https://doi.org/10.1186/s13059-023-03109-2

RESEARCH

Open Access

Quartet DNA reference materials and datasets for comprehensively evaluating germline variant calling performance

Luyao Ren[1†], Xiaoke Duan[1†], Lianhua Dong[2†], Rui Zhang[3†], Jingcheng Yang[1,4†], Yuechen Gao[1], Rongxue Peng[3], Wanwan Hou[1], Yaqing Liu[1], Jingjing Li[1,5], Ying Yu[1], Naixin Zhang[1], Jun Shang[1], Fan Liang[5], Depeng Wang[5], Hui Chen[6], Lele Sun[7], Lingtong Hao[8], The Quartet Project Team, Andreas Scherer[9,10], Jessica Nordlund[10,11], Wenming Xiao[12], Joshua Xu[13], Weida Tong[13], Xin Hu[14], Peng Jia[15], Kai Ye[15], Jinming Li[3], Li Jin[1], Huixiao Hong[13], Jing Wang[2*], Shaohua Fan[1*], Xiang Fang[2*], Yuanting Zheng[1*] and Leming Shi[1,14,16]

†Luyao Ren, Xiaoke Duan, Lianhua Dong, Rui Zhang and Jingcheng Yang contributed equally to this work.

*Correspondence: wj@nim.ac.cn; shaohua_fan@fudan.edu.cn; fangxiang@nim.ac.cn; zhengyuanting@fudan.edu.cn

[1] State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, Shanghai, China
[2] National Institute of Metrology, Beijing, China
Full list of author information is available at the end of the article

Abstract

Background: Genomic DNA reference materials are widely recognized as essential for ensuring data quality in omics research. However, relying solely on reference datasets to evaluate the accuracy of variant calling results is incomplete, as they are limited to benchmark regions. Therefore, it is important to develop DNA reference materials that enable the assessment of variant detection performance across the entire genome.

Results: We established a DNA reference material suite from four immortalized cell lines derived from a family of parents and monozygotic twins. Comprehensive reference datasets of 4.2 million small variants and 15,000 structural variants were integrated and certified for evaluating the reliability of germline variant calls inside the benchmark regions. Importantly, the genetic built-in-truth of the Quartet family design enables estimation of the precision of variant calls outside the benchmark regions. Using the Quartet reference materials along with study samples, batch effects are objectively monitored and alleviated by training a machine learning model with the Quartet reference datasets to remove potential artifact calls. Moreover, the matched RNA and protein reference materials and datasets from the Quartet project enables cross-omics validation of variant calls from multiomics data.

Conclusions: The Quartet DNA reference materials and reference datasets provide a unique resource for objectively assessing the quality of germline variant calls throughout the whole-genome regions and improving the reliability of large-scale genomic profiling.

Background

The detection of germline variants from high-throughput DNA sequencing (DNA-seq) is vital for biomedical research and molecular diagnostics of rare [1] and complex [2] genetic diseases. Well-characterized genomic reference materials can be used to benchmark measurement procedures, calibrate measuring systems and determine flagging criteria, and thereby support reliable application of genomic sequencing in basic research and clinical practice [3, 4].

Genome in a Bottle (GIAB) and other efforts have established various whole-genome reference materials and defined benchmark calls and regions to benchmark germline small variants (SNVs and indels) [5–8] and structural variants (SVs) [9–11]. However, all these efforts on genomic reference materials only evaluated variants identified inside the benchmark regions. Benchmark regions are partial of whole-genome region that are well-characterized and validated. When evaluating the performance of variant calling results using reference datasets based on standalone reference material, only variants within these benchmark regions can be assessed. However, the full extent of sequences generated and analyzed for a test genome is greater than what is defined by the boundaries of the benchmark regions. A substantial portion of variants detected outside the benchmark regions are overlooked, including many medically relevant variants [12]. Moreover, benchmark calls and regions are generally integrated from various sequencing technologies and bioinformatic pipelines, and thus biased toward easy-to-detect genomic contexts. Using variant calling performance inside the benchmark regions as a proxy will overestimate the overall performance of DNA assays or bioinformatic pipelines on the whole-genome region. Moreover, ignoring variants outside the benchmark regions will militate against objective understanding of the limitations of existing sequencing technologies, and thus hindering further method development.

Furthermore, in many practical applications of omics technologies, especially in large cohort studies, samples are often inevitably processed by multiple sequencing platforms at multiple centers over a relatively long period of time [13]. These large-scale projects usually suffer from batch effects due to the inconsistency of experimental conditions and sequencing machines [14, 15]. In DNA sequencing, batch effects are largely overlooked, but their widespread existence could lead to incorrectly taking batch-specific artifacts as real biological findings. Genomic reference materials are effective tools to identify and mitigate batch effects in DNA-seq [16]. Genomic reference materials can be sequenced along with test samples in every batch to determine whether batch effects exist. According to the properties of true positives and false positives detected from genomic reference materials, proper thresholds can be selected to remove batch-specific artifacts for each batch [17].

To address these challenges in DNA-seq and beyond, we established four DNA reference materials from Epstein-Barr virus (EBV)-immortalized lymphoblastoid cell lines of a Chinese Quartet family, including the biological parents and monozygotic twin daughters. The Quartet was recruited from the Fudan Taizhou cohort in Central China, possessing genetic features of both Northern and Southern Chinese populations [18]. We extensively sequenced the whole genomes of the Quartet reference samples using multiple short-read and long-read sequencing platforms. We integrated both small variant and structural variant benchmark sets for each of the Quartet reference samples for evaluating variant calling accuracy inside the benchmark regions. The genomes of the monozygotic twins are almost identical [19], and the expected number of germline de novo variants is fewer than 30 per generation and fewer than 1000 somatic mutations are introduced from cell culture [20]. The number of Mendelian violations in the detected variants is far more than the expected numbers of germline de novo variants and somatic mutations, indicating that most of the violations are sequencing or calling errors. Pedigree information of the Quartet members not only helped improve the specificity of benchmark sets by eliminating additional false positive variants with apparently high quality, but also facilitated the estimation of false positive rates of variants called outside the benchmark regions. The diverse sequencing data from the Quartet DNA reference materials also allowed us to identify batch effects present in whole-genome sequencing (WGS). The Quartet pedigree information was further used to develop a machine learning-based batch-specific filtration strategy to remove false positives and improve cross-batch reproducibility.

This study is part of the Quartet Project that aims for quality control and data integration of multiomic profiling (http://chinese-quartet.org/). Apart from the DNA reference materials, the Quartet Project also established matched RNA, protein and metabolite reference materials from the same culturing of the immortalized Quartet cell lines. Benchmark sets defined for the DNA reference materials facilitate evaluation of variant calling accuracy from RNA and protein data according to the principles of the central dogma. Accompanying papers on the overall project findings [21], transcriptomics [22], proteomics [23], metabolomics [24], batch effect monitoring and correction [25], and the Quartet Data Portal [26] can be found elsewhere.

Results

Study design with monozygotic twins and data generation

We established four immortalized lymphoblastoid cell lines of a Chinese Quartet family, including father (F7), mother (M8), and monozygotic twin daughters (D5 and D6) (Fig. 1a). The Quartet DNA reference materials are genomic DNA (gDNA) extracted from each immortalized lymphoblastoid cell line in large single batches. They have been certified by China's State Administration for Market Regulation as the First Class of National Reference Materials and are extensively being utilized for proficiency testing and method validation. The certified reference material numbers are GBW09900 (D5), GBW09901 (D6), GBW09902 (F7), and GBW09903 (M8).

To unbiasedly characterize germline small variants and SV benchmark calls, we sequenced all four Quartet genomes on four short-read (Illumina HiSeq and NovaSeq, MGI MGISEQ-2000, and DNBSEQ-T7 (30-60x coverage)) and three long-read (Oxford Nanopore Technologies (ONT) (100× coverage), Pacific Biosciences (PacBio) Sequel (80× coverage), and PacBio Sequel II (30× coverage)) sequencing platforms at seven centers. We then used four orthogonal technologies, including linked read sequencing (10× Genomics (30× coverage)), SNP array (the Axiom Precision Medicine Research Array (PMRA)), optical sequencing (BioNano), and PacBio circular consensus sequencing (CCS) reads (50× coverage) to validate and refine the benchmark calls (Fig. 1a and Additional file 2: Table S1).

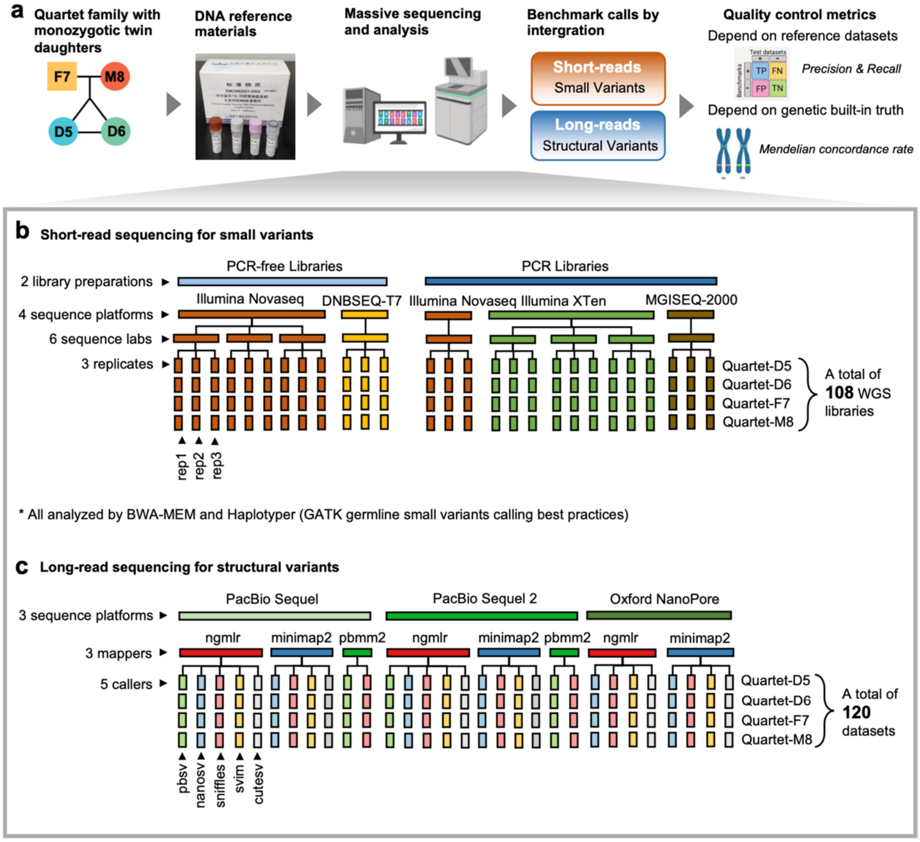

Fig. 1 Study design with monozygotic twins and data generation. **a** Overview of the study design. Briefly, DNA reference materials were constructed from immortalized cell lines of a Chinese Quartet with father (F7), mother (M8), and monozygotic twin daughters (D5 and D6). They were sequenced by four short- and three long-read platforms at seven labs. Small variant and structural variant benchmark calls were integrated from massive sequencing datasets. Performance of a test dataset can be evaluated by comparing with benchmark calls or genetic built-in truth within the Quartet family. **b** Schematic overview of short-read sequencing datasets. Three replicates for each of the Quartet DNA reference materials were sequenced in nine batches, by both PCR and PCR-free libraries on four sequencing platforms at six labs, resulting in 108 WGS libraries. **c** Schematic overview of long-read sequencing datasets. One replicate for each of the Quartet DNA reference materials was sequenced per batch by PacBio Sequel, PacBio Sequel II, and ONT. Eleven combinations of three mappers and five callers were used to detect structural variants, resulting in 120 variant calling datasets

A total of 108 germline small variants call sets were obtained from 27 short-read WGS libraries of each Quartet genome using the widely adopted GATK best practices (BWA-MEM and HaplotypeCaller (HC)) (Fig. 1b and Additional file 3: Table S2). A total of 120 germline SV call sets were obtained from three long-read WGS libraries of each Quartet genome with 11 combinations from three aligners (NGMLR [27], minimap2 [28], and pbmm2) and five callers (Sniffles [27], NanoSV [29], cuteSV [30], SVIM [31], and pbsv) (Fig. 1c and Additional file 4: Table S3 and Additional file 5: Table S4).

