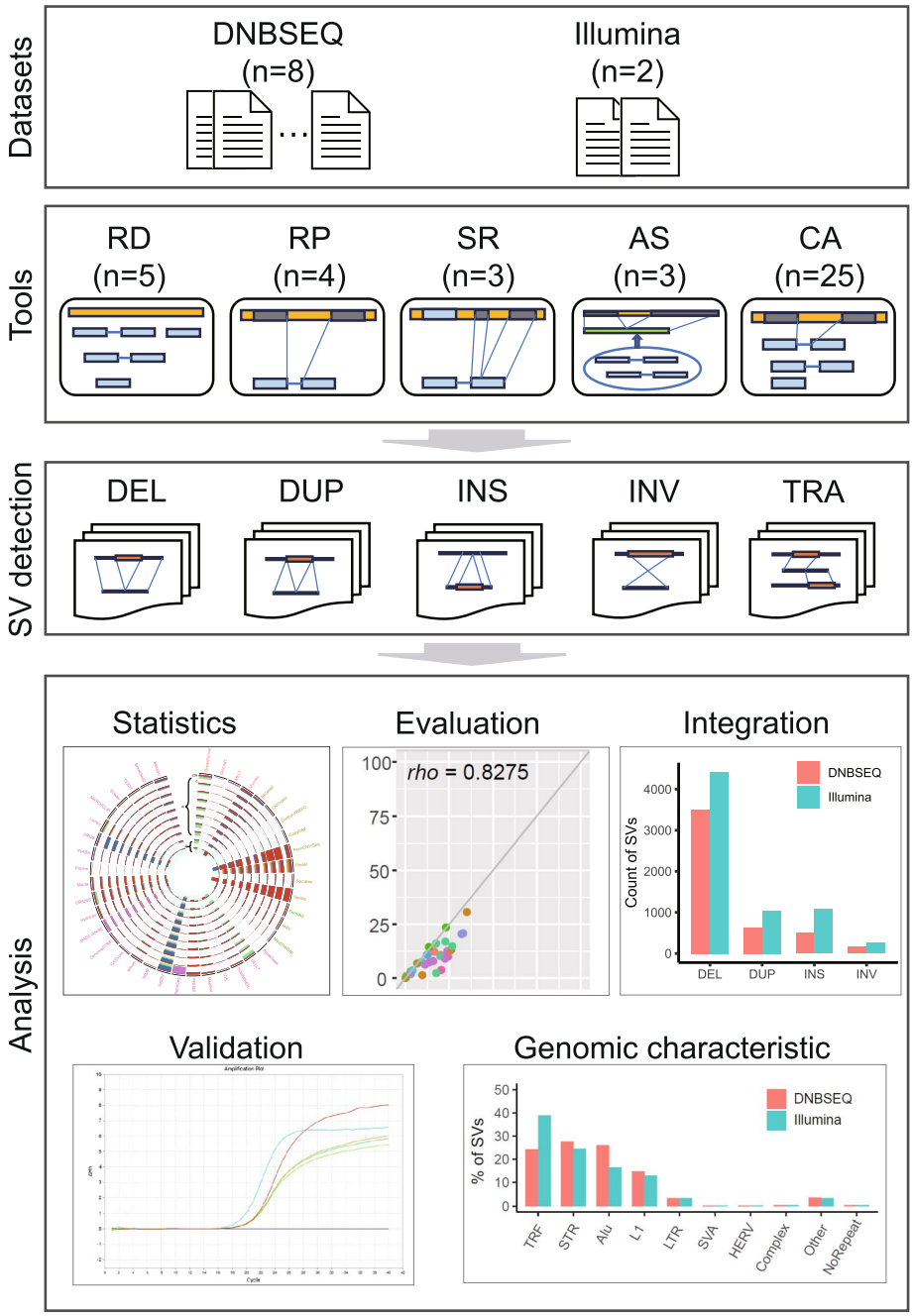

Fig. 1 The Overall Framework of SV Detection and Analysis in This Study. This study utilized ten datasets from both the DNBSEQ and Illumina platforms, along with 40 tools based on five distinct algorithms. SVs of five types were identified across the ten datasets using these 40 tools, respectively. Comprehensive analyses, including basic statistics, evaluation, integration, further validation, and genomic characteristic processing, were conducted based on the SV results obtained. RD—read depth; RP—read pair; SR—split read; AS—de novo assembly; CA—combination of approaches; DEL—deletion; DUP—duplication; INS—insertion; INV—inversion; TRA—translocation

datasets were 53.06% and 9.81%, respectively, while 19.86% and 5.52% of DUPs, 44.01% and 2.80% of INSs and 26.79% and 11.06% of INVs were detected, which is consistent with previous report [13] (Additional file 1: Fig. S2 and Additional file 2: Table S4). Analogously, the average precision and sensitivity of DELs detected on DNBSEQ datasets were 62.19% and 15.67%, respectively, 23.60% and 6.95% of DUPs, 43.98% and 3.17% of INSs and 25.22% and 11.58% of INVs (Additional file 1: Fig. S3 and Additional file 2: Table S4). In line with our

