BMC Genomics

RESEARCH

Open Access

Interlaboratory evaluation of high molecular weight DNA extraction methods for long-read sequencing and structural variant analysis

Alison S. Devonshire[1*], Jordi Morata[2,3], Claire Jubin[4], Rui Pedro Abreu Pereira[1], Laura Hernandez-Hernandez[1], Dilek Yener[1], Eric Cabannes[4], Steven McGinn[4], Marc Delepine[4], Cédric Fund[4], Raúl Tonda[2,3], Simon Heath[2,3], Marc Dabad[2,3], Javier Gutierrez-Cuesta[2,3], Ignacio Sanchez Escudero[2,3], Maria Cristina Frias-Lopez[2,3], Simon Cowen[1], Alexandra Whale[1], Thorsten Voss[5], Jean-François Deleuze[4], Ivo Gut[2,3], Marta Gut[2,3] and Carole A. Foy[1]

Abstract

Background Long-read sequencing technologies enable resolution of structural variants (SV) and long-range genome assembly, but require high molecular weight (HMW) DNA of both high quantity and quality to produce optimal sequencing results. New DNA extraction methods have been developed but these have not been assessed for use in routine testing. The interlaboratory study described here tested four commonly used methods: Fire Monkey, Nanobind, Puregene and Genomic-tip with a reference cell line containing known chromosomal alterations. Samples were assessed with commonly applied approaches for evaluating DNA purity and integrity as well as a method based on linkage using digital PCR. Sequencing performance was evaluated and the impact of extraction method on structural variant calling investigated.

Results All methods generally produced samples of acceptable purity although yield varied considerably between laboratories. Library preparation and sequencing were successful for all four methods, with Fire Monkey extracts achieving the highest N50 values, Genomic Tip giving the highest sequencing yields and Nanobind, the highest proportion of ultra-long reads (> 100 kb). The dPCR assay with duplexes at 100 kb and 150 kb distances was predictive of ultra-long reads and provides a more quantitative read-out (% linkage) than pulse-field gel electrophoresis (PFGE) which varied in performance between instruments and gel dyes. Neither PFGE nor dPCR were predictive of the proportion of short reads (< 10 kb). Coverage was a key factor in the success of SV calling, but this was dependent on SV caller. Megabase scale SVs were challenging to analyse with SV callers and required confirmation based on coverage plots and mapping of junction sequences, and the findings of earlier studies were only partially confirmed.

Conclusions This study highlights some of the challenges of HMW DNA extraction as well as the need for robust sample QC metrics to ensure optimal sequencing yield and read length which in turn influence the success of SV analysis. dPCR approaches for DNA integrity showed potential but require further development. As long-read

*Correspondence:
Alison S. Devonshire
alison.devonshire@lgcgroup.com

Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

methods are increasingly applied in routine settings such as clinical testing laboratories, cellular reference samples with well-characterised SVs are recommended as controls for the full long-read sequencing workflow.

Keywords Pre-analytical, Nanopore, Genetics, Diagnostics, QC, Standardisation, Digital PCR, Long read sequencing, Structural variation

Background

Long-read sequencing strategies have disclosed new possibilities in the knowledge of genetic information in genomes and facilitating the assembly of complex genomes [1–3]. Long reads enable characterization of structural variants (SV) such as inversions, translocations, duplications, or insertions of mobile repetitive elements [4] and *de novo* genome assembly of diverse plant and animal species [5–7]. Combining long-read technologies and other approaches has allowed previously intractable regions of the human genome to be unveiled, resulting in a new, complete reference sequence of a human genome [8].

Long-read sequencing technologies from Oxford Nanopore Technologies (ONT) and PacBio have led this 'third revolution' [9], with both approaches generating reads from single DNA molecules without the need for amplification. By controlled, continuous flow of a single-strand of a DNA double helix through a membrane-bound flow-cell, ONT flow-cells (FCs) are capable of generating sequence average read lengths ≥ 20 kb, with maximum reads in the "ultra-long" (> 100 kb) and "whale" scales (> 1 Mb) being observed [10, 11]. Alternatively, PacBio single molecule real-time (SMRT) sequencing utilises single-stranded circular structures ("SMRTbell"), which typically have insert size of 10–25 kb [12]. This size enables multiple rounds of sequencing the same molecule ("circular consensus sequencing"), which produce "HiFi" reads with high raw read accuracy of 99.9% [13]. Both techniques require large input quantities (µg-scale) of high-molecular weight (HMW) genomic DNA (gDNA) [14, 15]. Extraction of HMW gDNA suitable for long-read sequencing has been highlighted as a limitation for these sequencing strategies [16]. Starting library construction with high-quality gDNA is vital and deviations from typical manufacturers' recommended values (for example, fragment size > 40 kb) [12] can lead a reduction in library preparation efficiency [17].

Although some recent publications use traditional phenol/chloroform methods for long-read sequencing applications [3, 18], these chemicals are strong oxidizers and may lead to DNA damage [12] that may impact results as well as there being safety implications of using such approaches. Therefore, they may not be as well-suited to high-throughout, reproducible sample processing for applied testing such as clinical diagnostics. Consequently, alternative methodologies have been developed to facilitate the extraction of gDNA suitable for long read sequencing. This study investigated approaches to assessing the performance of alternative extraction methodologies through inter-laboratory assessment and testing their fitness for purpose in the application to human germline SVs analysis using nanopore sequencing. Established and novel sample QC methods for gDNA fragment length were evaluated and compared to sequencing metrics, as a means of developing standard approaches to measure sample quality and extraction performance.

Results

DNA extraction yield and purity

The performance of four HMW gDNA extraction kits: Nanobind CBB Big DNA (*NB*) kit (Circulomics, Baltimore, MD, USA), Fire Monkey kit (*FM*) (Revolugen, Hadfield, UK), Gentra Puregene Cell (*PG*) kit and Blood & Cell Culture DNA kit with Genomic-tip 20/G (*GT*) (QIAGEN, Hilden, Germany) were compared in an interlaboratory study (Fig. 1). A cell line sample, GM21886, with known copy number alterations [19, 20] was selected to test the impact of extraction methodology on a complete long-read sequencing workflow including SV analysis. Four laboratories of varying experience with the methods received the same sample containing cryopreserved GM21866 cells. Due to failed extractions ($n = 2$) with one kit (Table S1), Laboratory 1 repeated extractions for all four methods with additional GM21866 cells with an increased input per extraction (5 vs. 3.3 million cells). Figure 2 shows the yield and purity of isolated DNA from all extractions. Yield varied between an overall median 0.9 to 1.9 µg/million cells for the four kits (PG < GT < FM < NB) (Fig. 2A). A significant degree of between-laboratory variation was observed ($p < 0.001$), which showed an interaction with extraction method ($p < 0.001$), with no overall differences in yield between kits ($p = 0.077$). The NB kit had the most consistent yield (interquartile range (IQR) of 1.1 µg/million cells), with greater variation in yield observed for the other three methods (IQR of 1.7–1.8 µg/million cells for FM, PG and GT respectively). An average concentration of > 50 ng/µL was obtained for slightly over 50% of the extractions (26/48 samples) (Figure S1), which corresponded to the minimum input for the short fragment removal step (Short Read Elimination (SRE) kit, Circulomics) (3 µg) using a standard input volume of 60 µL. Comparing the fluorimetric measurements of triplicate aliquots from the top, middle and bottom of the sample tube (Figure S2), higher variability was observed for the GT, NB and PG extracts, indicating