Variants call sets of the monozygotic twins are expected to be the same, because the twins share the identical genome from their parents. When investigating the consistency of call sets from different sequencing platforms, variant calling methods, and Quartet samples (Additional file 1: Fig. S1), we observed that SNVs, small indels (<50 bp), large insertions, or large deletions (≥50 bp), were clustered distinctly based on the identity of the Quartet samples, and the monozygotic twins were grouped together as expected.

However, for large duplications, inversions, or translocations (≥50 bp), the call sets did not cluster by the identity of the Quartet samples, but revealed strong clustering by bioinformatic pipelines, indicating lack of reliability of or consistency in bioinformatic pipelines for these three types of SVs. Thus, these three types of SVs were not included in the benchmark sets.

Determining small variant benchmark calls and regions

To define germline variant benchmark calls, we first selected reproducible variants among call sets for each of the Quartet samples. Because the number of Mendelian violations was much higher than the expected number of de novo mutations or somatic mutations arising from cell culture, all Mendelian violations were assumed to be errors [20]. Thus, we excluded Mendelian violations from the benchmark calls, even when they were reproducible among call sets.

We generated one small variant benchmark dataset by integrating 108 call sets (27 call sets per sample) of all four Quartet samples based on short-read WGS. At the individual sample level, we obtained a total of 6 million variants of 27 call sets at the beginning, and an average of ~4.6 million consensus variants after voting (see " Methods") across triplicates in a batch, sequencing labs, and library preparation methods (PCR-free and PCR) (Fig. 2). To check Mendelian consistency of the remaining variants, genotypes should be confidently detected in all four Quartet samples for each variant. We then removed a total of 412,054 variant positions with no-call or conflict genotypes among the 27 call

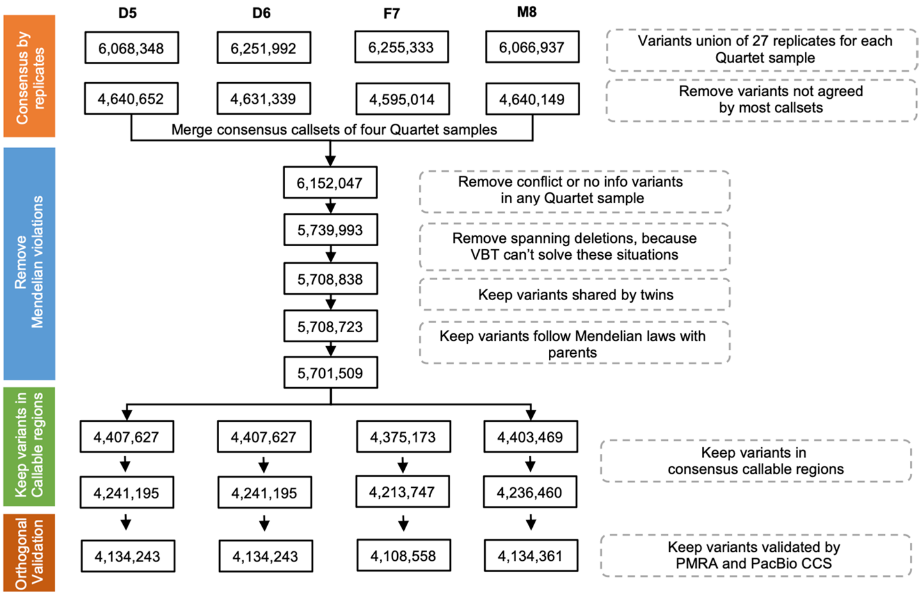
D5
D6
F7
M8
Consensus by replicates
6,068,348
6,251,992
6,255,333
6,066,937
Variants union of 27 replicates for each Quartet sample
4,640,652
4,631,339
4,595,014
4,640,149
Remove variants not agreed by most callsets
Merge consensus callsets of four Quartet samples
Remove Mendelian violations
6,152,047
Remove conflict or no info variants in any Quartet sample
5,739,993
Remove spanning deletions, because VBT can't solve these situations
5,708,838
Keep variants shared by twins
5,708,723
Keep variants follow Mendelian laws with parents
5,701,509
Keep variants in Callable regions
4,407,627
4,407,627
4,375,173
4,403,469
4,241,195
4,241,195
4,213,747
4,236,460
Keep variants in consensus callable regions
Orthogonal Validation
4,134,243
4,134,243
4,108,558
4,134,361
Keep variants validated by PMRA and PacBio CCS

Fig. 2 Integration workflow of Quartet small variant and structural variant benchmark calls. This workflow depicted the integration process to obtain small variant benchmark calls from 108 original GVCF call sets. Numbers in the boxes represented remaining small variants after each data processing step in the grey dotted boxes. Approximately 6 million small variants were discovered in 27 call sets for each Quartet reference sample. About 1.5 million small variants were removed by the voting process (" Methods"). We merged the four consensus call sets corresponding to the four Quartet samples, and discarded variants that did not reach agreement across 27 replicates in any Quartet sample. Only Mendelian consistent variants, which were shared by twins and following Mendelian inheritance laws and validated by PMRA and PacBio CCS datasets, were kept as small variant benchmark calls

sets of any Quartet sample. Compared with irreproducible variants filtered during the voting process, these removed variants showed higher variant allele frequency (VAF), read depth, mapping quality, and genotype quality (Additional file 1: Fig. S2). Therefore, they could not be removed by simply increasing variant filtration thresholds.

We identified 5,708,723 small variant positions with reproducible genotype calls among all four Quartet samples. These remaining variants were further examined for Mendelian consistency in the Quartet family, and 7329 (0.13%) of them were identified as Mendelian violations. We manually inspected 4761 variants located in the callable regions (see "Methods") with high mapping quality. Of the 3221 validated small variants, 1034 overlapped with large deletions. They were mistakenly considered as Mendelian discordant by Mendelian analysis software VBT [32], which was based on the hypothesis that variants always passed on diploid. Comparing with the variants detected in the matched blood samples of the Quartet family members, we found 95 pre-twinning germline de novo variants shared by the twins (homozygous reference in the parents and heterozygous or homozygous alternative in the twins), one postzygotic germline de novo variant specifically found in Quartet-D5, 1532 somatic variants (also found in blood), and 556 variants probably accumulated from cell culture (not found in blood) (Additional file 6: Table S5). Finally, we kept the Mendelian violations confirmed by manual curation into the initial catalog of benchmark calls. This process resulted in about 4.2 million well-supported small variants for each Quartet sample.

Previous studies show that PacBio CCS reads yield a higher variant calling accuracy compared with short-read NGS, especially when calling variants in the repetitive regions of the genome. When comparing with the variants based on 50× coverage of PacBio CCS reads, we found that 98.7% of SNVs and 95.0% of small indels in our benchmark dataset can be validated (Additional file 2: Table S6). The 89.7% unvalidated ones were found to be located in the repetitive regions of the genome, especially segmental duplications (41.6%) and centromere regions (27.9%).

We also validated the small variant benchmark dataset using 16 replicates of PMRA SNP array. We obtained 793,024 Mendelian consistent probes in the benchmark regions that were well-supported by most replicates from the 902,394 clinically related probes assayed on the PMRA array. Of those reliable probes, 99.99% homozygotic references, 98.6% SNVs, 95.7% small insertions, and 96.2% small deletions were the same with the NGS consensus variants (Additional file 2: Table S7). Among the 2845 discordant variants, 2704 were detected by the PMRA array but were absent from NGS. We manually inspected the read alignment and found that the remaining 141 calls were either missed by NGS or genotyped differently from the PMRA array, and only seven were obvious false positive in the NGS consensus calls due to misalignment of NGS reads. The seven obvious false positives were later removed from the small variant benchmark calls. Consequently, the two validation processes removed 61,532 SNVs and 61,152 indels from the benchmark call sets.

To enable the identification of false positive and false negative variants, we defined benchmark regions for small variants (Additional file 1: Fig. S3). These benchmark regions were derived by integrating callable regions, which are regions where short reads can be accurately mapped to the human reference genome with high mapping quality. Within the benchmark regions, we excluded high-confidence large deletions and insertions integrated from long reads, as well as their flanking regions (50bp). The benchmark regions were defined as high-confidence variant regions and homozygotic reference regions within the consensus callable regions, as determined by all Quartet samples. These regions covered approximately 87.8% of the GRCh38 reference genome (~2.66 G; chr1-22, X). Consensus and Mendelian consistent variants outside the benchmark regions were not included in the final benchmark call sets (Table 1).

We further compared the small variants benchmark calls with high-confidence call sets from two accompanying studies [33, 34] (Additional file 1: Fig. S4). These two high-confidence call sets provide orthogonal confirmation of our calls (FDU), since Pan et al. (NCTR) [33] integrated high-confidence calls from four mappers (Bowtie2, BWA, ISAAC, and Stampy) and eight callers (FreeBayes, GATK-HC, GATK-HC (sentieon), RTG, ISAAC, Samtools, SNVer, and Varscan), and Jia et al. (XJTU) [34]. constructed haplotype-resolved high-confidence calls by combining short-read and long-read technologies. We compared variants in the intersect of the three high-confidence regions of the three studies and found that 99.9% SNVs and 99.2% indels in our FDU callset could be confirmed by either the NCTR callset or the XJTU callset.

Table 1 Summary of quartet small variant and structural variant benchmark calls and regions

		Quartet-D5	**Quartet-D6**	**Quartet-F7**	**Quartet-M8**
Small Variant Benchmark Calls	Total variants [a]	4,122,817	4,122,817	4,097,306	4,123,162
	SNV	3,558,056	3,558,056	3,527,544	3,557,613
	sINS [b]	274,854	274,854	273,186	276,426
	sDEL [b]	281,212	281,212	278,427	280,765
	Block substitutions [c]	8695	8695	8149	8258
	Het/Hom ratio	1.37	1.37	1.30	1.35
	SNV Ti/Tv	2.08	2.08	2.07	2.07
	Benchmark region, chr1-22, X(bp)	2,658,688,832	2,658,688,832	2,658,688,832	2,658,688,832
Structural Variants Benchmark Calls	Total variants [d]	15,005	15,005	15,098	14,893
	INS [e] < 1kb	7216	7216	7353	7161
	INS ≥ 1kb	734	734	755	717
	DEL [e]< 1kb	6352	6352	6287	6324
	DEL ≥ 1kb	703	703	703	691
	Het/Hom ratio	1.43	1.43	1.45	1.57
	Longest INS (bp)	12,450	12,450	12,450	12,450
	Longest DEL (bp)	117,310	117,310	435,343	494,712
	Affected bases (bp)	7,796,176	7,796,176	8,821,558	8,550,650
	Benchmark region, chr1-22 (bp)	2,622,728,511	2,622,728,511	2,591,967,148	2,596,140,552

[a] All small variants benchmark calls are located in small variants high-confidence region, and false positive variants detected by orthogonal validation have been removed

[b] sINS and sDEL stand for short insertion and deletion with size less than 50 bp

[c] Block Substitutions are variants with length change between REF and ALT. It is not simple addition or removal of bases, for example, ATT -> CTTT

[d] Structural variant benchmark calls include variants not located in benchmark regions

[e] INS and DEL stand for long insertion and deletion with size over or equal to 50 bp

Determining structural variant benchmark calls and regions

A similar strategy was used to determine SV benchmark calls by integrating the 120 call sets obtained from the long-read WGS data (Fig. 3, " Methods"). For each individual in the Quartet family, the tool Jasmine was used to merge SV callsets from different sequencing platforms and SV calling pipelines based on the variant breakpoint and length [35]. This left about 90,000 isolated SVs of each Quartet sample. Then, SVs supported by the same pipeline from at least two sequencing platforms or by at least six pipelines from the same platform were determined as consensus SVs. Large SVs over 10 Mb and the ones located in centromeres, peri-centromere, and gap regions of the reference genome were excluded. The remaining 31,659 SVs were then re-genotyped in a pedigree using three genotypers (Sniffles [27], SVjedi [36], and LRcaller [37]) with the reads of PacBio Sequel and ONT. Consensus genotypes (23,891) from at least six of the ten genotype call sets were then determined as the consensus genotype calls for each of the Quartet samples. SVs with conflict genotypes had higher VAF (0.12–0.25 and 0.75-1.0) compared with discordant variants among replicates (0.12), but not as high as VAFs at peaks near 0.5 (heterozygous) or 1.0 (homozygous), respectively (Additional file 1: Fig. S5).