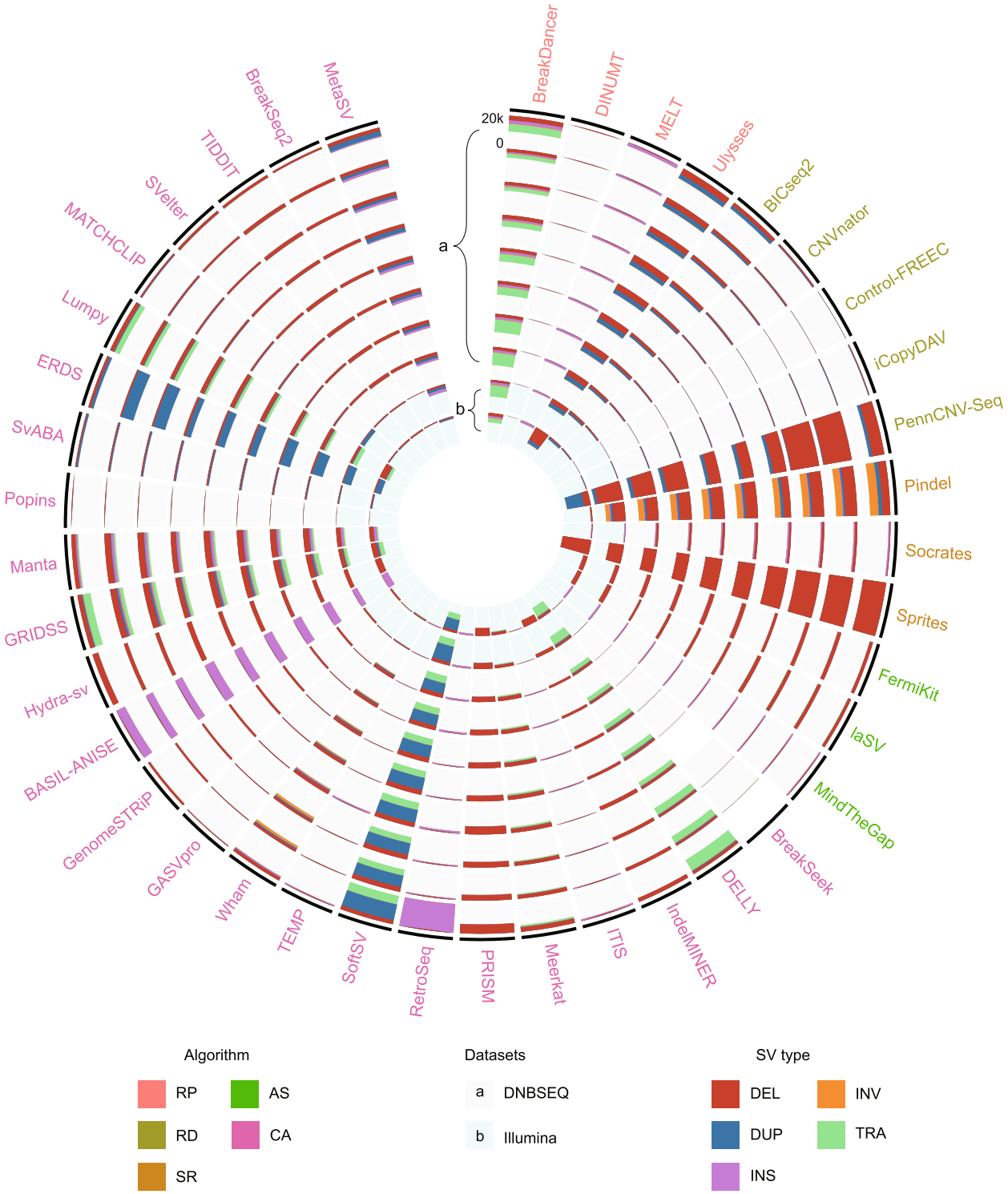

Fig. 2 Statistics of SVs Detected on Ten Datasets With 40 Tools. The circle graph visually represents the number of SVs detected using both DNBSEQ and Illumina datasets across various analytical tools. The outermost circle highlights the tools employed for SV detection, each distinguished by unique colors corresponding to their respective algorithms. The (**a**) and (**b**) circles depict the number of SVs identified in the DNBSEQ and Illumina datasets, respectively. The range of SV counts (y-axis) is consistent across the ten concentric circles, with the scale clearly labeled on the outer circle for reference. SV types are represented by distinct colors: DELs in red, DUPs in blue, INSs in purple, INVs in orange, and TRAs in green. RD—read depth; RP—read pair; SR—split read; AS—de novo assembly; CA—combination of approaches; DEL—deletion; DUP—duplication; INS—insertion; INV—inversion; TRA—translocation

prior results [29], the detection of DELs and DUPs on the DNBSEQ platform mirrored the performance observed on the Illumina platform. Meanwhile, the detection of INSs and INVs by the DNBSEQ platform showed a comparable level of performance to those identified using the Illumina platform.

To seek the consistency of genome-wide SV detection between DNBSEQ and Illumina platforms in deep, we compared the number, size, sensitivity and precision of SVs detected by the same tool on these two platforms. We found that both the number and size of various SVs were highly consistent between the DNBSEQ and Illumina platforms (Fig. 3a). Specifically, the consistency of the number and size of DELs were observed with Spearman's rank correlation coefficients of 0.88 and 0.97 (32 tools), respectively, 0.88 and 0.85 (21 tools) of DUPs, 0.95 and 0.92 (22 tools) of INSs, and 0.96 and 0.88 (16 tools) of INVs (Fig. 3a). Furthermore, the sensitivity and precision of SV detection were also highly consistent between the two platforms, with *rho* values of 0.83 and 0.91 for DELs (Spearman's rank correlation coefficient), 0.91 and 0.80 for DUPs, 0.96 and 0.97 for INSs, and 0.92 and 0.86 for INVs (Fig. 3b). However, the sensitivity and precision of DELs identified on the DNBSEQ platform (average 15.67% and 62.19%) were found to be marginally higher than those detected on the Illumina platform (9.81% and 53.06%, Fig. 3b, Additional file 1: Fig. S2 and Fig. S3). Overall, the DNBSEQ and Illumina platforms, both MPS platforms, showed similar SV detection performance, and the SV detection tools developed based on the Illumina platform dataset were also applicable to DNBSEQ platform dataset.

High validation rates of the integrated SV set from DNBSEQ and Illumina platforms

In our study, we built an SV set of DNBSEQ platforms (referred as "DNBSEQ" set) by integrating all SV results of eight DNBSEQ datasets detected by 40 different tools (Additional file 1: Fig. S4, see Supplementary "Methods" for details). "DNBSEQ" set is consisted of 4,785 SVs, including 3,499 DELs, 630 DUPs, 500 INSs and 156 INVs (Additional file 1: Fig. S5 and Additional file 2: Table S5). We also integrated SV results of two Illumina datasets and obtained 6,797 SVs, referred as "Illumina" set, including 4,424 DELs, 1,042 DUPs, 1,071 INSs and 260 INVs (Additional file 2: Table S6).

Presently, three research groups have compiled SVs collections for the NA12878 sample using distinct methodologies. Of which, Marta et al. built a collection of 8,236 SVs from the 1000 Genomes Project (referred as "1KGP" set), which was constructed using high-depth WGS data from the NovaSeq6000 [9]. Additionally, Jouni et al. created a collection of 11,089 SVs using the pan-genome approach tailored for short-read data (referred as "Giraffe" set) [33]. Peter et al. employed PacBio RSII long-read data to construct a set of 4,561 SVs for the Human Genome Structural Variation Consortium (referred as "HGSVC" set) [10]. Here, we compared the above-mentioned "DNBSEQ" and "Illumina" sets with "1KGP", "HGSVC" and "Giraffe" sets. We found that the "DNBSEQ", "Illumina", and "1KGP" sets had similar specific SV proportions of 33.38% (1,597/4,785), 51.63% (3,509/6,797) and 30.96% (2,550/8,236), respectively, likely due to their use of short-read MPS data (Additional file 1: Fig. S5 and Fig. S6). The "Giraffe" set had 66.37% (7,360/11,089) specific SVs, which may be attributed to improvements in pan-genome alignment and analysis methods applied to short-read MPS data, especially for INS detection [33]. The "HGSVC" set had 77.33% (3,527/4,561) specific SVs, likely because it uses long-read sequencing data, highlighting the advantages of long-read data in SV detection [34].

To enhance the credibility of the "DNBSEQ" and "Illumina" SV sets generated in our research, we employed real-time PCR to validate the SVs in both SV sets (see Supplementary "Methods" for details). Due to the complexity and variability of the breakpoint regions for DUPs and INVs made it challenging to design reliable primers for these SV types [1, 6], our validation efforts focused exclusively on DELs and INSs in this study. Additionally, we selected only those SVs detectable by Manta in "DNBSEQ" or "Illumina" SV sets to accurately identify breakpoint sequences. From these "Manta-supported" SVs, we randomly selected 17 SVs for real-time PCR validation, including six "DNBSEQ"-specific SVs (three DELs and three INSs), five "Illumina" -specific SVs (three DELs and two INSs), and six shared SVs (three DELs and three INSs, Additional file 2: Table S7, see Supplementary "Methods" for detail). We designed and synthesized 25 primer pairs targeting the breakpoints of the DEL and INS and conducted real-time PCR assay (Additional file 1: Fig. S7 and Fig. S8, see Supplementary "Methods" for details). In summary, all 12 SVs from the "DNBSEQ" set were validated via real-time PCR, whereas nine out of 11 SVs within the "Illumina" set were validated (Table 1). In detail, six DELs and six INSs from "DNBSEQ" set were successfully validated, with the exception of one INS breakpoint (chr10:134,865,273). For this INS, only the left breakpoint was validated (Additional file 1: Fig. S8m), while the right breakpoint could not be confirmed (Additional file 1: Fig. S8n). In contrast, within the "Illumina" set, four DELs and five INSs were confirmed, while two DELs failed to be validated by real-time PCR. Moreover, all six shared