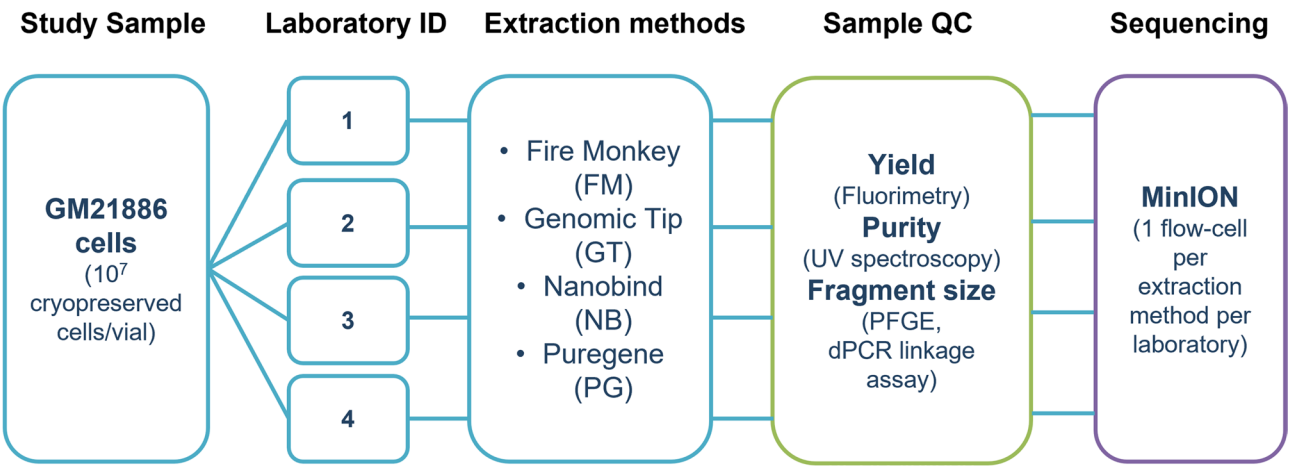

Fig. 1 Interlaboratory study design. All four laboratories (designated 1–4) isolated DNA from GM21886 cells with four alternative extractions methods as shown ($n = 3$ per method per laboratory using 3.3×10^6 cells input). DNA extracts from the same method underwent sample QC and were pooled for library preparation. Each library was sequenced using a single MinION flow-cell (except Laboratory 4 for FM and GT methods due to failed extractions). In addition, GT and PG extractions were repeated with an alternative source of cells (using 5×10^6 cells input, $n = 2$) at Laboratory 1. Laboratory 1 also sequenced combined extracts from each method using a PromethION FC at Laboratory 1

greater within-sample heterogeneity, compared to the FM samples.

DNA purity was assessed by UV spectrophotometry, with A260/280 and A260/230 wavelength absorbance ratios indicating protein and organic component contamination levels respectively. For sample concentrations >10 ng/µL, most A260/280 ratios were above the optimal level of ≥1.8 (Fig. 2B) with no differences observed between kits however differences between laboratory were indicated ($p < 0.05$). A260/230 ratios differed between kits ($p < 0.05$) and laboratories ($p < 0.05$) (Fig. 2C), with a variable proportion of extracts showing values <2.0 (10%, 25%, 33% and 45% of samples for GT, FM, NB and PG respectively).

DNA fragment size

DNA fragment size was assessed by pulsed field gel electrophoresis (PFGE) and dPCR-based linkage assay [21] (Fig. 3; Figure S3). PFGE indicated that all four extraction methods showed evidence of HMW DNA with the majority of the gel smear exceeding 50 kb (Fig. 3A). The NB extracts were characterized by a more prominent peak >80 kb, whilst the intensity of the gel smear for the other three extraction methods between 30 and 80 kb was the dominant 'peak' (Fig. 3A). For the PG extracts, it varied between sites whether a clear HMW (> 80 kb) peak was visible, with the Laboratory 3 individual and Laboratory 4 pooled PG samples showing a clearer peak than the PG samples (Figure S3B-C) at Laboratory 2 (Fig. 3A).

The percentage of linked (intact) molecules were measured by dPCR with duplex assays positioned at five different distances from each other: 33, 60, 100, 150 and 210 kb [21], which provides an indicator of the proportion of molecules which exceed this fragment size. Linked molecules segregate in a single dPCR partition and are observed in the double-positive cluster (Fig. 3B). Differences in observed linkage between extraction methods were observed ($p = 0.018$). Compared to the FM kit, NB samples demonstrated a significantly higher proportion of linked molecules at all linkage distances ($p = 0.0003$). For example, at 33 kb distance, NB samples showed a linkage range of 40–65%, followed by PG samples (15–60%), with similar results for GT (25–50%) and FM (20–40%) (Fig. 3C). Evaluating the 150 and 210 kb linkage distances as indicators of the proportion of very HMW DNA in the samples, at 150 kb, the median percentage (and range) of linked molecules were 4% (1–11%) in NB samples, 3% (0.7–8%) for PG, 0.4% (0 to 4%) for GT and 0.7% (0–3%) for FM samples (Fig. 3D).

Sequencing performance

Fourteen gDNA samples underwent SRE kit-based size selection followed by library preparation, corresponding to one library for each of the four extraction methods at Laboratories 1, 2 and 3, and for the Laboratory 4 NB and PG extracts. Laboratory 4's FM and GT extracts were not further processed due to limiting yield (Table S2). Extraction yield affected the amount of DNA available for the SRE step, with input ranging from (5.1–15.5 µg). DNA yield from SRE varied between laboratories, with a mean recovery of 60.7% ± SD 25.7%. Input to library preparation varied between 0.9 and 5 µg and resulted in a mean yield of 1.6 µg (range: 0.45–3.54 µg), corresponding to relative recovery rate of 56.7% ± 12.9%.

Each library was sequenced using one MinION flow cell with three FC loadings over a period of 72 h, with the exception of Laboratory 2 and 4 where the run period

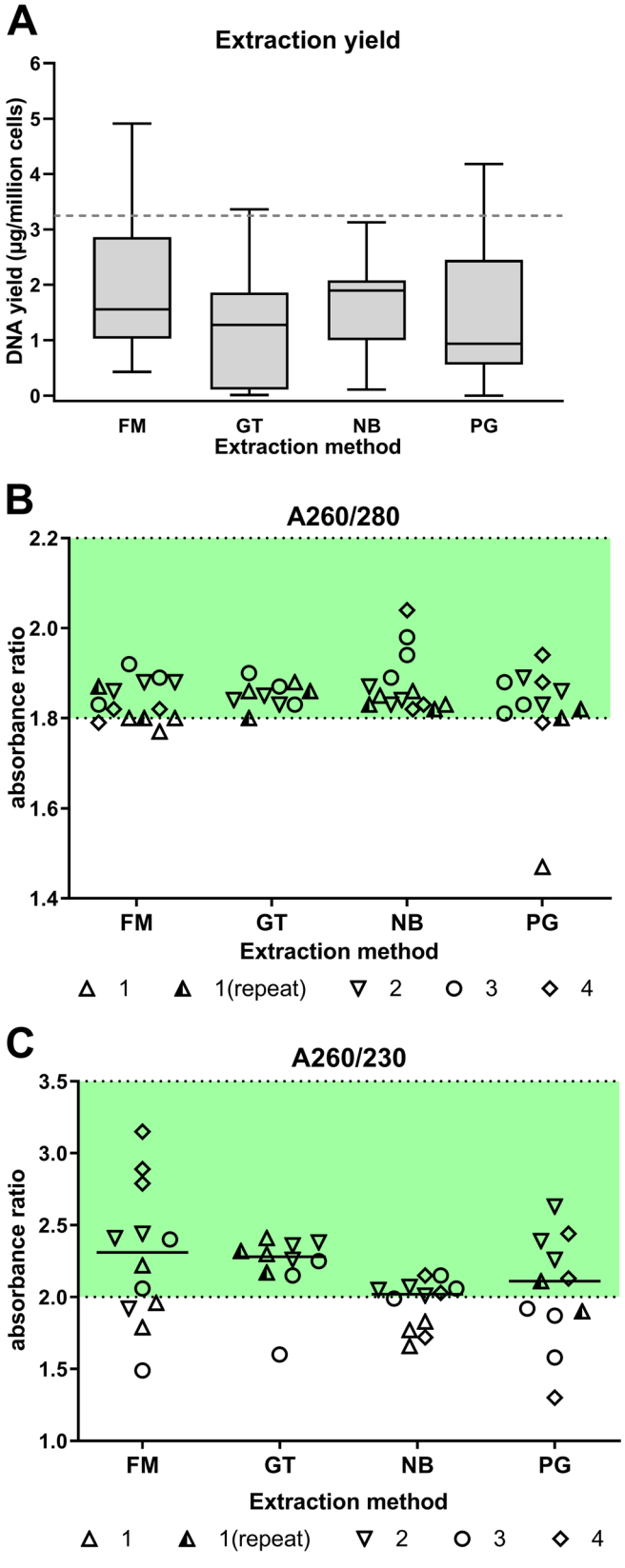

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 DNA sample yield and purity. **A** Yield based on fluorimetric assay for interlaboratory study and repeat performed at Laboratory 1 ($n = 14$ samples total). Boxplots show median (line), interquartile range (box) and range (whiskers). Dashed line indicates 50% of the theoretical yield based on cell input number. **B**, **C** A260/280 and A260/230 absorbance ratios are shown for all samples with concentration (based on Qubit) > 10 ng/μL. Datapoints reflect individual samples according to laboratory. Dashed line/green shading reflects acceptable A260/280 and A260/230 values of > 1.8 and > 2.0 respectively. Note: Four data points (Laboratory 1 repeat FM and NB kits) with A260/230 values > 4 are not shown due to a suspected blanking error