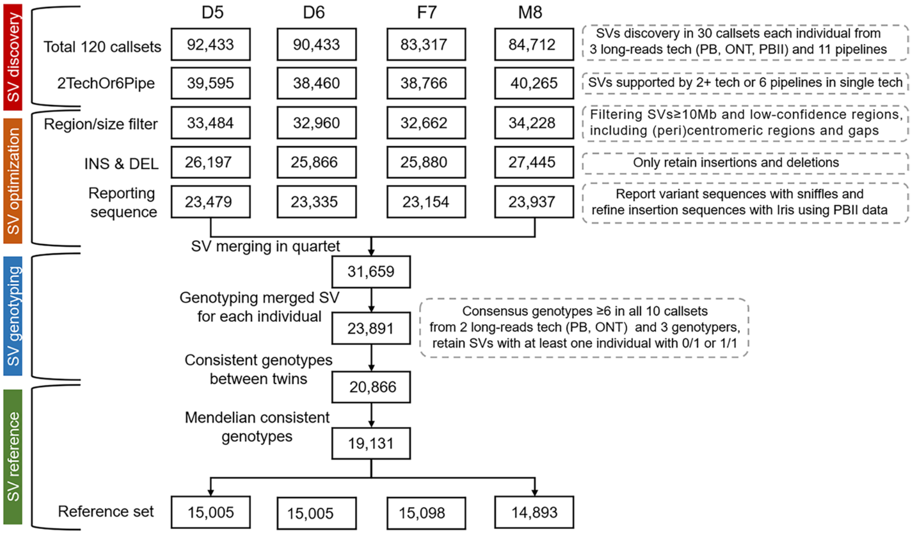
SV discovery
SV optimization
SV genotyping
SV reference
D5
D6
F7
M8
Total 120 callsets
92,433
90,433
83,317
84,712
SVs discovery in 30 callsets each individual from 3 long-reads tech (PB, ONT, PBII) and 11 pipelines
2TechOr6Pipe
39,595
38,460
38,766
40,265
SVs supported by 2+ tech or 6 pipelines in single tech
Region/size filter
33,484
32,960
32,662
34,228
Filtering SVs≥10Mb and low-confidence regions, including (peri)centromeric regions and gaps
INS & DEL
26,197
25,866
25,880
27,445
Only retain insertions and deletions
Reporting sequence
23,479
23,335
23,154
23,937
Report variant sequences with sniffles and refine insertion sequences with Iris using PBII data
SV merging in quartet
31,659
Genotyping merged SV for each individual
23,891
Consensus genotypes ≥6 in all 10 callsets from 2 long-reads tech (PB, ONT) and 3 genotypers, retain SVs with at least one individual with 0/1 or 1/1
Consistent genotypes between twins
20,866
Mendelian consistent genotypes
19,131
Reference set
15,005
15,005
15,098
14,893

Fig. 3 Integration workflow of structural variant benchmark calls. This workflow depicts the integration process to obtain structural variant benchmark calls from 120 call sets. Numbers in the box represented remaining structural variants after each data processing step in the grey dotted boxes. Briefly, approximately 90,000 structural variants were discovered in 30 call sets of each Quartet reference sample. We first kept structural variants supported by at least two sequencing platforms or at least six pipelines from one sequencing platforms, then removed SVs with length over 10 Mb or located on centromeric or pericentromeric regions and gaps. INSs and DELs were extracted for the construction of structural variants benchmark calls. Sniffles was used to report structural variants sequences, and structural variants that failed in reporting sequences were filtered. Iris was applied to refine variant sequences. After obtaining consensus of structural variants in multiple data sets, we merged four catalogs of reproducible variants of each Quartet reference sample and obtained 31,659 SVs in total. Three genotypers were used to determine genotypes of these SVs, and only SVs with consensus genotypes in at least six of all ten genotype call sets were kept for Mendelian analysis. The final structural variant benchmark calls were shared by twins and followed Mendelian inheritance laws with parents

After obtaining consensus genotyped SVs, we then removed Mendelian violated SVs. Of the 194 Mendelian violated SVs, we found that two de novo heterozygous variants shared by the twins, and four individual-specific heterozygous variants from one of the twins, which probably were somatic or arose from cell culturing (Additional file 2: Table S8). Following manual curation, we observed that the remaining 188 SVs were incorrectly genotyped. Most of them (91.7%) were located in regions of simple repeats over 100 bp or segmental duplications, or clustered with other variants. Finally, ~15,000 benchmark SVs were kept into the benchmark call set for each Quartet sample (Table 1). Consistent with prior studies, we observed three peaks near 300 bp, 2.1 kb and 6 kb, likely reflecting the activities of *Alu* elements, SVA elements, and full-length LINE1 elements in the human genome (Additional file 1: Fig. S6).

Validating based on Illumina short reads, 10X Genomics linked read, BioNano optical mapping, and whole-genome assemblies using PacBio CCS and PacBio CLR data, we found that our SV benchmark callset is of high quality (Additional file 2: Table S9). The overall validation rates of insertions and deletions were 95.24 and 95.78% by at least one technology. Although we integrated short-read SV validation callset using 15 SV callers, the validation rates by short-reads (48.7% INS and 76.0% DEL) were much lower than long-read assemblies (90.7% INS and 92.6% DEL). BioNano only validated 3.2% INS and 1.8% DEL over 1 kb, due to its low resolution (kb) by specific restriction enzyme cut sites and failure to accurately determine breakpoints [38]. We also validated our SV benchmark callset with Jia et al. [34] and found that 97.1% INS and 91.9% DEL were confirmed.

We also compared our SV benchmark calls to the SVs identified by GRC [39], HGSVC [40], and HX1 [41] with different groups of samples. The validation rates were 91.3, 77.8, and 54.7%, respectively. The high validation rate of GRC was because a Chinese sample was included, and the SVs were also detected from long-read data. Note that such comparison based on a limited number of samples will only detect the common SVs that are shared in different samples.

To define SV benchmark regions, we used ~100x PacBio Sequel CLR reads to establish haploid de novo assemblies for the parents F7 and M8 (2.94–2.99 Gb), and diploid de novo assemblies for the twins D5 and D6 (2.87–2.88 Gb). We then mapped de novo assemblies to the GRCh38 reference genome, and ~2.74~2.78 Gb callable regions were retained which were supported by reads larger than 50 kb and with mapping quality greater than 5. Regions of assembly-specific SVs, centromeres, and gaps were excluded from callable regions (Additional file 1: Fig. S7). The Quartet SV benchmark regions cover ~2.62 Gb of the reference genome (GRCh38; chr1-22) and contains ~12,705 (75.7–83.6%) SVs of the benchmark calls. Only SVs inside the benchmark regions are considered when we evaluate variant calling performance of test sets based on benchmark sets with precision and recall.

Applications of the Chinese Quartet genomic reference materials

Evaluating variant calling performance by pedigree information and benchmark sets

We used the whole-genome variant callsets derived from various library preparation methods, sequencing platforms, and bioinformatic tools to demonstrate the usability of the Quartet DNA reference materials in evaluation of variant calling performance.

Each callset was evaluated based on the F1 score in the benchmark regions and the Mendelian consistent rate (MCR) on the whole genome.

Four mappers (Bowtie2, BWA, ISAAC, and Stampy) and eight germline variant callers (HaplotypeCaller (GATK version and Sentieon version), RTG, ISAAC, Varscan, FreeBayes, Samtools, and SNVer) were compared based on ~30× Illumina short-read replicates from three sequencing centers (Detailed information can be found in our companion study [33]) (Fig. 4a). Callers had greater impact on variant calling accuracy compared with mappers. SNV calling performance was high and similar (F1 score 0.978±0.012, MCR 0.944±0.017) among different callers, while indels calling performance was lower and varied (F1 score 0.732±0.158, MCR 0.695±0.094). RTG,

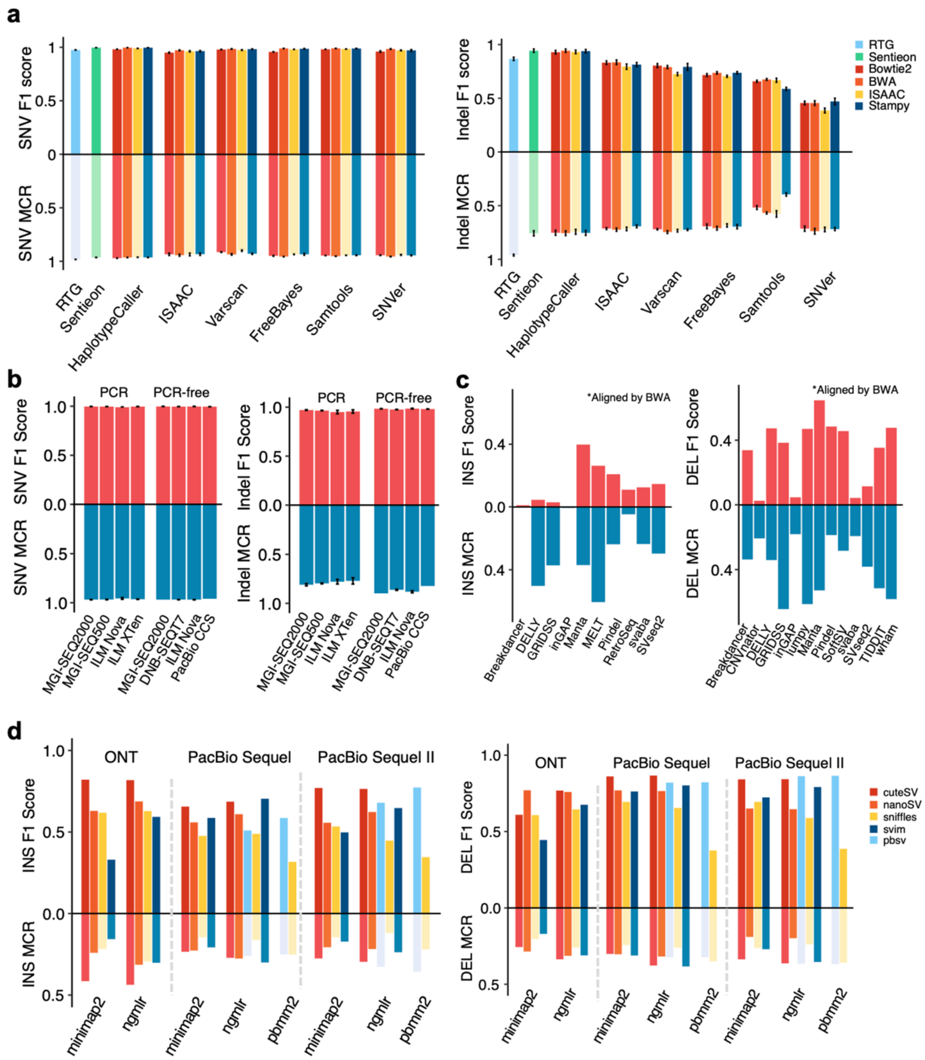

Fig. 4 Evaluating variant calling performance by pedigree information and benchmark sets. F1 score and MCR rate of different **a** mappers and callers for detecting small variants using Illumina short reads; **b** sequencing platforms and library preparation methods for detecting small variants; **c** callers for detecting SVs using Illumina short reads; and **d** sequencing platforms, combination of mappers and callers for detecting SVs using long-read data

Sentieon, and HaplotypeCaller showed higher F1 scores for indel calling, with Samtools and SNVer performing the worst.