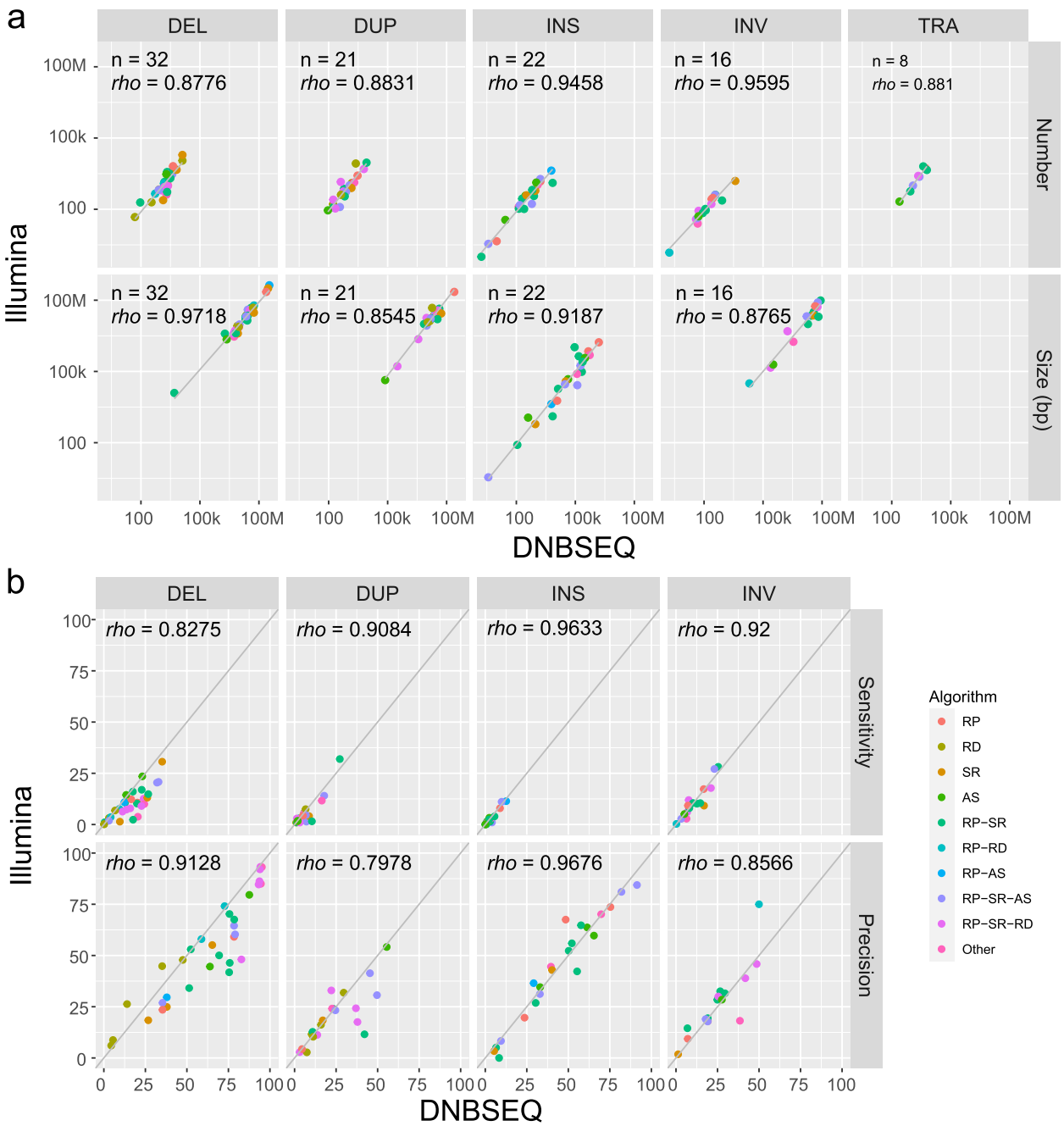

Fig. 3 Comparative Analysis of SVs Detected by DNBSEQ and Illumina Platforms. The dot plot presents both the number and size (**a**), as well as the sensitivity and precision (**b**), of SVs identified on the DNBSEQ platforms (x-axis) versus those detected on the Illumina platforms (y-axis) using various analytical tools. A total of 40 tools are differentiated by color based on their underlying algorithms. The types of SVs are organized into columns, while the attributes are arranged into rows. The symbol *rho* denotes the correlation coefficient as determined by Spearman's rank correlation, and n indicates the number of tools capable of detecting a specific type of SV

SVs (three DELs and three INSs) showed a validation rate of 100%. Given the high validation rates of both the "DNBSEQ" and "Illumina" SV sets, we further analyzed their genomic characterizes without additional modifications.

Genomic characterizing the SVs of the "DNBSEQ" and "Illumina" sets

Analyzing the genomic characteristics of SVs, such as repetitive DNA composition and GC content, provides valuable insights into the origins, mechanisms, and

Table 1 Validation of DNBSEQ and Illumina SV sets

Type	Set	Chr	Start	End	Validated or not
Deletion	DNBSEQ	chr3	10,397,475	10,397,832	Y
		chr10	91,547,385	91,547,736	Y
		chr17	35,143,423	35,143,742	Y
	Illumina	chr4	190,656,419	190,657,443	Y
		chr8	58,123,135	58,127,520	N
		chr19	30,388,815	30,393,167	N
	Common[a]	chr1	145,092,946	145,097,081	Y
		chr3	152,311,723	152,313,155	Y
		chr4	190,662,931	190,663,726	Y
Insertion[b]	DNBSEQ	chr9	114,736,428	114,736,428	Y / Y
		chr10	134,865,273	134,865,273	Y / N
		chr13	108,539,740	108,539,740	Y / Y
	Illumina	chr9	77,314,289	77,314,289	Y / Y
		chr11	22,262,800	22,262,800	Y / Y
	Common	chr5	173,192,947	173,192,947	Y / Y
		chr12	80,383,078	80,383,078	Y / Y
		chr13	90,943,356	90,943,356	Y / Y

[a] "Common" means this SV is found in both DNBSEQ and Illumina sets

[b] "Y/Y" indicates validation of two breakpoints per insertion

functional impacts of SVs, and even are crucial for disease research, evolutionary biology, and understanding genomic functions [35]. After evaluating the consistency of SVs from the DNBSEQ and Illumina datasets using various tools, we further analyzed and compared the genomic characteristics of SVs from both sets to explore their overall consistency. Firstly, we analyzed the repetitive DNA components and sequences of SVs in our "DNBSEQ" and "Illumina" sets. We observed that the size distribution of SVs was similar between these two sets, with both showing mobile element signatures of Alu (~300 bp) and LINE1 (L1, ~6 kb, Fig. 4a), which is consistent with previous report [9, 34]. These signatures might be driven by the fact that Alu and L1, as highly active mobile elements in the human genome, play a crucial role in generating mobile element insertions (MEIs) and driving non-allelic homologous recombination events, which contribute significantly to SVs [1, 36]. We also annotated the SVs to repeat regions and found that the majority were located in repeat regions (Fig. 4b). Specifically, 24.20% of SVs in the "DNBSEQ" set and 38.88% in the "Illumina" set were associated with tandem repeats (detected by Tandem Repeats Finder, TRF). This was followed by short tandem repeats (STR) at 27.61% and 24.51% in "DNBSEQ" and "Illumina", respectively, Alu elements at 26.04% and 16.57%, and L1 elements at 14.69% and 12.89%. Only a small percentage of SVs were not located in repeat regions, including 0.40% in "DNBSEQ" and 0.38% in "Illumina". These findings suggest that SVs are clustered in repeat regions on the human genome [34], and that both "DNBSEQ" and "Illumina" platforms are capable of detecting SVs in these regions.