was ~ 110–115 h, with a fourth FC loading in some cases (Table S2). In addition, four libraries corresponding to each extraction method were sequenced together on a PromethION flow cell at Laboratory 1. Raw data was base-called with *guppy* in fast, methylation, high accuracy and super high accuracy modes. Detailed yield and quality metrics per flow cell are listed in Table S3, with the results of analysing super high accuracy mode-called data being shown hereafter. All MinION sequencing experiments had a mean Qscore ≥ 13, with the pass reads (reads with Qscore ≥ 7) representing a mean 86% (SD ± 3%) of the total number of reads. The yield per MinION flow cell differed between extraction kits ($p < 0.01$), GT libraries had the highest yield (mean 22.9 ± SD 3.7 Gb), followed by FM (18.0 ± 3.8 Gb), NB (15.4 ± 5.3 Gb) and PG (7.3 ± 2.1 Gb) (Fig. 4A). The number of available pores per MinION flow cell varied markedly (835 to 1564 pores) (Table S2) but this variation did not explain the sequencing yields observed. In terms of the number of reads per MinION flow cell (Fig. 4B), no significant differences were found between extraction methods. A similar pattern to that for total yield was observed for the FM, GT and NB kits, with GT libraries also generating the highest total read count value (mean 1.0 ± SD 0.11 million reads). However, PG libraries showed higher interlaboratory variation in read count (63% CV) compared to total yield (29% CV), suggesting read length distribution was more variable between sites.

Mean read length, read N50 (median read length which corresponds to 50% sequencing yield) and maximum read length of quality-filtered and aligned reads were determined for each dataset (Table S3). Mean read length values per flow cell ranged from 6.3 kb (Laboratory 2, PG) to 26.9 kb (Laboratory 3, FM), with ten out of the 14 libraries sequenced presenting a mean read length within the 20–25 kb interval. In all samples, read N50 values were lower than the mean read length due to an asymmetric read length distribution, with higher differences between the two metrics observed in NB and PG samples. There were no overall significant differences between kits ($p = 0.06$); all three FM sequenced libraries' N50 values were in the highest quartile of the 14 libraries, with 23.2 kb being the highest N50 value in the study (Fig. 4C). Four libraries presented N50 values < 15 kb and were associated with PG (2.6 and 9.3 kb) and NB (10.7 and 11.7 kb) samples (Fig. 4C).

The proportion of reads of different size brackets varied with kit ($p < 0.01$). Aligned reads from NB and PG libraries showed a higher proportion of short reads < 10 kb, whilst the largest fraction of aligned reads for the FM and GT kits corresponded to the 20–50 kb read length bracket (Fig. 4D). The highest percentage of aligned reads of 100–200 kb were detected in NB libraries (Fig. 4D), which was consistent with the maximum read lengths in the study of > 300 kb being observed in three of the four NB libraries (Table S3).

Structural variant analysis

Sequencing data was merged according to extraction method for SV analysis using balanced data from Laboratories 1–3 (3 MinION FCs total) as there were missing results for the FM and GT methods in the case of Laboratory 4. In addition, the data for all four extraction methods (12 MinION FCs) were pooled to generate a high coverage 'truth-set' ("ALL"), since an independent truth-set does not exist for this cell line material. The number of SVs called for each extraction method was compared for alternative SV programs, CUTESV versions 1 and 2 [22, 23] and SNIFFLES version 2 [24] (Fig. 5A). In the ALL dataset, the total SV count was similar for both versions of CUTESV and SNIFFLES, with ~ 24,000 SV called, of which 31% were designated "imprecise" with SNIFFLES2 (only precise SVs were called by CUTESV) (Supplementary Table 4A). The majority of SVs were insertions (57–58%) and deletions (41%), whilst more inversions and duplications were found by CUTESV (657 variants) than SNIFFLES (89 variants) in the high coverage ALL dataset (Table S4A). Comparing the SV count in the smaller extraction method-level datasets, the ranking of SVs count followed the order GT > FM > NB > PG, however only marginal differences (< 500 SVs) between GT/FM/NB datasets were observed for SNIFFLES, whilst for CUTESV (version 2), ~ 3500, 5200 and 16,900 fewer SVs were called for the FM, NB and PG datasets respectively compared to GT (Fig. 5A, Table S4A).

As there is no gold standard truth-set of SVs for this cell line, the alternative SV-callers were compared in terms of their concordance with the high coverage 'ALL' dataset (Table S4B): concordance between the two versions of CUTESV was 91–94% (i.e. up to 9% of variants unique to each SV caller), and 82-86% between CUTESV and SNIFFLES2 (i.e. up to 18% of variants unique to each SV caller). To investigate the occurrence of putative false negatives and false positives, each extraction dataset was compared with the ALL truth-set (Table S4B). Recall (sensitivity) followed