To investigate the small variant calling performance of different sequencing platforms, we called small variants using the same pipeline (Sentieon) for short-read data and DeepVariant for PacBio CCS reads (Fig. 4b). Illumina platforms, MGI platforms, and PacBio CCS had similar performance, with no obvious differences. Sequencing platforms had smaller impact on variant calling accuracy compared with library preparation methods. PCR-free libraries were superior to PCR libraries for detecting Indels, with higher F1 scores (0.983±0.005 vs 0.958±0.016) and MCR rates (0.921±0.050 vs 0.873±0.094).

For investigating SV calling performance, we compared 15 common callers using short-read data (Fig. 4c). Different callers had various SV calling performance, with F1 scores ranging from 0 to 0.891 and MCR rates ranging from 0 to 0.645. Detection of DEL by short reads was slightly accurate than INS. Only Manta exhibited relatively high F1 score and MCR rate for both INSs and DELs compared to other callers. The MCR of INSs called by DELLY, GRIDSS, and MELT was much higher than F1 score evaluated by benchmark calls, because they detected fewer variants and had lower recall rates. We observed that most callers achieved high performance of DEL results, except for CNVnator, inGAP, and svaba. inGAP identified many more DELs (60,151) than benchmark calls, but had low precision and recall at the same time, indicating its low accuracy.

We also investigated SV calling performance of long-read sequencing platforms and bioinformatic pipelines, by retrospectively evaluating the performance of structural variants call sets used in this study to establish the benchmark sets (Fig. 4d). Generally, more SVs were detected from long reads (7726±3,203) than short reads (4922±9,604), and present sequencing technologies and algorithms display higher performance for DEL detection than INSs. Combination of mappers and callers should be carefully chosen according to sequencing platforms, since different combinations had F1 scores ranging from 0.374 to 0.856 and MCR rates ranging from 0.119 to 0.437. NGMLR with cuteSV showed high performance detecting both DELs and INSs on all three long-read sequencing platforms. Pbmm2 with pbsv, which was specifically developed for the PacBio platform, performed better on PacBio Sequel II than Sequel. Notably, DELs detected by pbmm2/sniffles had low F1 score but high MCR. Compared with the median het/homo ratio 2.2:1 in 30 call sets, het/homo ratio of pbmm2/sniffles was 0.02:1, which resulted in ~98% SVs of all four individuals with 1/1 genotypes, indicating that the genotypes of the pipeline were unreliable.

We found that an average of 9% SNVs, 40% indels, 33% DELs, and 20% INSs were located outside the benchmark regions, which could not be evaluated by benchmark sets. The F1 scores for variants inside the benchmark regions might not reflect the accuracy outside the benchmark regions (Additional file 1: Fig. S8a). As expected, the error rates were significantly higher outside of the benchmark regions. Moreover, the Quartet family design identified more false positive variant candidates compared to twins and trios and enabled a more precise measurement of error rates (Additional file 1: Fig. S8b).

Identifying and mitigating batch effects in genomic sequencing

To identify batch effects in WGS using the Quartet DNA reference materials, we performed principal component analysis (PCA) on genotype calls detected from various short-read sequencing platforms. Compared with RNA sequencing, DNA sequencing revealed a much smaller level of batch effect [22]. In the scatterplot of the first two eigenvectors, the monozygotic twin daughters were clustered together and located in the middle between the two parents in PC1 and above the parents in PC2, as expected (Additional file 1: Fig. S9). We observed a clear batch effect from the third and the fourth eigenvectors (Fig. 5a–d). The sequencing platforms played an important role in leading to such detectable batch effects. Large insertions exhibited the lowest reproducibility across the sequencing platforms compared with other variant types, because obvious batch effects were observed even from the first two eigenvectors. Variants called outside the benchmark regions showed larger batch effects than

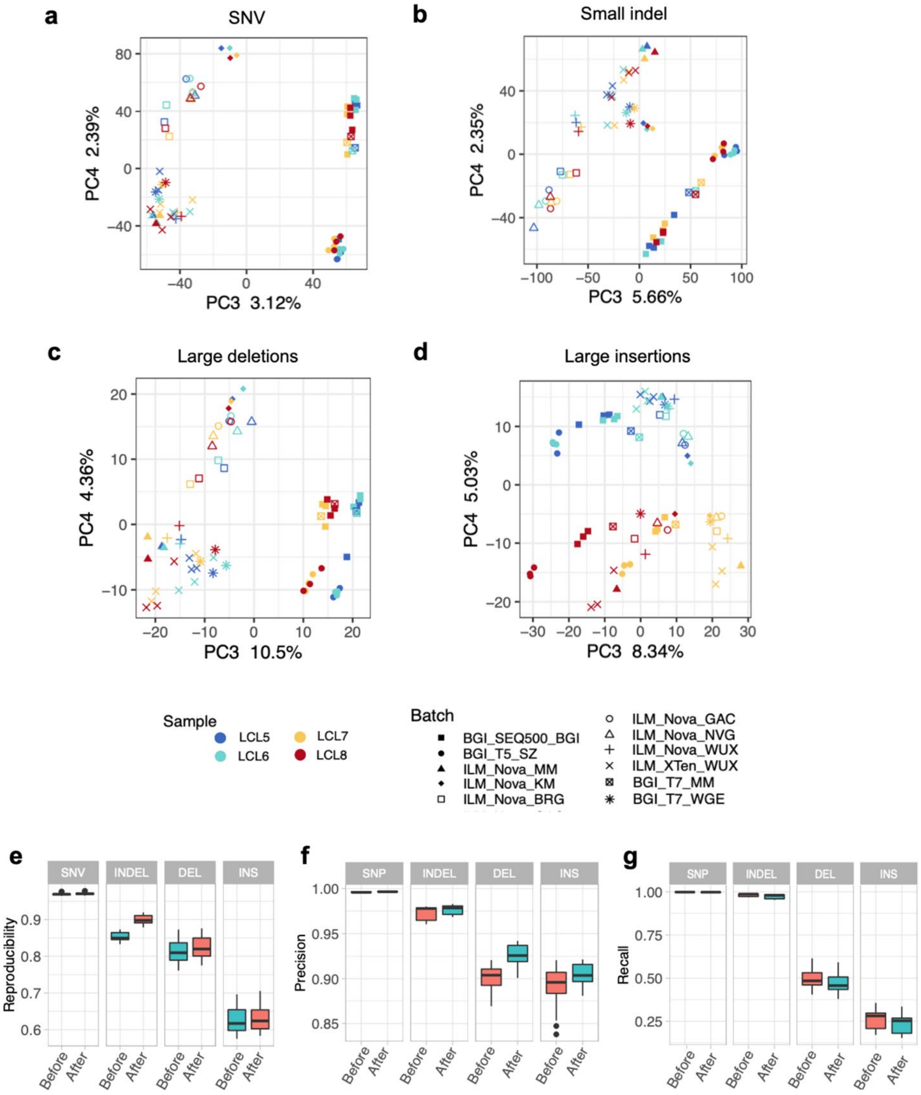

Fig. 5 Quartet DNA reference materials can be used to identify and mitigate batch effects in DNA sequencing. The scatterplots of the third and the fourth eigenvectors generated from PCA show batch effects in **a** SNVs, **b** small indels, **c** large deletions, and **d** large insertions. **e** Reproducibility of variants called on the whole-genome region before and after filtration. **f** Precision of variants called inside the benchmark regions before and after filtration. **g** Recall of variants called inside the benchmark regions before and after filtration

variants called inside the benchmark regions, as expected, because more variants outside the benchmark regions could not reach agreement among call sets (Additional file 1: Fig. S10).

Batch effects can be mitigated by removing false positive variants in each batch due to different variant quality metrics such as quality scores, read depth, and mapping quality scores. Pedigree information of the Quartet DNA reference materials can be used to select proper thresholds of those variant quality metrics for each batch to filter potential artifacts. We trained a one-class SVM (support vector machines) classifier using variant quality metrics of Mendelian consistent variants (reliable variants) from one of the three replicates for each batch (Additional file 2: Table S10, batches 5, 6, and 7). Then the trained models were applied on the other two replicates to filter potential false positives for each batch. The efficiency of batch-specific filtration method was assessed by precision, recall, and cross-batch reproducibility (Fig. 5e–g). After filtration, the cross-batch reproducibility was greatly improved. The precision compared with the benchmark calls increased, while the recall rates decreased, indicating that false positives were greatly reduced with inevitably sacrificing a small number of true variants.

Evaluating variants called from mRNA and protein

Apart from DNA reference materials, we also established RNA, protein, and metabolite reference materials from the same large batch of B-lymphoblastoid cell lines. Multiomic reference materials from the same resources of Quartet cell lines provide possibilities for cross-validating biological findings from one type of omics dataset by other levels of omics datasets, supporting quality assessment of a wide range of new technologies and bioinformatics algorithms.

We illustrated a cross-omics validation of variants detected using the Quartet genomics, transcriptomics, and proteomics datasets. As shown in Fig. 6a, an average of 15,580 RNA variants and 18 missense single-amino acid variants were detected in RNA-seq and LC-MS/MS-based proteomics of Quartet D5, respectively. We compared the variants called by GATK HaplotypeCaller and DeepVariant [42] in the intersected genomic regions of benchmark regions developed in this study, CDS regions, and regions with a minimum of 3× coverage. On average, GATK detected 16,305 SNVs and 1338 Indels, while DeepVariant detected 13,181 SNVs and 334 Indels in these intersected regions. DeepVariant called fewer variants, but yielded higher precision and recall than GATK for both SNVs and Indels (Fig. 6b). About 8.3% RNA variants from DeepVariant and 25.9% from GATK in RNA-seq could not be validated by DNA small variant benchmark calls. When comparing false positives against known RNA editing sites in REDIportal [43], we observed that DeepVariant disregarded RNA editing event (17 out of 1100), whereas GATK detected more RNA editing events (562 out of 4234) (Fig. 6c). This indicates that RNA editing does not have a significant contribution to the high level of inconsistency observed between variants identified from DNA and RNA sequencing datasets. Instead, the discrepancy is primarily attributed to technical artifacts. Figure 6d shows that a specific SNV benchmark call can be validated by both RNA and protein sequencing data. A