We also analyzed the GC composition of SVs in the "DNBSEQ" and "Illumina" sets to better understand their sequence components. We found that, regardless of the SV set, the GC distribution of DUPs and INVs were close to the reference distribution, with enrichment in the 40–50% GC content range, while DELs and INSs exhibited a different pattern (Fig. 4c). DELs in both "DNBSEQ" and "Illumina" sets exhibited a bimodal GC distribution with peaks at 40% and 55%. Specifically, 34.55% (1,209/3,499) of DELs in the "DNBSEQ" set and 28.28% (1,251/6,797) of DELs in the "Illumina" set had a GC content in the 50–60% range, corresponding to the 55% peak (Fig. 4d). Among these DELs with 50–60% GC content, 73.45% (888/1,209) in the "DNBSEQ" set and 61.47% (760/1,251) in the "Illumina" set were associated with Alu sequences (Fig. 4d). In contrast to DELs, INSs in both the DNBSEQ and Illumina sets exhibited GC content distributions with peaks at 0%, 40%, and 100%. Among INSs with 0% and 100% GC content, 63.64% in the "DNBSEQ" set and 50.63% in the "Illumina" set were STR sequences, suggesting that this enrichment in the extreme GC content of INSs was primarily driven by STRs (Fig. 4d). These findings suggest that the GC content of SVs varies depending on the type of SV and the repeat elements involved. Despite these differences, the GC content patterns of SVs were consistent between the "DNBSEQ" and "Illumina" sets.

Previous study had reported that SVs exhibit non-random distribution patterns in the genome, with enrichment in repeat regions and a notable bias towards chromosomal ends [34]. We analyzed the chromosomal location biases of the "DNBSEQ" and "Illumina" SVs. We observed a 1.26-fold ($p = 4.726e-06$, z-score = 4.58, permutation test) enrichment of "DNBSEQ" SVs and 1.58-fold ($p < 2.2e-16$, z-score = 8.31, permutation test) enrichment of "Illumina" SVs within 5 Mbp of telomere, respectively, which is consistent with the findings of earlier study [34]. However, we also found that 15.01% of "DNBSEQ" SVs (718/4,785) and 15.99% of "Illumina" SVs (1,087/6,797) were located within 5 Mbp of centromere, presenting a 2.03-fold ($p < 2.2e-16$, z-score = 15.37, permutation test) and 2.34-fold ($p < 2.2e-16$, z-score = 16.56, permutation test) enrichment, respectively (Additional file 1: Fig. S9). In conclusion, our analysis demonstrated a consistent chromosomal location bias in both the 'DNBSEQ' and 'Illumina' SV sets, which was observed not only towards chromosomal ends but also towards centromeres (Fig. 5a). Since SVs were known to be clustered