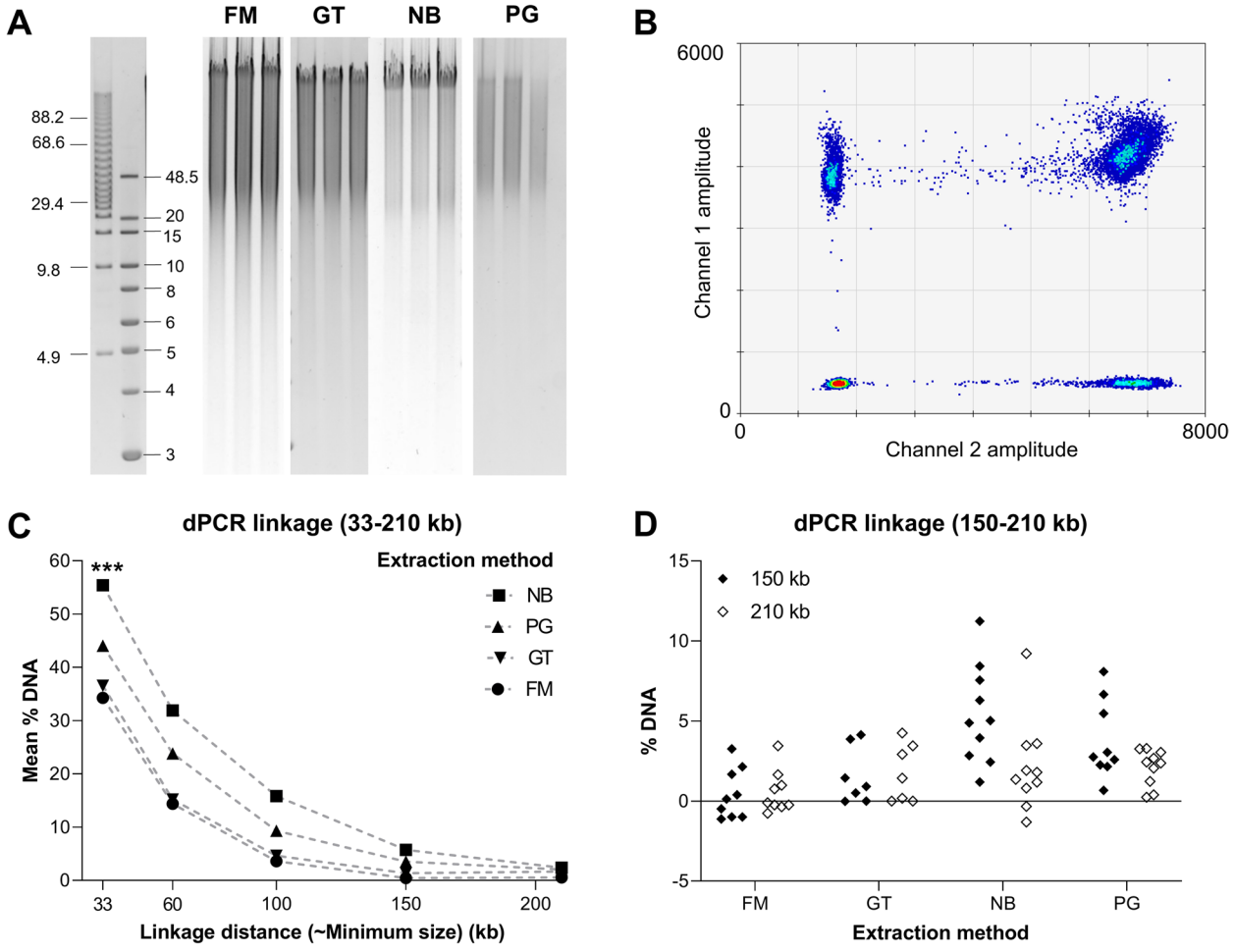

Fig. 3 DNA fragment size analysis. (**A**) A typical PFGE electropherogram of DNA samples from the four tested extraction methods. Samples were analysed using the Pippin Pulse system at Laboratory 2. Further PFGE analysis is shown in Figure S3. (**B**) Typical dPCR analysis of linked molecules with a higher occurrence of double-positive (FAM+, HEX+) partitions (~6000) than would be predicted based on random segregation (~300 double positive predicted based on 2700 single positive population and 22000 double negative partitions). Example is for a NB method sample for the 33 kb linkage assay. (**C**, **D**) Mean % linkage for (**C**) all distances (**D**) only 150 and 210 kb for four tested extraction methods based extracts which underwent MinION sequencing (see Methods). ***p=0.0003 compared to the FM kit (reflecting all linkage distances).

the same pattern as SV count, with the highest proportion of false negatives in the PG dataset, particularly in combination with CUTESV (Fig. 5B). Although it is not possible to definitively class SV unique to each extraction method dataset as false positives, precision (technical positive predictive value) metrics were ≥89% for FM, GT and NB, similar to that observed when comparing the SV sets from the alternative SV callers, however precision was 69% for PG with CUTESV(2)-SV calling which could indicate the occurrence of putative false positive calls (Fig. 5B). SV count was found to be proportional to coverage using CUTESV, whilst SNIFFLES was more robust in respect to this factor (Fig. 5C-D); however fewer duplications were found by SNIFFLES (27, 26, 22, 25 for FM, GT, NB and PG respectively) compared to CUTESV (252, 327, 217, 106 for FM, GT, NB and PG respectively) (Fig. 5E-F/Table S4A). SV length distribution was also compared between the extraction methods however no association was observed (Figure S4).

Chromosomal rearrangement analysis

Very large SVs on chromosomes 1, 18 and 22 of between 200 kb and 25 Mb in size, for which the cell line was annotated [19] and which were previously confirmed by short-read sequencing [20], were analysed using reference genome alignment (Table S3) and assembly approaches (Table S4C). Two datasets were analysed: the interlaboratory "ALL" combined MinION dataset (data from 12 FCs, 60X mean read depth) and data from a single PromethION flow cell with DNA extracts from all four methods performed at Laboratory 1 using the same approach of multiple FC loadings applied to the MinION FCs (resulting in 15X mean read depth) (Table S2).

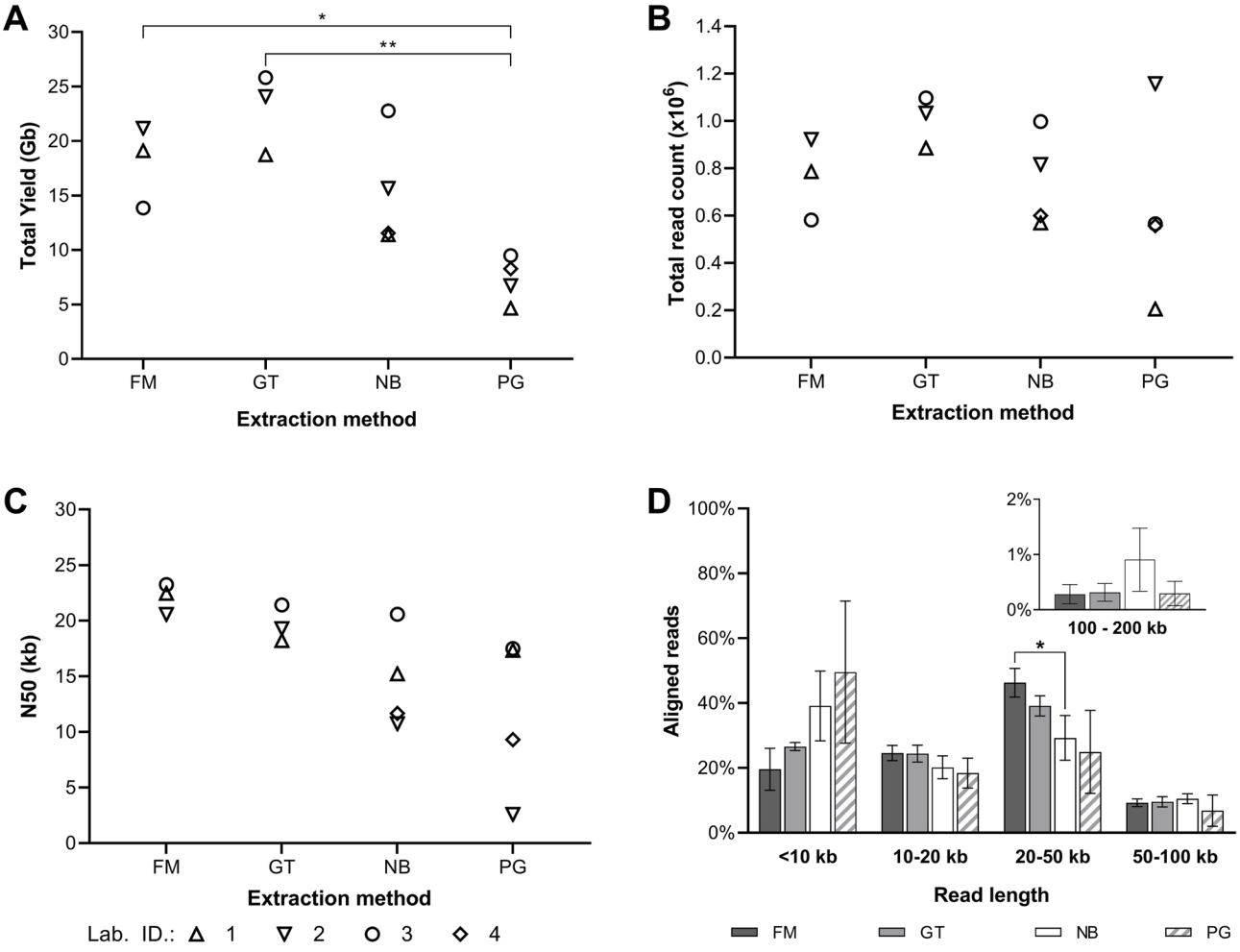

Fig. 4 Sequencing performance quality metrics (MinION). **A** Data yield **B** Total read number **C** N50. Individual datapoints are shown according to extraction method and laboratory which prepared the sample. **D** Read length distribution according to extraction method ($n = 3$ FM and GT, $n = 4$ NB and PG). Mean and SD are shown. *$p < 0.05$; **$p < 0.01$ for comparisons between pairs of extraction methods as shown

MinION sequencing provided better results than PromethION sequencing using alignment approaches (based on coverage, SV-calling with SNIFFLES version 1/2 [24, 25], or analysis of split read alignment pattern) (Table 1); this is likely due to the PromethION sequencing analyses being penalized by lower read depth; the commonly used standard read depth coverage is 30X [26, 27]. The ~210 kb copy number gain on chr1 was not detected based on coverage graphs and automated softwares in either MinION or PromethION datasets (Table 1), however a 1.37-fold-increase in read depth (lower than the theoretical 1.5-fold) was calculated between coordinates observed upon manual inspection of the region in the MinION dataset (Figure S5A(c)) and the junction sequence was found as split reads in the MinION dataset (Figure S5A(b-c)). Large 15 Mb and 25 Mb deletions at the *p*- and *q*-termini of chromosome 18 were confirmed based on coverage in both datasets (Table 1, Figure S5B(a)). Further evaluation was required to confirm the smaller 0.5 Mb deletion on the q-arm of chr18, by analysing split reads aligning to the regions either side of the putative deleted region (Figure S5C). This SV was also called in analysis of both datasets using SNIFFLES version 1 or 2 whilst the other large SVs were not (Table 1). Evidence for one copy of chr18 being a ring chromosome was found in reads aligned to the 3' end of the 15 Mb deletion and the 5' end of 25 Mb deletion in both datasets (Figure S5B(b)-(c)), whereas, as expected, circularization split reads pattern was not observed in analysis of the telomeric regions containing the non-deleted regions at the extreme 5' and 3' ends of chr18 (data not shown). The previously characterised copy number gain on chromosome 22 was not confirmed in the current study's long read datasets (both MinION and PromethION); this SV was found to overlap with a region containing multiple satellite DNA elements