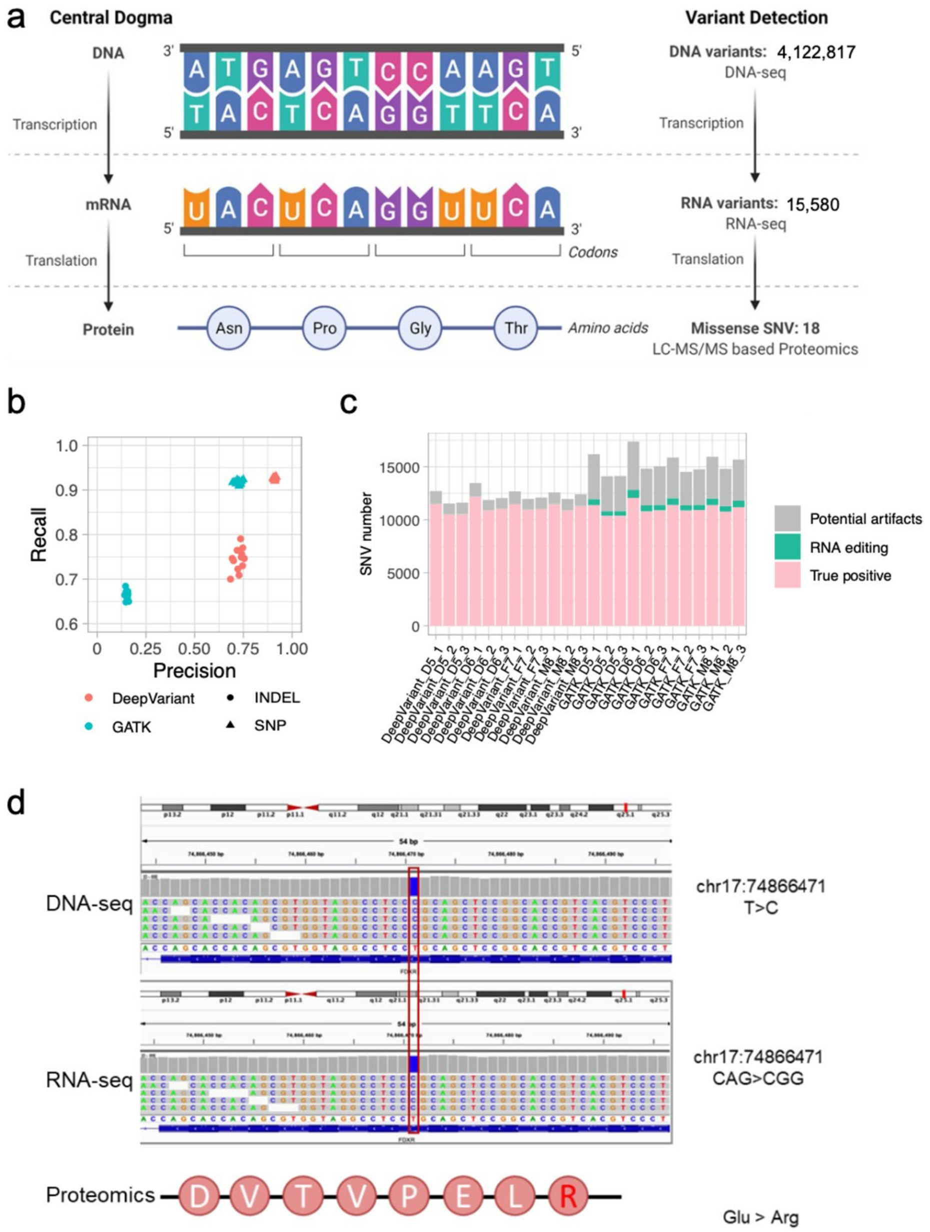

Fig. 6 Evaluating variant calling accuracy from RNA and protein data by benchmark variants constructed from DNA data. **a** Schematic of central dogma and the number of variants detected in the Quartet DNA-seq, RNA-seq, and LC-MS/MS-based proteomics datasets. **b** Validation of Quartet RNA variants using DNA reference datasets. True positive (TP) means RNA variants validated in DNA reference datasets, whereas false positive (FP) means the RNA variants not included in the DNA reference datasets. **c** Composition of RNA variant types in false positive (RNA_FP) and true positive RNA (RNA_TP) variant calls. **d** A T-to-C variant (located in chr17: 74866471) detected by both DNA-seq and RNA-seq is visualized in IGV. The corresponding Glu-to-Arg variant was also detected by LC-MS/MS-based proteomics

missense SNV (chr17:74,866,471 T>C) caused a single-amino acid mutation, changing from glutamic acid to arginine.

These preliminary cross-omics validation results implicated that current applications for variant detection from RNA sequencing and LC-MS/MS-based proteomics remain a challenge. The Quartet multiomics reference materials and datasets enable objective quality assessment of these emerging bioinformatics algorithms from cross-omics validations.

Discussion

One primary challenge of germline variants performance assessment by a single reference sample is that the benchmark sets focus on evaluating the performance of easily detected variants and genomic regions, but ignore difficult variants outside the benchmark regions. Here, we established four DNA reference materials from a Chinese Quartet with parents and monozygotic twins. We constructed high-quality germline benchmark calls, including SNVs, small indels, large insertions and deletions for each Quartet reference sample based on extensive short-read and long-read sequencing. The quality of the benchmark calls was improved through a series of data-filtering procedures including consensus voting of replicates, pedigree information, and orthogonal technologies.

We demonstrated that the use of the Quartet DNA reference materials together helps make a comprehensive performance assessment of variants across the whole genome. There are two aspects of "truth" related to the Quartet DNA reference materials. One aspect is related to the benchmark calls, where only highly confident variants were kept. Precision and recall are commonly used metrics to evaluate variant calling performance within benchmark regions. The benchmark regions currently cover 85.5–87.7% of the human reference genome, encompassing 92.3–93.6% of gene regions, 91.5–93.4% of exon regions, 92.2–94.4% of CDS regions, and 92.9–94.4% of regions containing medically relevant variants in ClinVar (Table 2). Another aspect is the genetic built-in truth of the monozygotic twins and their parents. The Mendelian concordance rate of variants among the Quartet members can be used to estimate the accuracy of the fraction of variants that are not included in the benchmark regions. Compared to other studies focusing on easy-to-detect variants in benchmark regions alone, difficult variants outside the benchmark regions not only reflect major discordances among different sequencing platforms and labs, but also help guide future development and optimization of sequencing technologies.

Table 2 Coverage of quartet benchmark regions on coding region and clinically related genes

Class	Total bases (bp)	Small variant benchmark regions of Quartet[1] v1.1 (%)	SV benchmark regions of Quartet D5&D6 v1.1 (%)	SV benchmark regions of Quartet-F7 v1.1 (%)	SV benchmark regions of Quartet-M8 v1.1 (%)
GRCh38 (chr1-22, X)	3,031,042,417	87.7	86.5	85.5	85.7
Gene	1,776,904,532	93.6	93.4	92.5	92.3
Exon	156,255,092	93.4	93.0	91.5	91.7
CDS	35,744,776	94.4	93.6	92.2	92.4
ClinVar	6,247,973	94.4	93.6	92.9	92.9

[1] The benchmark regions of the four Quartet reference samples are the same

[2] Genomic regions of gene, exon and CDS (coding sequences) are extracted from gencode v43

[3] Genomic regions of variants of ClinVar are extracted from ClinVar v2023-06-17

There are also drawbacks only using pedigree information instead of benchmark sets for performance assessment. For example, systematic sequencing or mapping errors, such as heterozygotic or homozygotic variants called on all Quartet samples, which are Mendelian consistent, will be mistakenly considered as true variants. In some cases, Mendelian concordance rate is low due to sequencing failure of one or more Quartet reference samples. Comparison with the benchmark calls can help identify which sample exhibits bad variant calling performance. Notably, pedigree information can be used to evaluate Mendelian concordance, but it cannot help determine false negatives. Therefore, benchmark sets are necessary to identify false negatives and measure recall rate, while the pedigree information provides additional tool for the assessment of variant calling accuracy outside the benchmark regions.

As part of the Quartet Project, the main objective of our study is to provide the scientific community with genomic DNA reference materials that can be used to assess and improve the accuracy of germline variant calling. However, we acknowledge that the initial version of the small variant and structural benchmark sets for Quartet DNA reference materials does not include complex variants and genomic regions. This limitation arises from the challenges associated with mapping short reads to repetitive genomic regions and the potential mapping errors that can occur when calling structural variants solely through mapping approaches. In our companion study [34], we addressed these limitations by generating haplotype-resolved whole-genome assemblies for the monozygotic twin daughters. Decoding the complete genome of a diploid sample, compared to the complete hydatidiform mole (CHM13), presents more challenges. Nevertheless, we achieved high quality in the assemblies, with 76% of the chromosomal arms being gap-free from telomere to centromere. The updated benchmark regions increased to cover 92.43% of the GRCh38, including more complex variants and regions, through the integration of short reads, long reads, and haplotype assemblies. In our analysis, we compare benchmark sets integrated from three different sources: (1) multiple technical replicates generated from various platforms, laboratories, and batches analyzed using GATK best practices (as described in our study); (2) three batches of datasets called from multiple pipelines [33] ; and (3) haplotype assemblies [34] (Additional file 1: Fig S4). Our findings revealed that while benchmark variants integrated from variants called from short reads and long reads by mapping approaches may miss complex variants, there is high consistency in the benchmark variants within the overlapping benchmark regions, with 99.99% agreement for SNVs, 99.51% for indels, 91.5% for large deletions, and 97.1% for large insertions.

To evaluate and monitor the performance of the data generation processes, sequencing all the Quartet genomes is not cheap, especially for long-read sequencing. If one is only interested in variants or regions in the benchmark calls and regions, we recommend sequencing one of the Quartet samples and making quality assessment using benchmark sets by precision and recall. If the aim is to improve current technologies in some challenging genomic regions, we recommend sequencing all four Quartet samples to estimate performance on those difficult regions. Since a new technology is often accompanied by advantages beyond what current technologies can offer, the Quartet based Mendelian concordance rate is independent of the benchmark calls and can provide a more objective evaluation.

To monitor and improve data quality across different sequencing centers in large-scale studies, we recommend sequencing all the Quartet DNA reference materials per batch. In an automated library preparation setup, 96 samples are routinely handled in a batch. Although including four quality control samples per batch increases experimental cost by ~5%, it can benefit the study tremendously by identifying and mitigating batch effects for the sake of discovering genuine biomarkers for precision medicine.

The Quartet DNA and other types of omics reference materials are publicly available to the community by requesting through the Quartet Data Portal website (http://chinese-quartet.org/). We encourage researchers to upload and share Quartet sequencing data, thereby hoping the rich collections of diverse datasets and analysis for the Quartet samples will enable optimization of the benchmark sets and regions.

Conclusions

In summary, the Quartet DNA reference materials and datasets are essential resources for objective and comprehensive evaluation of the quality of sequencing and bioinformatic methods, which will greatly improve the quality control awareness of the sequencing community and help overcome barriers to the translation of findings from genomic studies into clinical practices.

Methods

Establishing DNA reference materials

The Chinese Quartet DNA reference materials were extracted from four immortalized B-lymphoblastoid cell lines transfected by Epstein-Barr virus, including father (F7), mother (M8) and monozygotic twin daughters (D5/D6). We extracted two batches of DNA on August 6, 2016 and October 28, 2017 from two large expansions of the cell lines. We diluted DNA to 220 ng/μL and made >1000 aliquots for each DNA sample. Each vial contains 10 μg of DNA in TE buffer (10 mM TRIS, pH 8.0; 1 mM EDTA, pH 8.0). The Quartet DNA is stored at −80°C for long-term preservation, or at 4°C for short-term preservation. We checked the integrity of DNA (DIN) by Agilent 4200 and the distribution of DNA fragment length by Agilent 2200. The Quartet DNA is stable for at least 3 years at −80°C and for 3 weeks at 4°C during the entire duration of quality monitoring. This study focuses on germline variant calling quality control. Two batches (Lot 20160806 and Lot 20171028) of DNA reference materials were extracted from large expansion of cell lines, with 1000 tubes (10 μg, 220 ng/μL) for each Quartet reference sample at each batch. DNA reference materials are stable and in good quality. The peak size of DNA fragments is over 60 kb. The stability has been monitored monthly for 3 years, with DNA integrity number (DIN) over 8.5.

Library preparation and whole-genome sequencing

Short-read sequencing

Twelve tubes of Quartet DNA reference materials, with three replicates for each of the four Quartet sample types, were sequenced per batch. DNA reference materials were from Lot 20160806. We obtained datasets from four sequencing platforms in six sequencing labs by PCR and PCR-free library protocols, resulting in 27 replicates per sample and 108 libraries in total:

(1) ~50× paired-end, whole-genome sequencing with 2×100 bp reads of ~250 bp insert size from MGI MGISEQ-2000 with PCR library kit, performed at BGI.
(2) ~30× paired-end, whole-genome sequencing with 2×150 bp reads of ~300 bp insert size from Illumina HiSeq XTen with TruSeq Nano library kit, performed at ARD and NVG.
(3) ~30× paired-end, whole-genome sequencing with 2×150 bp reads of ~400 bp insert size from Illumina HiSeq XTen with TruSeq Nano library kit, performed at WUX.
(4) ~30–60× paired-end, whole-genome sequencing with 2×150 bp reads of ~300–400 bp insert size from Illumina NovaSeq6000 with PCR-free library kit, sequenced at ARD, BRG, and WUX.
(5) ~35× paired-end, whole-genome sequencing with 2×150 bp reads of ~380 bp insert size from MGI DNBSEQ-T7 with PCR-free library kit.