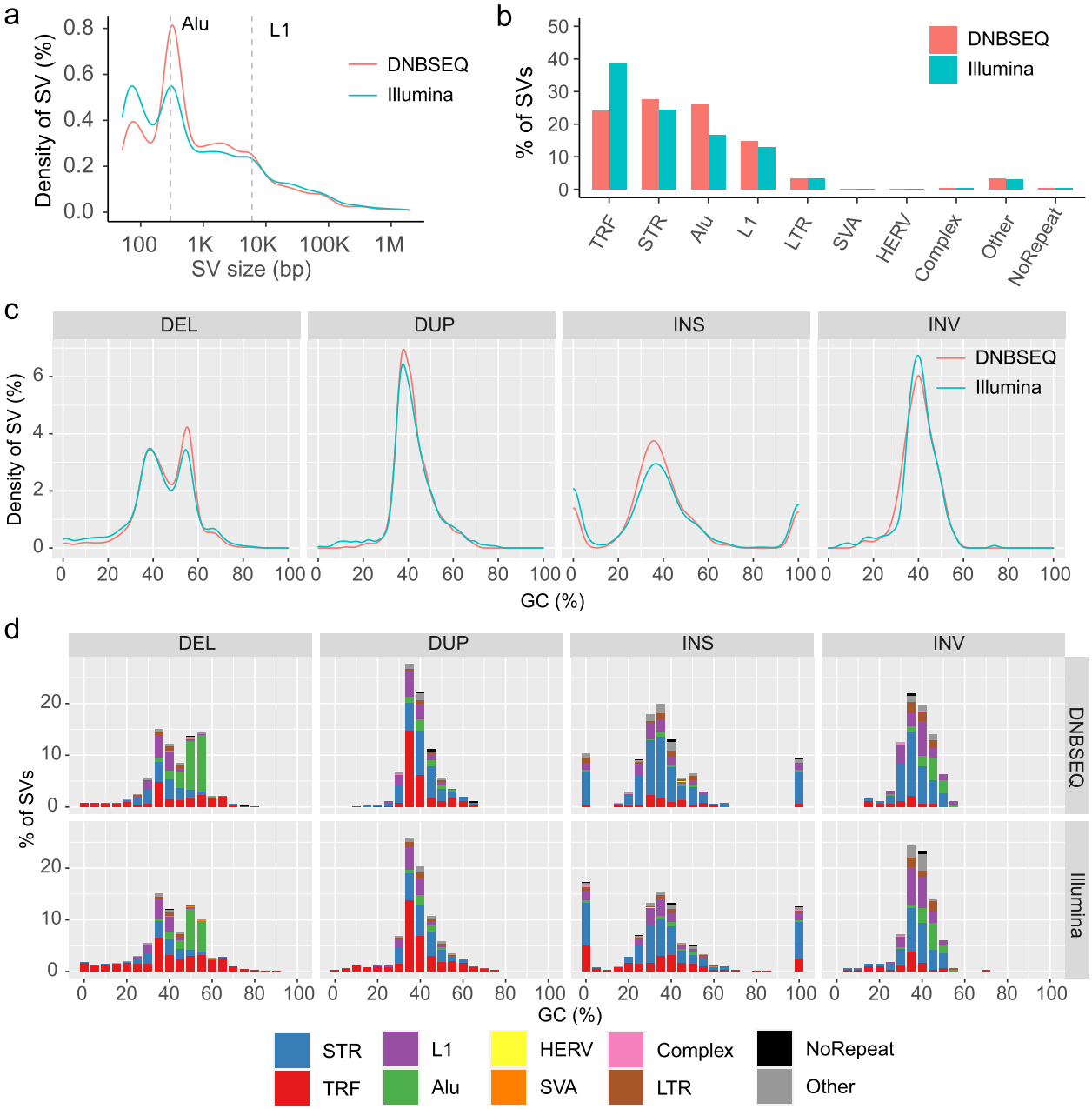

Fig. 4 Characteristics of "DNBSEQ" and "Illumina" SV Sets. **a** The size distribution density for DNBSEQ and Illumina SVs reveals a peak at 300 bp corresponding to Alu elements and a 6 kb peak associated with L1 elements. **b** The bar chart illustrates the percentage of DNBSEQ and Illumina SVs (y-axis) that are annotated within specific repetitive regions (x-axis). **c** The line density plot depicts the distribution of GC content across each type of SV (column), with the y-axis representing the proportion of SVs at a given GC content level (x-axis). **d** The bar chart details the repetitive components within SVs of varying GC content for each SV category (column), showing the percentage of repetitive regions (y-axis) to which SVs of particular GC contents (x-axis) are annotated. The GC content is categorized into bins at intervals of 5% GC. The DNBSEQ and Illumina SV sets are distinguished by color: DNBSEQ set, red; Illumina set, green. Different colors denote the repetitive regions: STR, short tandem repeats, blue; TRF, tandem repeats detected by Tandem Repeats Finder, red; L1, LINE1, purple; Alu, green; HERV, human endogenous retroviruses, yellow; SVA, orange; Complex, low complexity, pink; LTR, long tandem repeats, brown; Other, gray; NoRepeat, black

[34], we further identified hotspots in the "DNBSEQ" and "Illumina" sets and analyzed the chromosomal distribution of these hotspots. We identified 26 SV hotspots in the "DNBSEQ" set and 51 SV hotspots in the "Illumina" set (Additional file 2: Table S8 and Table S9). Of these, 30.77% (8/26) hotspots in the "DNBSEQ" set and 31.37%

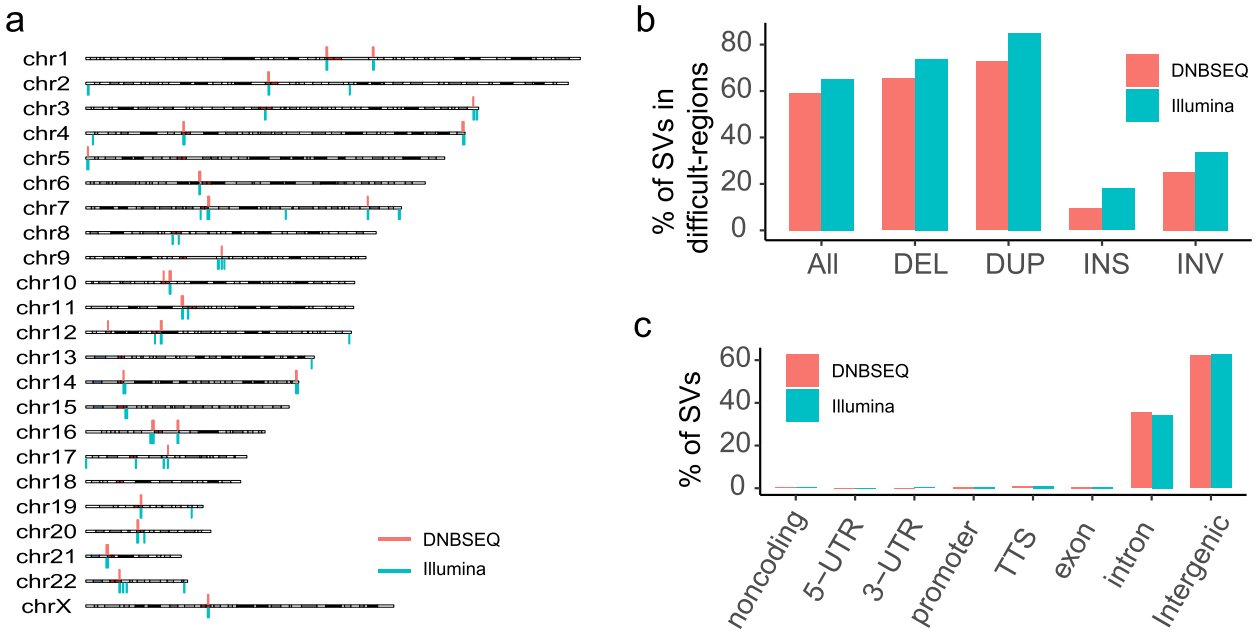

Fig. 5 Distribution of SVs. **a** The bar chart displays the hotspots of DNBSEQ and Illumina SV sets on human genome. The gray and white bands on ideogram indicate different genomic regions, and the red bands represent centromeres. **b** The bar chart indicates the percentage of DNBSEQ and Illumina SVs located within genomic regions that are typically difficult to sequence or analyze. The proportions are shown on the y-axis, while the x-axis categorizes the different types of SVs. **c** The bar chart illustrates the distribution of DNBSEQ and Illumina SVs across various gene regions, with the y-axis indicating the proportion of SVs mapped to specific gene regions on the x-axis. The DNBSEQ and Illumina SV sets are differentiated by color: DNBSEQ, DNBSEQ SV set, red; Illumina, Illumina SV set, blue

(16/51) hotspots in the "Illumina" set were previously reported in research using long-read datasets [10]. As expected, these hotspots in both the "DNBSEQ" and "Illumina" sets were mostly located near centromeres or telomeres.

Given that it is challenging to clone and sequence regions with GC-bias [34], we sought to assess the performance of "DNBSEQ" and "Illumina" SVs in these difficult genomic regions. To achieve this, we classified the "DNBSEQ" and "Illumina" SVs according to the difficult- and easy-to-sequence regions defined by the Genome in a Bottle (GIAB) Consortium [37]. We found that 59.16% of "DNBSEQ" SVs (2,831/4,785) and 65.18% of "Illumina" SVs (4,430/6,797) were located in difficult-to-sequence regions, even though these regions only make up 18.00% of the human genome (Fig. 5b, Additional file 2: Table S10). In detail, 65.36% of DELs (2,287/3,499) and 72.70% of DUPs (458/639) in the "DNBSEQ" set were enriched in difficult-to-sequence regions, whereas only 9.40% of INSs (47/500) and 25.00% of INVs (39/156) were found in these regions (Fig. 5b). Similarly, in the Illumina set, 73.73% of DELs (3,262/4,424) and 85.12% of DUPs (887/1,042) were located in difficult-to-sequence regions, compared to 18.02% of INSs (193/1,071) and 33.85% of INVs (88/260). These findings suggest that SVs are more likely to occur in specific genomic regions, including repeats and difficult-to-sequence regions, and exhibit GC-biases in their chromosomal locations.

Additionally, we annotated SVs to functional region of human genes to gain a better understanding of their impact on functional regions. We found that 62.28% of "DNBSEQ" SVs (2,980/4,785) and 62.70% of "Illumina" SVs (4,262/6,797) were located in intergenic regions, followed by 35.42% of "DNBSEQ" SVs (1,695/4,785) and 34.38% of "Illumina" SVs (2,337/6,797) in intronic regions (Fig. 5c). Only 2.03% of "DNBSEQ" SVs and 2.91% of "Illumina" SVs intersected with functional elements such as exons (26 "DNBSEQ" SVs and 40 "Illumina" SVs), promoters (25 and 46), and UTRs (8 and 22). Interestingly, 81.44% of "DNBSEQ" SVs (3,897/4,785) and 81.48% of "Illumina" SVs (5,538/6,797) were located within 0.25 Mb of transcription start site (TSS, Additional file 1: Fig. S10). Our analysis revealed that SVs in both the "DNBSEQ" and "Illumina" sets are predominantly located in intergenic regions and depleted in functional regions of genes, such as exons and introns. In conclusion, we further validated the consistency of SV sets from both DNBSEQ and Illumina platforms across multiple genomic characteristics, including repetitive regions, GC distribution, difficult-to-sequence regions, and gene features. This consistency underscores the robustness of our

comparative analysis and highlights the value of both platforms in understanding the genomic context of SVs.