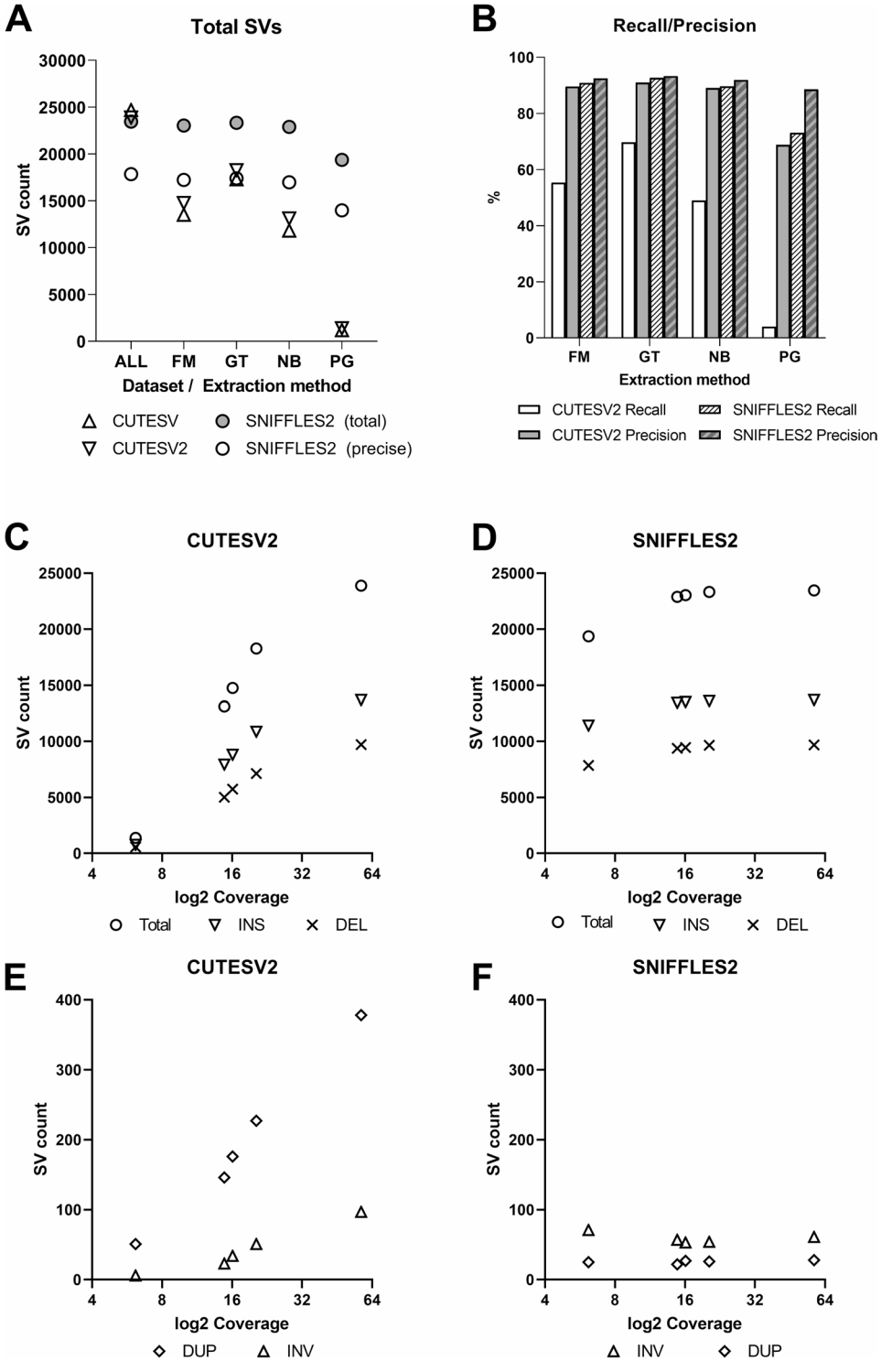

Fig. 5 SV analysis. The number and type of SVs are compared for all pooled data ("ALL") and pooled data for each extraction method ($n = 3$ MinION FCs) with two SV callers: CUTESV (version 1 or 2) and SNIFFLES version 2. **A** Total SV count for CUTESV versions 1 and 2 and SNIFFLES2. For SNIFFLES2, counts of precise and total (including imprecise) SVs are shown. **B** Recall and precision based on comparison with the ALL dataset. SV in the extraction method dataset which were present in the ALL dataset were designated true positives (Recall) whereas SVs which were absent from the ALL dataset were designated false positives (Precision). **C**-**F** SV count compared to coverage of the dataset shown for (**C**, **D**) total SVs and insertions and deletions or (**E**, **F**) inversions and duplications (average read depth) as called by CUTESV2 (**C**, **E**) or SNIFFLES2 (**D**, **F**) (i.e. all variants are included, regardless of whether they were concordant with the 'ALL' dataset)

Table 1 Analysis of GM21886 cell line-annotated large SVs using alignment and assembly approaches

A. MinION merged dataset (ALL)

	SV	Start	End	LEN	Coverage graph	Sniffles 1	Sniffles 2	Ragoo	Junction	Start refinement	End refinement
chr1	DUP	237,067,989	237,277,853	209,864	No	No	No	No	Yes	237,060,971	237,288,156
chr18	DEL_circ	15,000,001	54,944,892	40,428,395	/	/	/	/	Yes	ambiguous	54,938,121
chr18	DEL	10,000	15,000,001	14,990,001	**Yes**	No	No	No	/	No change	ambiguous
chr18	DEL	54,944,892	80,257,174	25,312,282	**Yes**	No	No	No	/	54,938,121	No change
chr18	DEL	54,292,719	54,816,820	524,101	No	No	**Yes**	No	Yes	54,293,057	54,809,784
chr22	DUP	18,794,020	19,019,471	225,451	No	No	No	No	No	Not applicable	Not applicable

B. PromethION merged dataset

	SV	Start	End	LEN	Coverage graph	Sniffles1	Sniffles2	Ragoo	Junction
chr1	DUP	237,067,989	237,277,853	209,864	No	No	No	No	No
chr18	DEL_circ	15,000,001	54,944,892	40,428,395	/	/	/	/	Yes
chr18	DEL	10,000	15,000,001	14,990,001	**Yes**	No	No	No	/
chr18	DEL	54,944,892	80,257,174	25,312,282	**Yes**	No	No	No	/
chr18	DEL	54,292,719	54,816,820	524,101	No	**Yes**	No	No	Yes
chr22	DUP	18,794,020	19,019,471	225,451	No	No	No	No	No

Approaches based on alignment (coverage graph, SV calling using SNIFFLES version 1 or 2, detection of a junction sequence as a split read) or assembly (Flye followed by Ragoo SV calling) were used for detection of the listed SVs. Where the junction sequence was detected, the start and end of the SV coordinates are shown when it allowed refinement (MinION dataset only). Key: **Yes**, SV observed using the approach; No, SV not observed using the approach. For further information, see Figure S5

which is probably the source of a false positive duplication call in NGS analysis (Figure S5D).

In addition, an assembly approach using Flye and Ragoo [28, 29] for identification of the large SVs was tested. Assembly metrics exhibited a very good contiguity for the two datasets with N50 of 37.8 Mb for MinION dataset and 25.7 Mb for PromethION dataset (Table S4C). However, Ragoo did not permit to call any large SVs in the GM21886 cell line (Table 1). Further investigation of the presence of the SVs in assemblies were conducted using junction reads identified with help of mapping information but the identification of contigs containing the whole continuous junction read failed. Despite Flye not being expected to provide diploid assemblies, it was explored if there were contigs containing allele sequences of the heterozygote duplication (chr1) or smaller deletion (0.5 Mb in chr18) but only contigs with the partial sequence or the sequence not duplicated were found in the case of chr1 (Fig. S5A(d)), and the sequence of the junction read for chr18 deletion was partially observed on two separate contigs (Fig. S5C(d)). The rearrangements appeared not to have been assembled into contigs.