Long-read sequencing

To establish structural variant benchmark calls, the four Quartet DNA reference materials, one replicate for each sample, were sequenced on three long-read platforms, resulting in three libraries per sample and 12 libraries in total:

(1) ~100×, whole-genome sequencing with 11–14 kb mean read length and 20–25 kb N50 read length from Oxford Nanopore Technologies (ONT). DNA reference materials were from Lot 20171028.
(2) ~100×, whole-genome sequencing with 8–11 kb mean read length and 13–18 kb N50 read length from PacBio Sequel (CLR). DNA reference materials were from Lot 20160806.
(3) ~30×, whole-genome sequencing with 16–18 kb mean read length and 26–28 kb N50 read length from PacBio Sequel II (CLR). DNA reference materials were from Lot 20160806.

We also generated sequencing datasets from BioNano, 10x Genomics, and PacBio CCS reads to validate benchmark calls:

(1) BioNano Genomics: ~200X for D5, ~300X for D6, F7 and M8 BioNano Genomics data with average fragment length 260~300 kb. DNA reference materials were from Lot 20160806.
(2) 10x Genomics: ~30X Genomics data with average fragment length ~150 kb. DNA reference materials were from Lot 20160806.
(3) ~50×, whole-genome sequencing with 13–14 kb mean read length and 13–14 kb N50 read length from PacBio Sequel II (CCS HiFi reads). DNA reference materials were from Lot 20160806.

Reads mapping and variant calling for short-read sequencing for developing benchmark variants

Sequences were mapped to GRCh38 (https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files). We used Sentieon Genomics software (https://www.sentieon.com/) to analyze short-read WGS datasets from raw fastq files to GVCF files. This workflow was derived from recommended germline small variant calling pipeline by the Broad Institute (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-), including read mapping by BWA-MEM, duplicate removing, indel realignments, base quality score recalibration (BQSR), and variant calling by HaplotyperCaller in GVCF mode. Then we performed joint variant calling using Sentieon GVCFtyper to merge all 108 GVCF files. We used default settings for all processes.

Different from regular VCFs, GVCF files have records and extra information for all genomic sites. A site is recorded as a variant call, homozygotic reference, or with no reads covered. In a regular VCF, we cannot distinguish a site with no information from a homozygotic reference. GVCF files enable us avoid mistaking no-call sites as homozygotic references and facilitate representation of complex variants as well.

To keep as many variants as possible and not to remove any potential true variants with low qualities, we did not filter variants from the original GVCF call sets by empirical variant quality or machine learning-based variant quality score recalibration (VQSR).

Reads mapping and variant calling for long-read sequencing for developing benchmark variants

We used three mappers (NGMLR, minimap2, and pbmm2) and five callers (cuteSV, NanoSV, Sniffles, pbsv, and SVIM) to call structural variants, resulting in 11 combinations. Reads were mapped to human genome version hg38 (GCA 000001405.15) from UCSC Genome Brower (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/).

PacBio Sequel-based call sets were generated as follows:

(1) Reads were aligned with NGMLR v.0.2.7 with -x pacbio parameter, minimap2 v.2.17-r941 with -x map-pb –MD -Y parameters and pbmm2 v.1.0.0 with –sort –median-filter –sample parameters separately.

(2) Structural variant calling was performed using cuteSV v.1.0.4 with –genotype parameter, NanoSV v1.2.4 with per chromosome pattern and an ancillary file containing random positions in hg38, Sniffles v.1.0.11 with default parameter, and SVIM v.1.2.0 with –minimum_depth 10 parameter based on BAM files created by NGMLR v.0.2.7 and minimap2 v.2.17-r941 separately. Additionally, Sniffles v.1.0.11 was also run on pbmm2 v.1.0.0 and pbsv v.2.2.1 was run on pbmm2 v.1.0.0 and NGMLR v.0.2.7. The pbsv discover stage was run with –tandem-repeats parameter using tandem repeat annotation file human_GRCh38_no_alt_analysis_set.trf.bed (https://github.com/PacificBiosciences/pbsv/tree/master/annotations). The pbsv discover and call stages were both run on the full genome.

PacBio Sequel II-based call sets were generated as follows:

(1) Reads were aligned with NGMLR v.0.2.7 with -x pacbio parameter, minimap2 v.2.17-r941 with -x map-pb –MD -Y parameters and pbmm2 v.1.0.0 with –sort –median-filter –sample parameters separately.
(2) Structural variant calling was performed using cuteSV v.1.0.4 with -s 3 –genotype parameters, NanoSV v1.2.4 with per chromosome pattern and an ancillary file containing random positions in hg38, Sniffles v.1.0.11 with -s 3 parameter and SVIM v.1.2.0 with –minimum_depth 3 parameter based on BAM files created by NGMLR v.0.2.7 and minimap2 v.2.17-r941 separately. Additionally, Sniffles v.1.0.11 with -s 3 was also run on pbmm2 v.1.0.0 and pbsv v.2.2.1 was run on pbmm2 v.1.0.0 and NGMLR v.0.2.7. The pbsv discover stage was consistent with PacBio Sequal-based process.

Nanopore-based call sets were generated as follows:

(1) Reads were aligned with NGMLR v.0.2.7 with -x ont parameter and minimap2 v.2.17-r941 with -x map-ont –MD -Y parameters.
(2) SVs were called using cuteSV v.1.0.4 with –genotype parameter, NanoSV v1.2.4 with per chromosome pattern and an ancillary file containing random positions in hg38, Sniffles v.1.0.11 with default parameter and SVIM v.1.2.0 with –minimum_depth 10 parameter based on BAM files created by NGMLR v.0.2.7 and minimap2 v.2.17-r941 separately.

In addition to the parameters of mappers and callers mentioned above, the others are default.

Detecting structural variants from Illumina-based short-read sequencing

Illumina NovaSeq WGS short-read sequencing with ~40× 2×150 bp and 420 bp insert size was performed at ARD and used to call structural variants. The reads were mapped to the GRCh38.d1.vd1 reference genome by Sentieon BWA. According to previous studies [44], 15 algorithms with relatively high precision and/or recall were selected for structural variants discovery, including Breakdancer [45], CNVnator [46], DELLY [47], GRIDSS [48], inGAP-sv [49], LUMPY [50], Manta [51], MELT [52], Pindel [53], softSV [54], SvABA [55], Svseq2 [56], tardis [57], TIDDIT [58], and Wham [59]. Consequently, 15 Illumina-based call sets were generated for each Quartet reference sample. Structural variants were filtered based on the number of reads supporting structural variants (RSS), types, and lengths. For several algorithms, RSS value was not available and other values such as quality scores were used to simulate RSS. Only five types of structural variants were retained (INSs, DELs, DUPs, INVs, and BNDs). Structural variants under 50 bp were removed except for BNDs. The filtered output file for each algorithm was converted to a VCF format with SVMETHOD, END, SVTYPE, and SVLEN tags in the information field. All 15 call sets for each individual in Quartet were merged into a single call set based on the same type and with breakpoints distance of 1 kb using SURVIVOR v.1.0.7.

Detecting small variants and structural variants from 10x Genomics linked reads

Small variants and structural variants were called by longranger-2.2.2 (https://support.10xgenomics.com/genome-exome/software/downloads/latest) with default parameters from 10x Genomics linked read data sets. Small variants were from phased_variants.vcf.gz. Structural variants ≥ 50bp were from dels.vcf.gz and large_svs.vcf.gz were retained.

Detecting structural variants from BioNano

The structural variants were called by BioNano Solve v3.1 (bnxinstall.com/solve/Solve3.1_08232017) with default parameters.

Detecting small variants and structural variants from PacBio CCS reads

The small variants were called by DeepVariant (https://github.com/google/deepvariant) with default parameters. The structural variants were called by pbsv (https://github.com/PacificBiosciences/pbsv) with default parameters.

Detecting structural variants from PacBio assembly alignments

The complete diploid assembly was reconstructed based on trio binning of canu v1.8 from PacBio Sequel CLR data (~100X) of twins and Illumina NovaSeq ~40× 2×150bp WGS short-read sequencing data with 420 bp insert size performed at ARD for the two parents. Trios are formed by twins D5 and D6 and their parents respectively. Each trio is then assembled independently. The assembly was performed with canu -p prefix -d prefix genomeSize=3.1g -pacbio-raw pacbio.fasta.gz -haplotypeF7 F7.NGS.fastq.gz -haplotypeM8 M8.NGS.fastq.gz. The diploid assembly results of parents were generated by FALCON v0.4 with default parameters based on ~100x PacBio Sequel CLR sequencing data.

Two methods of assembly alignment were used, including MUMmer v4.0.0beta2 and minimap2 v.2.17-r941. MUMmer assembly alignments were performed with the commands nucmer -maxmatch -l 100 -c 500 ref.fa –prefix haplotype.contigs.fasta. Minimap2 assembly alignments were performed with the commands minimap2 -cx asm5 -t12 –cs ref.fa haplotype.contigs.fasta. Three assembly-based callers were used including Assemblytics V1.2.1, SVMU V0.4, and Paftools (https://github.com/lh3/minimap2/tree/master/misc). Assemblytics was run with the parameters unique_length_required=10000 min_size=20 max_size=1000000 by MUMmer alignment. Results were transformed into VCF format using SURVIVOR. SVMU was run with default parameters by MUMmer alignment. Paftools was run with default parameters to identify structural variants from the CS tags generated by Minimap2 alignment. Results of SVMU and Paftools were transformed into VCF format using a custom script. Structural variants of two contigs of the twins were merged into a single call set, and then structural variants shared between twins are used to validate structural variant benchmark calls.

Preprocessing and filtering of structural variants call set from long-read sequencing

Due to considerable diversity in the number, type, and size of structural variants and the format of VCF files created by different caller algorithms, it was difficult to merge the original VCF files directly for downstream analysis. In order to unify the standard and facilitate the analysis, structural variants call sets were preprocessed as follows:

(1) Only five types of structural variants (INS, DEL, DUP, INV, and BND) were retained for each call set. For Sniffles, complex structural variant types were excluded. For SVIM, DUP_INT, and DUP:TANDEM were converted to DUPs. For pbsv, CNVs were filtered.
(2) All structural variants under 50 bp were removed except for BNDs.
(3) All structural variants call sets were filtered if they do not meet the minimum number of supporting reads. For ~100X PacBio Sequel and ONT sequencing datasets, structural variants call sets from cuteSV and Sniffles were filtered with tag RE $\geq$10. SVIMs were filtered with tag SUPPORT $\geq$10. Structural variants called by NanoSV were filtered with tag DV $\geq$10. For pbsv, structural variants were filtered based on read depth of variant allele $\geq$3 of tag AD. The parameter median-filter in pbmm2 v.1.0.0 only aligns the subread closest to the median subread length per ZMW and significantly reduces the number of reads supporting structural variants, thus a lower filtering threshold should be used. Otherwise, pbsv will lose too many true variants. For ~30X PacBio Sequal II, heuristically, the minimum number of reads supporting structural variants in all call sets from cuteSV, Sniffles, SVIM, and NanoSV was adjusted to three. The filtering threshold of pbsv was the same as that of PacBio Sequel for the parameter –median-filter.
(4) All structural variants call sets were assigned a unique ID based on sequencing platform, sample name, pipelines, serial number, and structural variant type for backtracking easily.

Integration of small variant benchmark calls

The construction process of high-confidence variant calling can be divided into three steps. Firstly, select the variants that are reproducibly detected in multiple datasets. Secondly, retain the variants that adhere to Mendelian inheritance patterns in Quartet family. Lastly, keep the variants within callable regions.