Resource consumption

We recorded the time and memory of SV detection on ~ 30X WGS datasets using various tools (Additional file 2: Table S11). We found that Sprites (mean = 229.77 h and mean = 439.68 h on DNBSEQ and Illumina datasets, respectively, Additional file 1: Fig. S11), Pindel [38] (216.68 h and 163.32 h), MindTheGap [39] (166.92 h and 190.65 h), and laSV [40] (114.69 h and 121.17 h) had large time consumption, while laSV (mean = 150.91 GB and 143.47 GB on DNBSEQ and Illumina datasets, respectively), FermiKit [41] (73.52 GB and 54.42 GB), and MindTheGap (33.61 GB and 39.93 GB) had large memory consumption. This differences in time and memory consumption were related to the input format and algorithms of tools. However, we found high consistency of time consumption and memory consumption between the DNBSEQ and Illumina datasets using the same tools under similar ~ 30X data size (*rho* = 0.97 for time consumption and 0.95 for memory consumption, Spearman's rank correlation coefficient). These findings suggest that the choice of SV detection tool may depend on the specific needs of the study, such as the desired balance between time and memory consumption.

Discussion

Although SV detection on the Illumina platforms has increasingly demonstrated the importance of SVs, there is insufficient information regarding the performance of SV detection on another widely used MPS platform: the DNBSEQ platforms. In this study, we detected and characterized SVs, including DELs, DUPs, INSs and INVs, using DNBSEQ and Illumina datasets with 40 tools for the first time. Overall, our systematic analysis demonstrated that the DNBSEQ platform exhibits performance in SV detection that is consistent with the Illumina platform across various aspects, including the number, size, precision, and sensitivity of detected SVs, as well as their composition in repeats, genomic element distribution, and genomic localization.

Various tools have been designed to detect SVs using short-read signatures based on dataset sequenced on Illumina platforms. These tools have shown varying levels of precision and sensitivity for different SV types. This study demonstrated that SV detection tools developed for Illumina dataset are also compatible with the DNBSEQ dataset, as the results and performance of SV detection were consistent between the DNBSEQ and Illumina datasets using the same tool. However, we also observed notable performance differences between different tools, regardless of whether they were applied to DNBSEQ or Illumina datasets (Additional file 2: Table S4 and Table S12). For example, Manta, GRIDSS, SoftSV, and MetaSV demonstrated higher precision and sensitivity in detecting DELs and DUPs. Specifically, Manta excels due to its efficient use of RP and SR signals, enabling robust identification of breakpoints [42]. GRIDSS leverages a combination of SR and AS approaches, enhancing its ability to resolve complex SVs [43]. SoftSV integrates multiple alignment signals, including discordant read pairs and split reads, to improve detection accuracy [44]. MetaSV combines calls from multiple tools and refines them using local assembly, increasing its reliability for DELs and DUPs [45]. For INSs detection, MELT is a strong choice because it is specifically designed to identify mobile element insertions, utilizing both SR and RP evidence to accurately pinpoint insertion sites [46]. For INVs, TIDDIT, DELLY, and GRIDSS are particularly effective. Despite the varying performance of specific tools for particular types of SVs, when using the same tool for SV detection across datasets, we observed comparable concordance rate between the DNBSEQ datasets (mean = 55.40%, range 0.09%—96.93%) and the Illumina datasets (mean = 40.29%, range 1.61%—89.62%, Additional file 1: Fig. S12). These results not only confirm the consistency between the DNBSEQ and Illumina platforms but also highlight the importance and necessity of carefully selecting SV detection software, regardless of the data platform used.

Recent advancements in sequencing and data analysis technologies have significantly enhanced our detection and understanding of SVs. Long-read sequencing technology can fully sequence large DNA fragments (> 10 kb), providing continuous sequence that spans entire SVs [34]. For short-read dataset, pangenomic technology, particularly Giraffe, reduces reference allele bias and improves SV detection performance by mapping reads to a haplotype-resolved graph that includes references from thousands of human genomes [33]. In this study, we demonstrated the ability to detect SVs on short-read WGS data using DNBSEQ platforms and characterized the genomic features of these SVs. Our results showed a high consistent ratio of NA12878 SVs between "DNBSEQ", "Illumina", and "1KGP" sets, all of which were sequenced on short-read WGS data based on MPS platforms. For example, 73.34%, 53.05%, and 78.67% of DELs in the "DNBSEQ", "Illumina", and "1KGP" sets were shared, respectively (Additional file 1: Fig. S6). However, these three SV sets exhibit notable differences compared to the "HGSVC" set, which uses long-read data, and the "Giraffe" set, which employs pangenomic technology. For instance, 76.31% of DELs and 77.92% of INSs in the "HGSVC" set, and 61.07% of DELs and 70.53% of INSs in the "Giraffe" set, could not be detected in any of "DNBSEQ", "Illumina", or "1KGP" sets. This result highlights

the limitation of SVs detection, especially INSs detection, by normally short-read datasets, which is consistent with previous research [12, 33]. The distributions of SVs on short- and long-read platforms were also found to be inconsistent. Most of the SVs on long-read platforms were concentrated within 5 Mbp of the telomere [34], while the SVs on short-read platforms were enriched within 5 Mbp of both the telomere and centromere (Fig. 5a; Additional file 1: Fig. S9). These results confirm the complexity of SV detection and illustrate the stability and limitation of SV detection based on short-read MPS platforms. In our future work, we will continue to explore the performance of SVs detection in the DNBSEQ datasets combined with pangenomic technology, as well as analyze the SV detection performance of long-read platforms, such as PacBio [47], Oxford Nanopore [48] and CycloneSEQ [49].

In conclusion, we systematically analyzed the performance and characteristics of germline SVs detected in WGS datasets sequenced on the DNBSEQ platform. By evaluating the performance of SV detection with the same tool and integrating the results of all tools to assess the genomic characteristics of SV sets, our study demonstrated the consistency of SV detection between the DNBSEQ and Illumina platforms. Furthermore, we provided a benchmark reference for future SV detection based on the DNBSEQ platforms.

Methods

Sequencing data resources

Ten WGS germline datasets of NA12878 were utilized for SV detection, which were publicly available and analyzed in our previous article [29]. SOAPnuke (version 1.5.6) [50] was used to filter low-quality reads based on the following criteria: 1) adapter contaminations, 2) more than 10% of bases having a quality score < 10, and 3) more than 10% N bases. All reads that passed the quality filtering were subsequently aligned to the human reference genome (hg19) using BWA-MEM (version 0.7.10-r789) [51]. The resulting alignment files were processed using SAMtools (version 0.1.19) [52] and Picard (version 1.96, https://github.com/broadinstitute/picard) for further analysis. After data preprocessing, ten WGS datasets of NA12878 were obtained, with an average coverage of approximately 30X, mapping rate of over 99%, and genome coverage of over 99%.