Discussion

In parallel with rapid developments in sequencing technologies, recent years have seen the development of new methods targeting production of HMW DNA extraction intended for long-read whole genome sequencing (WGS) and other long range genome characterization techniques. In this study, recently developed methods for HMW DNA extraction using a paramagnetic disc (NB) [30] and novel spin column chemistry (FM) [31] were compared with established methods based on precipitation (PG) [32] and ion exchange chromatography (GT) [33] within the context of an interlaboratory study. Although HMW DNA extraction methods have been evaluated for a range of applications such as metagenomics, there have not been comparison of methods for human genomics applications. The choice of extraction methods investigated here was shown to have an impact on the DNA fragment size profile, sequencing yield, read length profile and SV detection rate, underscoring the importance of selecting the optimal protocol and testing its performance with the intended sequencing method [34]. For clinical applications such as rare disease diagnosis and genetic analysis, it is essential to implement validated methods and appropriate QC criteria for performance monitoring [35]. The Food and Drug Administration (FDA) has developed guidance for design, development and validation of NGS in vitro diagnostics for Germline Diseases [36] and the Clinical and Laboratory Standards Institute (CLSI) has developed guidance for validation and quality management of clinical tests based on NGS [37]. In other fields such as forensics, guidance for quality assurance of DNA testing have been developed [38]. Such guidance documents may assist in development, validation and QC of long-read WGS assays.

Key validation criteria for extraction include DNA yield (quantity), purity (absence/minimal concentration of potential downstream reaction inhibitors) and fragment size profile as well as sequencing metrics such as yield and N50. Although there was some interlaboratory variation in QC and sequencing steps within the current study (Table S1), this provides information on the robustness and reproducibility of the methods in question. Although better precision may be demonstrated in a single laboratory closely following a series of standard operating procedures (SOPs) with a homogeneous control material, the level of interlaboratory variation reflected in the

current study may better reflect the magnitude of variation which arises from processing heterogeneous samples such as clinical specimens.

It is important to consider both magnitude and consistency in DNA yield: in our interlaboratory study, high variability in DNA yield was observed (CV of 47–95%) compared to, for example, extraction cfDNA from plasma performed at a single laboratory [39], with laboratory being main source of variation. Longer multistage protocols such as the GT [40] may be more susceptible to between analyst effects/user error, compared to faster methods such as FM and NB. Therefore validation experiments should include both sufficient replication and testing by alternative analysts [41]. Improved yield and consistency may be obtained with higher cell input, particularly for precipitation-based approaches such as PG, as was observed in repeat experiments performed by Laboratory 1 with 5 million vs. ~3 million cells. Sample purity based on UV absorbance has been shown to indicate the presence of contaminants which are associated with library preparation efficiency [42]. Sample purity was within optimal ranges in most cases for all four extraction methods tested here, where sample yield was acceptable (i.e. not failed extractions with close to zero yield), however more variability was observed in A260/230 ratios. Indeed, in general the kits with the highest proportion of A260/230 results < 2.0 (NB, PG) also showed the lowest sequencing yield and read count. The NB library from laboratory 1 (average A260/230 = 1.75) also had the lowest yield and read count of the four NB libraries sequenced, suggesting that carryover of organic extraction kit components via A260/230 ratios may be a factor which should be monitored as a potential cause of pore blocking, as suggested by Nanopore users [43].

Whilst optimal preparation of HMW gDNA is important for long-read applications, library preparation and sequencing will clearly also influence the final results, therefore comparing sequencing quality metrics using DNA from a particular extraction method with manufacturer's or other examples of optimal outputs for the specific library preparation and sequencing chemistries is advised. In our study which used ligation-based adapter addition, N50 values of 10–20 kb were observed, with the FM and GT kits being the more consistent with N50 > 20 kb and mean N50 = 20 kb respectively. This is similar to other reports using the ligation-based adapter addition approach [44]. Transposase-based adapter addition (which is used in the ONT Ultra-Long DNA Sequencing Kit) has been shown to generate higher read lengths and N50 values of 50–100 kb [3, 45]. However, other workflows building on one of the extraction methods used here (NB) together with ligation-based adapter addition have been developed which prove to be efficient for analysis of the majority of types of variants in human cells [46]. One step in this workflow which is different to the current study is the use of shearing to produce a more homogeneous fragment size distribution with a peak of ~ 30 kb; this makes more molecules available for passing through nanopores and a higher N50 ~ 30 kb was achieved.

As both extraction and sequencing reagents are evolving rapidly, it is important to utilise QC methods to monitor pre-analytical performance, however QC processes and metrics themselves may need to be adapted for working with HMW DNA. Fluorimetry (normally Qubit (Thermo Fisher Scientific)) is the gold standard in DNA quantification for long-read workflows, however high variability in fluorimetric replicates measurements of the same sample was noted for some the HMW gDNA samples which may lead to inaccurate input into PFGE as well as downstream SRE and library preparation. A recent protocol for shearing of HMW gDNA samples prior to Qubit measurements has been developed and recommends the use of human gDNA standards, as opposed to the Qubit lambda DNA standard [47].

Assessment of DNA fragmentation are also important in determining DNA quality for long-read sequencing and the four extraction methods tested here showed evidence of the majority of DNA fragments exceeding 50 kb. PFGE is considered the gold standard for assessing HMW DNA fragmentation, however interpretation of results can be subjective. Klingstrom et al. have proposed the characteristic fragment length (main peak) and the smear ratio (the proportion of DNA with a fragment length of less than 500 bp) [48]. However these metrics are likely to be influenced by the PFGE method and, although the current study suggests that PFGE results can vary considerably between instruments (such as Pippin Pulse and Agilent Femtopulse), it is difficult to make definitive conclusions due to no common samples being run on the various instruments. Profiles appeared to vary with the use of alternative DNA stains, with fluorescent dyes giving worse peak resolution than ethidium bromide. Relative to the HMW main peak, visualisation of lower MW smear is challenging due to low signal intensity, making it difficult to predict the occurrence of shorter fragments which may lead to a reduction in sequencing yield.

dPCR-based analysis of linked molecules [21] constitutes a novel quantitative approach to characterisation of DNA integrity, with this study being the first to apply this approach to evaluation of HMW gDNA extracts for long-read sequencing. Similar to the previous authors, differences in proportion of linked molecules were observed between alternative extraction methods; all four methods in our study produced a higher percentage of linked molecules compared to standard silica column extraction used by Regan et al. [21], which confirms that the

specialised HMW extraction methods outperform standard methods. The profile of the extracts in this study were similar to those isolated using the PrepFiler Forensic DNA Extraction Kit (Life Technologies) by Regan et al., although the average linkage in our study at 60 kb distance was somewhat lower than 42% observed by Regan et al. This may be attributable to DNA degradation during sample transport between laboratories which was necessary for our study. The highest percentage of linked molecules, suggesting the most intact gDNA, were obtained with the NB kit in our study particularly at higher linkage distance, which correlated with the maximum read lengths obtained by nanopore sequencing. However, the higher proportion of linked molecules observed for the NB and PG kits did not generally mirror the read length profile obtained by sequencing, with a higher proportion of short reads < 10 kb obtained by these kits. This may be due to longer fragments being processed less efficiently in passage through the nanopore channels, leading to pore blocking [43]. Alternatively, dPCR-linkage provides a minimum fragment size indicator, as opposed to an average fragment size (as electrophoretic methods do), therefore it may be that minimum size thresholds are not optimal to predict sequencing performance. Additional linkage distances lower than the 33 kb distance which was the minimum size in this study are however likely to be necessary to interrogate the presence of smaller fragments. Also combining 'bins' between linkage markers (e.g. calculation of the %linkage at 60 kb minus %linkage at 33 kb) markers [49] may reflect the physical size range more effectively than a minimum size indicator. While this study followed the calculation model developed by Regan et al., which accounts for two genomic dPCR targets coinciding in the same dPCR partition by chance as well as due to linkage, other mathematical models have recently been proposed for calculation of the proportion of intact DNA which also provide an estimate of the error (measurement uncertainty) in the value [49], which may refine the fragment-size measures produced by this approach further.