GVCF files of 108 libraries were merged by joint variant calling process for each chromosome separately (chr1-22, X), with samples in columns and variants in rows. This process was run for each chromosome. Since variants detected in only one or a few datasets have a higher probability of being false positives, we kept variants that are consistently detected in multiple datasets. We first integrated the three technical replicates generated in each batch, and variants supported by at least two out of three replicates proceed to the next round of integration. For example, if a variant in a specific batch had genotypes of [“0/1”, “1/1”, “0/1”] across the three technical replicates, after integrating the replicates, the genotype detected for that batch at the locus was determined as “0/1.” We next integrated the voting results across nine batches. Genotypes supported by at least four out of five PCR library preparation batches were considered as the integration result for PCR libraries, while genotypes supported by at least three out of four PCR-free library preparation batches were considered as the integration result for PCR-free libraries. For example, after integrating the technical replicates, the genotype results for the five PCR library preparation batches were ["0/1", "1/1", "1/1", "1/1", "1/1"]. After batch integration, the resulting genotype for PCR libraries was "1/1." We lastly integrated the voting results across library preparation methods. If the consistent genotype voting results were the same in both PCR and PCR-free results, the variant is considered "reproducible." The GVCF files of the 27 technical replicates for each Quartet sample were merged into one VCF result for each of four Quartet samples. Each variant was annotated with voting status. "Conflict" refers to genotypes that did not pass the voting integration across technical replicates, batches, and library preparation methods. "./." indicates no call in most replicates. If a variant's genotype was determined through the above voting process in the 27 replicates, it was annotated as the integrated genotype.

The four integrated VCF files were merged, excluding any loci annotated as "Conflict" in any of the Quartet samples, and loci voted as "0/0" or "./." in all four Quartet samples. A total of 31,155 small variant positions overlapping deletions were removed in all four Quartet samples, which represented with "*" in gvcf files, because downstream analysis tools cannot deal with * allele. Mendelian inheritance status of remaining sites was checked by VBT [32] with the parameter "-no-call explicit." We split Quartet into two "trios" (D5-F7-M8 and D6-F7-M8), and performed Mendelian analysis by VBT separately. Only variants shared between twins and Mendelian consistent with parents were retained.

Lastly, we retained the variants in the callable regions as high-confidence variants. Callable regions are characterized by having sufficient coverage and quality of sequencing reads, enabling reliable variant detection. Callable regions are typically defined by specific criteria, such as a minimum read depth and mapping quality. By focusing on callable regions, researchers can ensure that their variant calling analysis is performed on regions of the genome with reliable data, enhancing the accuracy and confidence of the results. We described the way we defined callable regions for Quartet samples in the "Methods" section "Defining benchmark regions."

Integration of structural variant benchmark calls

The benchmark structural variants were constructed based on all 120 long-read sequencing structural variants call sets described above, only including chr1-22:

(1) Structural variant callers with different detection algorithms lead to the same variant being called with different breakpoints and lengths. Moreover, due to the scoring systems of aligners and different clustering methods of callers, some large structural variant events were split into several smaller INSs/DELs in a local region. These redundant variants inflated the number of structural variants and hindered subsequent merging calls between different callers for the same sample. Jasmine v.1.0.1 (https://github.com/mkirsche/Jasmine) uses an improved minimum spanning forest algorithm to merge different variants within a single caller or between callers. Each variant was represented by a breakpoint (start, length) in two-dimensional space. The distance between the two variants was equal to the Euclidean distance (default) by their breakpoints. When the distance between variant breakpoints met the max_dist value (default value 1000, Euclidean distance: [(start1-

start2) ^2+ (length1-length2) ^2]1/2 ≤1000), these close variants with the same variant type were clustered into a single structural variant event.

(2) We used Jasmine with –allow_intrasample, --keep_var_ids and –ignore_strand parameters to merge structural variants between callsets for each sample.
(3) The integrated structural variants set of each individual sample was subsequently filtered to retain structural variants supported by at least two long-read sequencing platforms or at least six call sets in a single technology.
(4) The four integrated structural variant sets in Quartet were merged into one call set by Jasmine with –keep_var_ids and –ignore_strand parameters.
(5) Structural variants were excluded if their size is over 10 Mb and in low-confidence regions, including centromeres, pericentromeric region, and gaps in hg38 reference genome.
(6) Structural variants frequently occur on repeats, which seriously hinders accuracy of detecting breakpoints and sequences on the alternative allele. Structural variants with explicit sequences were also helpful for subsequent genotyping. Therefore, we used Iris v1.0.1 to report alternative allele sequences of INSs and DELs. It extracted breakpoints by racon or falcon_sense to get consensus sequences. Then NGMLR or minimap2 was used to re-align these sequences of the breakpoints to the reference genome for refining the variant breakpoints and sequences. The read names of supporting structural variants and allele sequences were obtained by Sniffles with -n -1 -s 2 –Ivcf parameters. We refined INSs and DELs by Iris with max_out_length=1000000, --also_deletions and –pacbio parameters. In addition, the minimap2 bam files from PacBio Sequel II of each Quartet sample were adopted for reporting sequence and refining breakpoints, because PacBio Sequel II sequencing datasets had lower mismatch rates.
(7) We re-genotyped merged structural variants from two long-read sequencing platforms (PacBio Sequel and ONT) by three long-read genotypers (LRcaller v0.1.2, Sniffles v1.0.11 and SVJedi v1.1.0) with default parameters. The bam files from NGMLR and minimap2 of PacBio Sequel and ONT were used by Sniffles and LRcaller. The fasta files of PacBio Sequel and ONT were used by SVJdei. Thus, for each Quartet sample, a total of 10 genotyping call sets were produced, four from LRcaller, four from Sniffles, and two from SVJedi. SVs were considered successfully and concordantly genotyped if at least six of the ten genotypes were the same.
(8) The structural variants successfully genotyped as heterozygous variants or homozygous variants in at least one of four Quartet samples were retained as input of Mendelian analysis. We retained structural variants that were shared by twin daughters and Mendelian consistent with parents, using bcftools v.1.9-224-g96ef00a.

Defining benchmark regions of small variants

First, we obtained callable regions from bam files using GATK V3.8-1 CallableLoci for each of the 108 short-read libraries, with –maxDepth 300 –maxFractionOfReadsWithLowMAPQ 0.1 –maxLowMAPQ 1 –minBaseQuality 20 –minMappingQuality 20 –minDepth 10 –minDepthForLowMAPQ 10 parameters.

We next selected consensus callable bed regions for each Quartet reference sample, if bed regions were denoted as callable (1) at least 2/3 replicates in one batch, (2) at least 4/5 batches by PCR library preparation and 3/4 batches by PCR-free, and (3) both PCR and PCR-free library preparation methods. Then we kept regions callable in all Quartet samples.

We obtained reproducible invariant genomic positions by the same voting process. We then converted reproducible invariant genomic positions and high-confidence small variant positions to bed region, and kept regions where all Quartet samples had concordant voting results.

Benchmark regions include positions of small variants benchmark calls and invariant homozygotic reference positions in consensus callable regions mention above. Thus, we got regions which were callable and had consistent calling results among replicates and all Quartet samples.

Defining benchmark regions of structural variants

When evaluating analysis methods using structural variant benchmark calls, structural variants were limited in the benchmark regions, which could assess the accuracy of genotyping about INSs and DELs.

The process for constructing benchmark regions was as follows:

(1) We first identified callable regions covered by exactly one contig from output of Paftools based on trio binning genome assembly of Canu, as described in PacBio assembly-based structural variant detection. By default, Paftools used assembly-to-reference alignment longer than 10 kb to generate callable regions.
(2) For each individual in twins, we got the union of the regions from each parental haplotype. Then we obtained the intersection of callable regions between twins.
(3) We compared the benchmark calls and PacBio-based assembly structural variants from Paftools in twins through Jasmine with –keep_var_ids and –ignore_strand parameters, and then retained assembly-specific structural variants.
(4) We applied svanalyzer widen command to extend the repetitive genomic coordinates surrounding assembly-specific structural variants, and then added 50 bp on each side of these regions.
(5) Based on the regions obtained in step 2, we removed the regions in step 4. Finally, we constructed the benchmark regions for benchmark set in twins.
(6) The process for constructing benchmark regions in parent was similar to that of twins except for step 2, because there were no biological replicates of the parents.

Validation of small variants benchmark calls by PMRA

We performed 16 replicates for each Quartet reference material on the Applied Biosystem™ Axiom™ Precision Medicine Research Array (PMRA). Genotypes were called by Axiom Analysis Suite v4.0.1.

We selected genotype calls using the following criteria: (1) less than two replicates with missing calls; and (2) more than 80% genotype calls are the same.

The PMRA probes were annotated by hg19, but the reference datasets were mapped based on GRCh38. To avoid converting errors, we only compared variants annotated in dbSNP by dbSNP RefSNP ID.

Validation of structural variant benchmark calls by independent technologies

We validated the structural variant benchmark calls using four Illumina short reads, 10× Genomics linked reads, PacBio CLR long reads, and BioNano Genomics optical mapping. The structural variant datasets corresponding to each technology were generated through the data generation section above. In each technology, the shared structural variants between twins were used for validation of benchmark call structural variants in twins. The structural variants benchmark calls in parents were separately validated by the corresponding structural variant datasets. The validation process used Jasmine with –keep_var_ids and –ignore strand parameters.

We also randomly selected 40 structural variants including 20 insertions and 20 deletions that have not been validated by other technologies and manually checked their accuracy through IGV.

In addition, the datasets from three other independent researches based on long-read sequencing were employed to validate our benchmark calls using Jasmine with –keep_var_ids and –ignore strand parameters.

Training batch-specific machine learning models

Variant quality metrics of Mendelian concordant variants from one D5 replicate for each batch (Additional file 2: Table S10, Batches 5, 6, and7 with three replicates) were used to train one-class SVM classifier (https://scikit-learn.org/stable/modules/svm.html#). For small variants, variant quality, depth, BaseQRankSum, QualByDepth, FisherStrand, SrandOddsRatio, RMSMappingQuality, MappingQualityRnakSunTest, ReadPosRankSumTestg, genotype quality, and membership of dbSNP were used. For structural variants, variant quality, genotype quality, and the raw counts of paired reads supporting alternate allele were used. The three trained models were applied for each batch respectively to classify high-quality variants and low-quality variants.

Variant calling from RNA-seq

Sequences were mapped to GRCh38. We used Sentieon Genomics software to analyze short-read RNA datasets from raw fastq files to VCF files. This workflow was derived from recommended RNA-seq short variant discovery pipeline by the Broad Institute (https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNA-seq-short-variant-discovery-SNPs-Indels-), including read mapping by BWA-MEM, duplicates removing, split reads at junction, base quality score recalibration (BQSR), and variant calling by HaplotyperCaller. Additionally, we employed DeepVaraint with RNA-seq models to call variants (https://github.com/google/deepvariant/blob/r1.5/docs/deepvariant-rnaseq-case-study.md).

Variant detection from LC-MS/MS proteomics

XML file contained peptide identification results generated by an open-source search engine X!Tandem. The software needs to input the Mascot Generic Format (MGF) file, which is the most common format for MS/MS data encoding in the form of a peak list. Then PGA R packages (v1.18.1) were used to identify variant peptides from the XML file.

We constructed custom protein databases from RNA-seq datasets containing SNVs and Indels, then searched the database to detect variant peptides and their corresponding variant's locations on the genome from LC-MS/MS datasets.

Reproducibility

The Jaccard Index represents the concordance of variant detection between two sequencing datasets by measuring the proportion of shared variants detected by both datasets relative to the total number of variants detected by both datasets. Shared variants are defined as variants with the same position and variant sequence, where the "CHROM," "POS," "REF," and "ALT" and "GT" fields in the VCF file must be identical. The Jaccard Index ranges from 0 to 1, with values closer to 1 indicating a higher level of consistency in the variants detected by the two datasets. The formula for calculating the Jaccard Index is as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Precision and recall

Precision is the fraction of called variants in the test dataset that are true, and recall is the fraction of true variants that are called in the test dataset. True positives (TP) are true variants detected in the test dataset. False negatives (FN) are variants in the reference dataset failed to be detected in the test dataset. False positives (FP) are variants called in the test dataset but not included in the reference dataset. Precision and recall are defined as below:

$$Presicion = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

For small variants, we compared variants with benchmark small variants using hap.py with "vcfeval" as the comparation engine (https://github.com/Illumina/hap.py). For structural variants, we merged and compared variants in different callsets using Jasmine with parameters max_dist=1000 –keep_var_ids –ignore_strand. When considering the genotype of structural variants, an additional parameter –output_genotypes needs to be used. When comparing with small variant benchmark calls and structural variant benchmark calls, genotypes of the variants were considered.