Public SV sets of NA12878

Three public SV sets of NA12878, based on the reference genome GRCh38, were downloaded and extracted: 8,469 SVs detected using NovaSeq6000 by Marta et al. (referred to as "1KGP") [9], 4,718 SVs detected using PacBio RSII by Peter et al. (referred to as "HGSVC") [34] and 11,320 SVs (genotype quality ≥ 60) detected using Giraffe by Jouni et al. (referred to as "Giraffe") [33]. All three SV sets were converted to BED format and lifted over from GRCh38 to hg19 using the LiftOver tool (UCSC) to match the reference genome hg19 used in data alignment. This resulted in 8,236, 4,557 and 11,046 SVs of "1KGP", "HGSVC" and "Giraffe" sets, respectively.

SV detection

To provide a comprehensive performance assessment of SV detection in WGS germline datasets, we meticulously selected forty tools to ensure broad coverage across all five fundamental algorithms. The selection criteria were based on the following key factors: the ability to process individual WGS data, the capability to detect SVs in real datasets, compatibility with MPS short-read data, and ease of accessibility. Our curated selection includes five RD-based tools, four RP-based tools, three SR-based tools, three AS-based tools, and 25 tools based on a CA algorithm. These 40 tools were applied to the ten WGS datasets of NA12878 for SV detection, respectively. The SV calling were processed according to the approach described by Shunichi et al.[13]. SVs meeting the following criteria were filtered out: (1) the size of DEL, DUP and INV > 2 M bp or < 50 bp, (2) the number of reads supporting the called SV (RSS) < 3, (3) not located on autosomal or chrX chromosomes, or (4) overlapping a gap in the reference genome.

SV evaluation

A reference dataset of SV in NA12878 based on the reference genome hg19, as described in Shunichi et al. [13] was downloaded to evaluate SV performance (https://github.com/stat-lab/EvalSVcallers/blob/master/Ref_SV/NA12878_DGV-2016_LR-assembly.vcf). The dataset contained 9,241 DELs, 2,611 DUPs, 13,669 INSs and 291 INVs. The evaluation of SVs was performed using a script from https://github.com/stat-lab/EvalSVcallers based on a benchmark. DEL, DUP or INS was judged as true positive if it had a reciprocally overlap (RO) of ≥ 50% with the reference DEL, DUP or INS, respectively. INS was judged as true positive if the breakpoints of the called INS were located within ± 200 bp of those of the reference INS. The precision, sensitivity and F1-score were calculated using the following equations:

$$Precision = TP/(TP + FP)$$

$$Sensitivity = TP/(TP + FN)$$

$$F1 - score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$$

where TP is the true positive of SV, FP is the false positive, and FN is the false negative.

SV integration

We created a union set of SVs by integrating all SV results for each SV variant type (i.e., DELs were integrated with other DELs, and the same for DUPs, INSs and INVs, Additional file 2: Fig. S4). We integrated the SV set by merging pairwise comparison results sequentially in the order of the sample list in Additional file 2: Table S2. Specifically, the result of pairwise comparison between the first two datasets was compared with the third dataset, and so on, until all data were used. DELs, DUPs and INVs were excluded if they had a ≥ 50% RO with other DELs, DUPs or INVs in pairwise comparison, but INSs were excluded if they were located within 200 bp of other INSs. The common SVs detected in all datasets sequenced on the same platform (i.e., DNBSEQ or Illumina platforms) using the same tool were defined as candidate SV set of that tool, and the candidate SVs detected by two or more tools were merged as the SV set of the platform. Finally, we obtained 4,785 SVs on DNBSEQ platforms (referred as "DNBSEQ" set, Additional file 2: Table S5) and 6,797 SVs on Illumina platforms (referred as "Illumina" set, Additional file 2: Table S6), respectively. The low number of SVs in the "DNBSEQ" set is primarily due to the inherently low consistency ratio between SVs detected on different datasets using the same tool (mean = 55.40%, range 0.09%—96.93%, Additional file 1: Fig. S12). Therefore, the "DNBSEQ" set from DNBSEQ platforms with more integrated datasets has a lower number of SVs.

Comparison between SV sets

We conducted pairwise comparisons of the SV sets from "DNBSEQ", "Illumina", "1KGP", "HGSVC" and "Giraffe" for each SV type. DELs, DUPs and INVs were considered shared with another set if they had a ≥ 50% RO with DELs, DUPs or INVs in that set. For INSs, they were considered shared with another set if they were located within 200 bp of INSs in that set. For each specific SV set, the shared SV ratio was calculated as the proportion of SVs in this set that overlapped with all of the remaining sets, normalized by the total number of SVs in this set.

SV validation with real-time PCR

Manta was chosen to accurately identify the breakpoint sequences of these selected SVs due to its demonstrated excellent performance in SV detection reported in previous report [13] and this study (Additional file 2: Table S12), and its ability to output easily interpretable SV sequences in VCF format. Following the integration of SVs from both the DNBSEQ and Illumina datasets, we exclusively selected SVs detectable by Manta from the "DNBSEQ" and "Illumina" sets, respectively. Subsequently, we randomly selected DELs and INSs from these "Manta-supported" SVs for validation, extracting the breakpoint sequences from the VCF files generated by Manta. However, DUPs and INVs were not included in the validation process due to the unavailability of their breakpoint sequences necessary for primer design. Primer sequences were crafted using Primer Premier 6.0 and synthesized by BGI-Write (Additional file 2: Table S7). Human genomic DNA of NA12878 was purchased at the Coriell Institute. Real-time PCR assays were conducted using SYBR Green, with all procedures performed on the StepOne Real-Time PCR System (Applied Biosystems) in accordance with the manufacturer's protocol. For each breakpoint, seven real-time PCR reactions were executed as per the manufacturer's guidelines. This included triplicate reactions for the target primers, a no-template control to ensure specificity, and two positive controls: one using a standard sample and the other employing the GAPDH gene as an internal reference.

Identification of SV hotspots

The midpoint of the SV region was extracted and used to identify hotspot on hg19 genome. The midpoint sites were transformed using the 'makeGRangesFromDataFrame' function from the GenomicRanges package(v1.24.1) and then submitted to the 'hotspotter' function from the primatR package (with parameters: bw = 200,000, num.trial = 1000). The hotspots were displayed on hg19 genome using the karyoploteR package (v1.20.0).

Annotation of difficult regions

The difficult regions of hg19 were downloaded from the GIAB (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/genome-stratifications/v3.1/GRCh37/Union/GRCh37_alldifficultregions.bed.gz), which including 5,427,803 regions spanning 557.28 Mbp. We classified SVs in the "DNBSEQ" and "Illumina" sets as being in the difficult region if they were located in any difficult region with > 50% size overlap, and the remaining SVs were defined as being in the easy region.

Annotation of genes

The SVs in the "DNBSEQ" and "Illumina" sets were annotated to hg19 genes using HOMER annotatePeaks.pl (v4.11). We extracted the gene annotation and distance to TSS from the output of HOMER annotatePeaks.pl.

Annotation of repeat regions

The TRF and rmsk regions were downloaded from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/), while the SVA regions were obtained from MELT (v2.2.2.2). The SVs in the "DNBSEQ" and "Illumina" sets were annotated to these repeat regions using BEDTools (v2.30.0). Specifically, DELs, DUPs and INVs were annotated to TRF, rmsk and SVA regions, respectively, when > 50% of the SV size was located in a repeat region. INSs were annotated if they were within 50 bp of any repeat regions. In case where a single SV was annotated to multiple repeat regions, we selected the superior annotation based on the following priorities: TRF > STR > Alu > L1 > SVA > HERV > LTR > Low Complexity > Other Repeats.