As well as ensuring QC methods and metrics are reliable indicators, the availability of appropriate QC materials is also an aspect which supports translation of long-read based workflows into routine laboratory testing. While whole cell materials and synthetic clinical matrices are available for metagenomic applications [18, 34, 50], QC samples for human genomics applications are lacking. Samples such as cell line suspensions like that used here or pooled clinical material such as PBMCs or whole blood could be used. Where possible, QC samples should be matrix-matched to test samples as different potential inhibitors may be present which could impact assay performance in different ways. If QC materials are not matrix-matched to test samples then commutability and fitness for purpose of such materials needs to be demonstrated. In terms of internal controls, indicators of extraction efficiency can include spike-in approaches, which have been applied to cell-free DNA extraction [51]. Development of spike-in methods may prove useful in QC of HMW DNA extractions; however cellular materials are needed to mirror the processes which cellular samples undergo such as cell membrane/cell wall lysis, as well as methods to quantify the spike-in post-recovery.

Cell lines and other well-characterized samples are also an important source of reference data for evaluating the performance of later bioinformatic steps in a human genomic workflow [52], such as alignment and variant calling [53, 54]. Although the reference cell line analysed here was previously characterized by karyotyping, microarray and short read NGS in terms of large (100 kb to Mb scale) CNVs/chromosomal rearrangements [19, 20], a genome-wide SV reference dataset was not available. Therefore pooled data with high coverage was used as a reference to evaluate the impact of extraction method on SV calling. These results showed that the differences in yield associated with extraction method also influenced SV calling success rate due to resultant differences in coverage [55], however this effect was dependent on SV caller, with CUTESV more influenced by coverage than SNIFFLES2, highlighting the need to compare SV callers as part of developing a bioinformatic pipeline. In the high coverage pooled data, the number of insertions, deletions and inversions identified were similar between CUTESV and SNIFFLES. More duplications were found by CUTESV2 than SNIFFLES2 in our study, though it is not possible to determine definitively if these are true or false positives. Another report utilizing earlier versions of CUTESV and SNIFFLES showed lower accuracy (both recall and precision) for duplication-calling by CUTESV [56]. Recent validation of CUTESV version 2 demonstrated a higher rate of duplication calling compared to SNIFFLES2, as in the current study, and although recall (rate of consistent genotype calling) appeared to be improved, an appreciable rate of inconsistent (False positive) calls with the ground truth was apparent [23].

Through the process of verifying the large CNVs/chromosomal rearrangements annotated in reference cell line GM21886, this study highlights some of the challenges to current long-read SV calling methods. SV callers using either mapping or assembly approaches did not identify large SVs, with the exception of sporadic detection of a ~ 0.5 Mb deletion in chr18 by SNIFFLES. In terms of applying assembly-based methods, an improvement in SV identification may be seen using an alternative genome assembler such as phasebook [57]. Further analysis of mapped reads showed that the nanopore reads contained SV information, enabling further characterization than by NGS sequencing [20] by refining the coordinates

of chromosome 1 duplication, 0.5 Mb deletion in chromosome 18 and one extremity of the chromosome 18 deletion/circularization (Table 1). The strategy of junction reads extraction employed here demonstrated that SV information is contained in reads but softwares are not adapted to their optimal use. Specifically, they cannot resolve cases where one extremity of the split read is uniquely aligned and the other extremity multi-aligned as it is the case of chromosome 18 deletion/circularization. Also, this highlights that nanopore long reads are still affected by uncertainty due to genomic repetitions, which were also observed in the probably false positive chromosome 22 duplication which was described in a very repetitive region. The recently published human reference genome T2T-CHM13 clarifies the sequences of many regions which are rich in repeats, which includes centromeric satellites and segmental duplications, in which the *p*-terminal half of chromosome 22 is rich, and annotates new features within these regions [8]. Other techniques such as optical mapping would be complementary in confirming the presence/absence of large SVs [58, 59]. However, there is a limited number of reference samples with 'truth-set' SV calls available, especially large indels, with lack of broad scope consent for reference cell lines such as GM21886 also hampering availability of reference data [60].

Conclusions

For clinical diagnostic purposes, the entire workflow from sample collection, DNA extraction, library preparation, sequencing and bioinformatic analysis need to be validated and quality controlled [61]. Pre-analytical aspects, especially sample integrity and NA extraction, are known to be critical steps with massive impact on downstream sequencing success [62, 63]. This study shows that both pre-analytical and sequencing quality metrics need to be evaluated in selecting the optimal HMW DNA extraction method. It illustrates that, while established DNA QC methods are informative, there is scope for improvement and standardization as well as potential for novel approaches for DNA integrity to be utilized as part of monitoring extraction performance or characterizing control samples. While ISO standards for pre-examination processes relating to DNA isolation from clinical specimens exist [64, 65], specific guidance for HMW DNA and long-read sequencing/optical mapping technologies does not yet exist. The EU IVDR [66] also includes the requirement to include the pre-analytical steps as part of the overall examination, underscoring the need for diagnostic developers to consider the assurance of high quality sample preservation and HMW DNA isolation within a diagnostic solution. This study shows that reference cell lines can be utilised to test the impact of pre-analytic factors on downstream SV identification which supports the need for greater availability of cellular reference samples with well-characterised SVs as controls for the full long-read sequencing workflow.

Methods

Interlaboratory study design

For the Interlaboratory Study, cryopreserved human lymphoblastoid cells (cell line GM21886) were purchased from the Coriell Institute (Camden, NJ, US) and distributed by the coordinating laboratory (Laboratory 3) to the other participating laboratories (four vials with each vial containing 10^7 human cells per laboratory). Cells were stored at −80 °C. The coordinating laboratory also provided an Interlaboratory Study Protocol and a Reporting Form to record experimental information and sample, library and sequencing QC (Supplementary Methods 1, Supplementary Table 2). The Study Protocol was based on the process developed by ONT for DNA extraction from whole blood [14] where DNA extraction is followed by removal of shorter DNA fragments using the Short Read Eliminator (SRE) Kit (Circulomics, Baltimore, MD, US). The laboratories tested four HMW gDNA extraction kits (FM, GT, NB and PG) as described in the Results. PG and GT kits were provided by QIAGEN (Hilden, Germany). DNA extractions were performed in triplicate ($\approx 3.3 \times 10^6$ cells/replicate) according to the manufacturer's specified kit instructions for the number of cells used (any deviations from the manufacturer's/Interlaboratory study protocol are shown in Table S1. GT, NB and PG DNA extracts were incubated overnight at room temperature with (GT, PG) or without shaking (NB) to fully dissolve DNA. Thereafter extracts were stored at 4 °C. Laboratory 4's DNA extracts were sent to Laboratory 2 for fragment size QC, library preparation and sequencing.

Following initial sample QC, Laboratories 1 and 4 reported issues with some of the DNA extracts, i.e. low yield. Laboratory 1 repeated the DNA extractions for PG and GT kits, following the same protocol as before, but with in-house prepared cryopreserved GM21886 cells, with an input of 5×10^6 cells/extraction. All DNA extracts were stored and/or transported at 4 °C for downstream analysis.

QC of DNA extracts

Purity

To assess the purity of the extracted DNAs, spectrophotometry analysis was performed with NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA). UV spectrum absorbance was measured at 230, 260 and 280 nm and absorbance ratios A260/280 and A260/230 calculated.

Yield

Mass concentration (ng/μL) of DNA extracts was measured by Qubit dsDNA BR kit with Qubit fluorimeter (Thermo Fisher Scientific).

Fragment size (PFGE)

Fragment length profile of gDNA samples extracted by Laboratory 1 was assessed using the capillary-based Femto Pulse system with the Genomic DNA 165 kb Kit (Agilent Technologies, Santa Clara, CA, USA). Prior to running the system, the genomic DNA samples were diluted to 1 ng/μL, based on concentration values determined using Qubit dsDNA HS Kit (Life Technologies, Carlsbad, CA, USA). Gels were prepared according to the Femto Pulse system specifications and run using FP-1002-22– gDNA 165 kb programme settings. The data was visualised and analysed using Agilent PROSize 3.0 software (Agilent Technologies, Santa Clara, CA, USA).