Mendelian violation rate of Quartet family

Mendelian violation rate is the number of variants not following Mendelian inheritance laws divided by the total number of variants called among the four Quartet samples. Mendelian violated variants are the variants not shared by the twins or following

Mendelian inheritance laws with parents. When calculating Mendelian violation of small variants, variants on large deletions defined by structural variants benchmark calls were not included, because VBT (https://github.com/sbg/VBT-TrioAnalysis) takes these true variants as Mendelian violations. For structural variants, Mendelian analysis was only done for Quartet-D5, because we could not distinguish homozygotic references and no-call sites. We did not consider genotype information; therefore ,Mendelian discordant variants are variants not shared by Quartet twins or specifically identified in twins but not in parents.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03109-2.

Additional file 1. Supplementary figures.

Additional file 2: Table S1. Data from multiple short-read and long-read sequencing platforms were obtained to detect and validate small variant and structural variant benchmark calls in the Quartet reference samples. **Table S6.** Validation of small variants by PacBio CCS, which are reproducible among call sets and Mendelian consistent in the Quartet family. **Table S7.** Validation of small variants by PMRA, which are reproducible among call sets and Mendelian consistent in the Quartet family. **Table S8.** Statistics of de novo SVs for the Quartet family. **Table S9.** Validation of SV benchmark calls by Illumina short-reads, 10x Genomics, BioNano, and assembly PacBio reads. **Table S10.** Datasets for proficiency test analysis and batch effect analysis [65].

Additional file 3: Table S2. Mapping and calling statistics of short-read sequencing datasets [65].

Additional file 4: Table S3. Statistics of long-read raw sequencing datasets [65].

Additional file 5: Table S4. Mapping statistics of long-read sequencing datasets [65].

Additional file 6: Table S5. Statistics of de novo and somatic small variants for the Quartet [65].

Additional file 7. Review history.

Acknowledgements

We thank the Quartet Project team members, who contributed their time and resources to the design and implementation of this project. We also thank CFFF (Computing for the Future at Fudan) and the Human Phenome Data Center of Fudan University for computing support.

Review history

The review history is available as Additional File 7.

Peer review information

Veronique van den Berghe and Anahita Bishop were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team

Disclaimer

This article reflects the views of the authors and does not necessarily reflect those of the US Food and Drug Administration

Authors' contributions

Y.Z., X.F., S.F., J.W., H.H., and L.S. conceived and oversaw the study. Y.Z., L.D., R.Z., and W.H. prepared the biosamples and coordinated NGS library preparation and sequencing. L.R., X.D., J.Y., Y.G., R.P., Y.L., J.L., Y.Y., N.Z., J.S., F.L, D.W, H.C., L.L.S., L.H., A.S., J.N., W.X., J.X., W.T., X.H., P.J., K,Y., J.M.L., L.J., L.S., H.H., J.W., S.F., X.F., and Y.Z. performed data analysis and/or interpretation. J.Y., L.R., Y.L., and J.S. managed the datasets. R.Z., L.R., R.P., J.M.L, L.S., and Y.Z. coordinated proficiency testing with the Quartet DNA reference materials. L.R., X.D., Y.Z., S.F., H.H., L.S., W.T., J.X., and W.X. wrote and/or revised the manuscript. All authors reviewed and approved the manuscript. Dozens of participants of the Quartet Project freely donated their time and reagents for the completion and analysis of the project.

Funding

This study was supported in part by Shanghai Sailing Program (22YF1403500), the National Natural Science Foundation of China (32300536, 32170657 and 31720103909), National Key R&D Project of China (2018YFE0201600), State Key Laboratory of Genetic Engineering (SKLGE-2117), and the 111 Project (B13016). Some of the illustrations in this paper were created with BioRender.com.

Availability of data and materials

Quartet DNA reference materials can be requested from the Quartet Data Portal (http://chinese-quartet.org) [26]. The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (GSA) (accession number: HRA001859) [60]. Variant calling files are deposited in the European Variation Archive (accession number: PRJEB66342) [61]. Quartet DNA benchmark sets are available at the Quartet Data Portal (http://chinese-quartet.org) [26] and Zenodo at http://doi.org/10.5281/zenodo.10075391 [62].

All scripts used for statistical analyses have been publicly available on GitHub:https://github.com/LuyaoRen/Quartet_DNA (under the GNU General Public License v3.0) [63] and Zenodo: http://doi.org/10.5281/zenodo.10076814 [64].

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board (IRB) of the School of Life Sciences, Fudan University (BE2050). It was conducted under the principles of the Declaration of Helsinki. Four healthy volunteers from a family quartet participated in the Taizhou Longitudinal Study conducted in Taizhou, Jiangsu, China. Peripheral blood samples were collected from these individuals to establish human immortalized B-lymphoblastoid cell lines. All four donors provided their informed consent by signing consent forms.

Consent for publication

Not applicable.

Competing interests

F.L. is an employee of Nextomics Biosciences Institute. D.W. is the co-founder of Nextomics Biosciences Institute. H.C. is an employee of OrigiMed Co., Ltd. L.L.S. is the co-founder of Sequanta Technologies Co., Ltd. L.H. is an employee of Genome Wisdom Inc. All other authors declare that they have no competing interests.

Author details

[1]State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, Shanghai, China. [2]National Institute of Metrology, Beijing, China. [3]National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital, Beijing, China. [4]Greater Bay Area Institute of Precision Medicine, Guangzhou, Guangdong, China. [5]Nextomics Biosciences Institute, Wuhan, Hubei, China. [6]OrigiMed Co., Ltd, Shanghai, China. [7]Sequanta Technologies Co., Ltd, Shanghai, China. [8]Genome Wisdom Inc, Beijing, China. [9]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [10]EATRIS ERIC-European Infrastructure for Translational Medicine, Amsterdam, the Netherlands. [11]Department of Medical Sciences, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [12]Office of Oncologic Diseases, Office of New Drugs, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA. [13]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, USA. [14]Shanghai Cancer Center, Fudan University, Shanghai, China. [15]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. [16]International Human Phenome Institutes, Shanghai, China.

Received: 8 November 2022 Accepted: 13 November 2023
Published online: 27 November 2023

References

1. Turro E, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583:96–102.
2. Flannick J, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature. 2019;570:71–6.
3. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet. 2017;18:473–84.
4. Gargis AS, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol. 2012;30:1033–6.
5. Zook JM, et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol. 2019;37:561–6.
6. Chin CS, et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. Nat Commun. 2020;11:4794.
7. Zook JM, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32:246–51.
8. Huang C, et al. An integrated Asian human SNV and indel benchmark established using multiple sequencing methods. Sci Rep. 2020;10:9821.
9. Zook JM, et al. A robust benchmark for detection of germline large deletions and insertions. Nat Biotechnol. 2020;38:1347–55.
10. Du X. et al. Robust Benchmark Structural Variant Calls of An Asian Using the State-of-art Long Fragment Sequencing Technologies. Genomics Proteomics Bioinformatics (2021).
11. Liu Z, et al. Towards accurate and reliable resolution of structural variants for clinical diagnosis. Genome Biol. 2022;23:68.
12. Goldfeder RL, et al. Medical implications of technical accuracy in genome sequencing. Genome Med. 2016;8:24.
13. Telenti A, et al. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A. 2016;113:11901–6.
14. Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.
15. Tom JA, et al. Identifying and mitigating batch effects in whole genome sequencing data. BMC Bioinformatics. 2017;18:351.
16. Ren L. et al. Genomic reference materials for clinical application, Clinical Genomics, Chapter32. Second edition (2022). Editors: Kulkarni S and Roy S. ISBN: 9780323900249

17. Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet. 2014;15:56–62.
18. Wang X, et al. Rationales, design and recruitment of the Taizhou Longitudinal Study. BMC Public Health. 2009;9:223.
19. Jonsson H, et al. Differences between germline genomes of monozygotic twins. Nat Genet. 2021;53:27–34.
20. Conrad DF, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011;43:712–4.
21. Zheng Y, et al. Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials. Nat Biotechnol. 2023. https://doi.org/10.1038/s41587-023-01934-1.
22. Yu Y, et al. Quartet RNA reference materials and ratio-based reference datasets for reliable transcriptomic profiling. Nat Biotechnol. 2023. https://doi.org/10.1038/s41587-023-01867-9.
23. Tian S, et al. Quartet protein reference materials and datasets for multi-platform assessment of label-free proteomics. Genome Biol. 2023;24:202.
24. Zhang N. et al. Quartet metabolite reference materials and datasets for inter-laboratory reliability assessment of metabolomics studies. bioRxiv (2022).
25. Yu Y, et al. Correcting batch effects in large-scale multiomic studies using a reference-material-based ratio method. Genome Biol. 2023;24:201.
26. Yang J, et al. The Quartet Data Portal: integration of community-wide resources for multiomics quality control. Genome Biol. 2023;24:245.
27. Sedlazeck FJ, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.
28. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
29. Cretu Stancu M, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017;8:1326.
30. Jiang T, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020;21:189.
31. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. 2019;35:2907–15.
32. Toptas BC, Rakocevic G, Komar P, Kural D. Comparing complex variants in family trios. Bioinformatics. 2018;34:4241–7.
33. Pan B, et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. Genome Biol. 2022;23:2.
34. Jia P. Haplotype-resolved assemblies and variant benchmarks of a Chinese Quartet. Genome Biology, accepted (2023).
35. Kirsche M, et al. Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods. 2023;20:408–17.
36. Lecompte L, Peterlongo P, Lavenier D. & Lemaitre C. SVJedi: Genotyping structural variations with long reads. Bioinformatics (2020).
37. Beyter D, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet. 2021;53:779–86.
38. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329–46.
39. Chaisson MJP, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784.
40. Audano PA, et al. Characterizing the major structural variant alleles of the human genome. Cell. 2019;176(3):663-675 e19.
41. Shi L, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun. 2016;7:12065.
42. Daniel E. Cook, A.V., Dennis Yelizarov, Yannick Pouliot, Pi-Chuan Chang A Deep-learning based RNA-seq Germline Variant Caller. bioRxiv (2023).
43. Mansi L, et al. REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. Nucleic Acids Res. 2021;49:D1012–9.
44. Kosugi S, et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology. 2019;20:117.
45. Chen K, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81.
46. 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;21:974–84.
47. Rausch T, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.
48. Cameron DL, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27:2050–60.
49. Qi J, Zhao F. inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic Acids Res. 2011;39:W567-575.
50. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.
51. Chen X, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
52. Erikson GA, et al. Whole-Genome Sequencing of a Healthy Aging Cohort. Cell. 2016;165:1002–11.
53. 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;25:2865–71.
54. 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;17:51–62.

55. Wala JA, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018;28:581–91.
56. Zhang J, Wang J, Wu Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC Bioinformatics. 2012;Suppl 6(Suppl 6):S6.
57. Soylev A, Kockan C, Hormozdiari F, Alkan C. Toolkit for automated and rapid discovery of structural variants. Methods. 2017;129:3–7.
58. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000Res. 2017;6:664.
59. Kronenberg ZN, et al. Wham: Identifying Structural Variants of Biological Consequence. PLoS Comput Biol. 2015;11:e1004572.
60. Quartet Project Team. Raw sequencing data from Quartet Project. Datasets. Genome Sequence Archive. https://ngdc.cncb.ac.cn/gsa-human/browse/HRA001859 (2023).
61. Ren L. Quartet genomics variants. Datasets. European Variation Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB66342 (2023).
62. Ren L. Quartet DNA benchmark sets for germline small variants and structural variants. 2023. Zenodo. https://doi.org/10.5281/zenodo.10075391.
63. Ren, L. Scripts for Quartet DNA Manuscripts. Github. https://github.com/LuyaoRen/Quartet_DNA (2023).
64. Ren L. 2023. Scripts for Quartet DNA Manuscripts. Zenodo. https://doi.org/10.5281/zenodo.10076814.
65. Ren L. Supplementary tables for Quartet DNA manuscript. 2023. Zenodo. https://doi.org/10.5281/zenodo.10076948.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.