Abbreviations

SNV	Single-nucleotide variant
INDEL	Insertion and deletion
WGS	Whole-genome sequencing
DEL	Deletion
DUP	Duplication
INS	Insertion
INV	Inversion
STR	Short tandem repeat
MPS	Massively parallel sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11494-0.

Additional file 1.

Additional file 2.

Acknowledgements

There are no additional acknowledgements to include.

Authors' contributions

J.R., L.P. and F.C. designed the project. H.L. performed the experiment. J.R., H.L., D.A. and X.L. performed the data analysis and involved in the data curation. J.R. and H.L. wrote the manuscript. J.R., L.P. and F.C. revised the manuscript. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (2021YFF1200105).

Data availability

The datasets generated and/or analyzed during the current study are available in the National Center for Biotechnology Information (NCBI) repository (https://www.ncbi.nlm.nih.gov/; the accession number SRX1049768 - SRX1049791, SRX1049832 - SRX1049855), the GigaDB repository (http://gigadb.org; the accession number 100274), and the CNGB Nucleotide Sequence Archive (CNSA; the accession number CNX0001345, CNX0001346, CNX0001989, CNX0023580, CNX0023581, CNX0023582, CNX0023583, CNX0001347). The code scripts used in this study are available on GitHub (https://github.com/Jh-Rao/SV-analysis).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 30 December 2024 Accepted: 17 March 2025
Published online: 25 March 2025

References

1. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363–76.
2. Abyzov A, Li S, Kim DR, Mohiyuddin M, Stutz AM, Parrish NF, Mu XJ, Clark W, Chen K, Hurles M, et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nat Commun. 2015;6:7256.
3. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85–97.
4. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–55.
5. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):aab3761.
6. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75–81.
7. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14(2):125–38.
8. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444–51.
9. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022;185(18):3426-3440 e3419.
10. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372(6537):eabf7117.
11. Collins RL, Glessner JT, Porcu E, Lepamets M, Brandon R, Lauricella C, Han L, Morley T, Niestroj LM, Ulirsch J, et al. A cross-disorder dosage sensitivity map of the human genome. Cell. 2022;185(16):3041-3055 e3025.
12. Zhao X, Collins RL, Lee WP, Weber AM, Jun Y, Zhu Q, Weisburd B, Huang Y, Audano PA, Wang H, et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. Am J Hum Genet. 2021;108(5):919–28.
13. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019;20(1):117.
14. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–84.
15. Holtgrewe M, Kuchenbecker L, Reinert K. Methods for the detection and assembly of novel sequence in high-throughput sequencing data. Bioinformatics. 2015;31(12):1904–12.
16. Zhang Z, Wang J, Luo J, Ding X, Zhong J, Wang J, Wu FX, Pan Y. Sprites: detection of deletions from sequencing data by re-aligning split reads. Bioinformatics. 2016;32(12):1788–96.
17. Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20(1):246.
18. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1784.

19. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14(Suppl 11):S1.
20. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78–81.
21. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–3.
22. Wang K, Wang J, Zhu C, Yang L, Ren Y, Ruan J, Fan G, Hu J, Xu W, Bi X, et al. African lungfish genome sheds light on the vertebrate water-to-land transition. Cell. 2021;184(5):1362-1376 e1318.
23. Liu S, Liu Y, Gu Y, Lin X, Zhu H, Liu H, Xu Z, Cheng S, Lan X, Li L, et al. Utilizing non-invasive prenatal test sequencing data for human genetic investigation. Cell Genom. 2024;4(10):100669.
24. Zhu H, Xiao H, Li L, Yang M, Lin Y, Zhou J, Zhang X, Zhou Y, Lan X, Liu J, et al. Novel insights into the genetic architecture of pregnancy glycemic traits from 14,744 Chinese maternities. Cell Genom. 2024;4(10):100631.
25. Xiao H, Li L, Yang M, Zhang X, Zhou J, Zeng J, Zhou Y, Lan X, Liu J, Lin Y, et al. Genetic analyses of 104 phenotypes in 20,900 Chinese pregnant women reveal pregnancy-specific discoveries. Cell Genom. 2024;4(10):100633.
26. Liu S, Yao J, Lin L, Lan X, Wu L, He X, Kong N, Li Y, Deng Y, Xie J et al. Genome-wide association study of maternal plasma metabolites during pregnancy. Cell Genom. 2024;4(10):100657.
27. Guo J, Guo Q, Zhong T, Xu C, Xia Z, Fang H, Chen Q, Zhou Y, Xie J, Jin D et al. Phenome-wide association study in 25,639 pregnant Chinese women reveals loci associated with maternal comorbidities and child health. Cell Genom. 2024;4(10):100632.
28. Jin X, Xu X, Zhou A, Zhu H, Li Q, Liu S, Liu S, Huang S, Zhang J, Wang T, et al. Advances in using non-invasive prenatal testing to study genomics related to maternity. Cell Genom. 2024;4(10):100677.
29. Rao J, Peng L, Liang X, Jiang H, Geng C, Zhao X, Liu X, Fan G, Chen F, Mu F. Performance of copy number variants detection based on whole-genome sequencing by DNBSEQ platforms. BMC Bioinformatics. 2020;21(1):518.
30. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, Qu S, Mei X, Chen H, Yu T, et al. A reference human genome dataset of the BGISEQ-500 sequencer. Gigascience. 2017;6(5):1–9.
31. Patch AM, Nones K, Kazakoff SH, Newell F, Wood S, Leonard C, Holmes O, Xu Q, Addala V, Creaney J et al. Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. PLoS One. 2018;13(1):e0190264.
32. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8.
33. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang PC, Carroll A, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science. 2021;374(6574):abg8871.
34. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. Characterizing the major structural variant alleles of the human genome. Cell. 2019;176(3):663-675 e619.
35. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020;21(3):171–89.
36. Deininger P. Alu elements: know the SINEs. Genome Biol. 2011;12(12):236.
37. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37(5):555–60.
38. Ye K, Schulz MH, 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. 2009;25(21):2865–71.
39. Rizk G, Gouin A, Chikhi R, Lemaitre C. MindTheGap: integrated detection and assembly of short and long insertions. Bioinformatics. 2014;30(24):3451–7.
40. Zhuang J, Weng Z. Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. Nucleic Acids Res. 2015;43(17):8146–56.
41. Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics. 2015;31(22):3694–6.
42. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32(8):1220–2.
43. Cameron DL, Schroder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27(12):2050–60.
44. Bartenhagen C, Dugas M. Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. Brief Bioinform. 2016;17(1):51–62.
45. Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HY. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics. 2015;31(16):2741–4.
46. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Genomes Project C, Devine SE. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27(11):1916–29.
47. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133–8.
48. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12(8):733–5.
49. Zhang J-Y, Zhang Y, Wang L, Guo F, Yun Q, Zeng T, Yan X, Yu L, Cheng L, Wu W et al: A single-molecule nanopore sequencing platform. bioRxiv. 2024:2024.2008.2019.608720.
50. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience. 2018;7(1):gix120.
51. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
52. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

Publisher's Note