QC of fragment length profile of DNA samples extracted by Laboratories 2, 3 and 4 was performed using the Pippin Pulse system (Sage Science, Beverly, MA, USA). To analyse fragments of 5–80 kb a pre-set protocol was used with the following conditions: run time: 16 h and 45 min; voltage: 75 V; buffer: 0.5× TBE, SeaKem® GOLD Agarose 1% (Lonza, Rockland, ME, USA). Laboratory 2 used λ DNA-Mono Cut Ladder (New England Biolabs, Ipswich, MA, USA) and PFGE Mid-Range ladder (New England Biolabs, Ipswich, MA, USA) and Laboratory 3 used the CHEF DNA Size Standard Ladder (Bio-Rad, Hercules, CA, USA) and Quick-Load 1 kb Extend DNA Ladder (New England Biolabs, Ipswich, MA, USA). After the completion of the run, the gel was stained with ethidium bromide solution (Bio-Rad, Hercules, CA, USA) and visualised using NuGenius imaging system (Syngene, Cambridge, UK) by Laboratory 2. At Laboratory 3, the gel was stained with GelGreen® (Biotium, Hayward, USA) for 1 h and visualised with Quantum-CX5 (Vilber Lourmat, France).

Digital PCR

Digital PCR (dPCR) molecular linkage methodology developed by Regan et al. [21] was tested at Laboratory 3 as an indicator of fragment size using the QX200™ Droplet Digital™ PCR (ddPCR™) System (Bio-Rad, Pleasanton, CA). Physically linked targets on the same molecule are detected by the occurrence of dPCR partitions which are positive for both a FAM-labelled "mile marker" assay and a HEX-labelled assay *RPP30* "anchor sequence" separated at different distances (33 kb, 60 kb, 100 kb, 150 kb and 210 kb) on chromosome 10. As a negative control, one mile-marker assay (33 kb-distant) was paired with an HEX-labelled assay targeting *RPPH1* gene on a different chromosome (chr14). All primers were obtained from Sigma Aldrich (Merck KGaA, Darmstadt, Germany). RPP30 anchor, 60 kb, 150 kb, and 210 kb probes were obtained from IDT (Integrated DNA Technologies Inc., Iowa USA) as Iowa Black quenched probes. RPP30 33 kb and 100 kb probes were obtained from Applied Biosystems (Themo Fisher Scientific Inc., Waltham, MA, USA) as TaqMan-MGB probes (Supplementary Methods 2). The ddPCR assay conditions were identical to the ones previously described by Regan et al. Human gDNA (catalogue no. G3041 Promega, Madison, WI, USA) was used as control to monitor inter-experiment variability. Sample input per reaction was adjusted to a maximum of 60 ng per reaction. Laboratories sent their DNA extracts (1 aliquot per partner) to Laboratory 3 for dPCR analysis. Single measurements of each extract for each linkage distance were performed, except for Laboratory 3 where NB and FM pooled extracts were used ($n = 3$ dPCR replicates).

Library preparation and sequencing

Following sample QC, the three extract replicates for each kit were pooled if QC data was comparable in terms of yield and fragment size. When considerable variation was observed, the replicates with the best yield and purity values were selected for downstream analysis. In respect to FM extracts, only Fraction A was further processed at all laboratories due to considerably higher yields than Fraction B. The pooled DNA extracts were size-selected (input: 5–8 μg) using the SRE kit (progressive depletion: <25 kb; near complete depletion: <10 kb). The SRE protocol was followed, with the respective SRE buffer being adjusted, when required (low DNA concentration), to the volume of the DNA extract on a 1:1 ratio. Next, 3 to 6 μg of SRE-treated HMW DNAs were end-prepped and adapter-ligated for sequencing using the Ligation Sequencing kit (SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK). The libraries were prepared using a modified version of the ligation sequencing protocol, with the bead-based library purification elution steps consisting of 90 min incubation at 37°C instead of 10 min or less at RT. The sequencing libraries were run in the MinION platform using R9.4.1 flow cells (FLO-MIN106D; Oxford Nanopore Technologies, Oxford, UK). The MinION was run for 72 h and, with the objective of maximising output, three FC loadings were performed with aliquots from the same sequencing libraries, with the second loading performed at 24–30 h and the third loading at 48–54 h, using the Flow Cell Wash Kit (EXP-WSH004; Oxford Nanopore Technologies, Oxford, UK) together with the Flow Cell Priming Kit (EXP-FLP002; Oxford Nanopore Technologies, Oxford, UK). Variations on this are detailed in Table S1/S2.

In addition to MinION sequencing, Laboratory 1 also sequenced their SRE-treated libraries with PromethION

Beta platform (Oxford Nanopore Technologies, Oxford, UK). Here, sequencing libraries corresponding to the four extraction kits were multiplexed using the Ligation Sequencing kit (SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK) in conjunction with the Native Barcoding kit (EXP-NBD104; Oxford Nanopore Technologies, Oxford, UK) and simultaneously sequenced. The protocol followed the manufacturer's recommendations.

Data analysis

Following sequencing, raw fast5 reads from Laboratory 1 (MinION and PromethION) and Laboratory 3 (MinION) were transferred to Laboratory 2 for processing and analysis. FastQ files were generated from fast5 reads after base calling using Guppy v.5.0.11 (Oxford Nanopore Technologies, Oxford, UK). Data was base called with four different models: fast, high-accuracy, high accuracy - modified bases/5mC and super high-accuracy) (Table S3). Sequence data was aligned against a human reference genome (GRCh38) [67] with Minimap2 v.2.18.27 [68]. QC metrics were calculated using summary sequencing statistics from Guppy and paf file from Minimap2. Super high accuracy base-called data was used for SV analysis.

For SV analysis, BAM files from Laboratories 1, 2 and 3 were merged for each extraction kit and all four extraction kits ("ALL" dataset). Laboratory 4 results were not included due to missing data for FM and GT methods which would have led to an unbalanced number of flowcells for the four extraction methods.

SV analysis at Laboratory 2 was performed with CUTESV, CUTESV version 2 and SNIFFLES [22, 23, 25] using super high accuracy (*SUP*)-based called data. SVs were filtered with the following parameters: SV filtering: all SV calling was restricted to SV > 50 nt; missing genotypes or homozygote reference genotypes were removed; SV variants PASS and BND variants were not evaluated. SV data for each extraction was compared with the ALL 'truth-set' using *truvari bench* (version 3.4.0) [69]. The ALL SV datasets for each SV caller were also compared using *truvari consistency* to establish concordance between the alternative SV callers.

Further analysis of large chromosomal SV was performed by Laboratory 1 with the MinION 'ALL' and PromethION datasets using both mapping and assembly approaches. For the mapping approach, reads were aligned as performed as above followed by SV calling using Sniffles 2 [24]. A strategy to specifically target reads covering junctions of known SVs was implemented using the mapping information and split reads aligned each part of the targeted SV were extracted. Following this, it was checked that split alignments were consistent with the expected pattern for the given SV type.

For the assembly approach, reads were assembled with the Flye assembler [28], followed by SV calling with RaGOO [29]. Assembly metrics are shown in Table S4C. Read junctions, identified as above, were aligned with minimap2 against contigs of the assembly. Then we searched for alignments indicating that these reads were completely contained in a contig. In such a case, this would indicate that the SV can be characterized with the assembly.

Statistics

Analysis of sample and sequencing QC metrics was performed using Graphpad Prism version 9.4 (Graphpad software LLC, Boston, MA). Sample QC metrics (yield, A260 ratios) were analysed by two-way ANOVA with factors extraction method and laboratory. Sequencing metrics (yield, read count, N50) were analysed by one-way ANOVA with Tukey's multiple comparison test (yield results). Read length distribution was analysed by two-way ANOVA with matching according to flow-cell.

dPCR linkage data was analysed in R version 4.2.1 and RStudio version 2022.7.1.554 using a linear mixed effects model [70, 71]. Only data corresponding to extracts which were sequenced with MinION FCs were included in the analysis. Extraction method and distance (33, 60, 100, 150, 210 kb) were treated as fixed (categorical) effects and laboratory and method/laboratory interaction as random effects. Variance was stratified within-group by extraction method. As linkage results for the FM kit at the 33 kb distance were set as the intercept of the model, differences between distances and kits were analysed relative to the 33 kb and FM kit respectively.

Abbreviations

chr	Chromosome
FC	Flow cell
FM	Fire Monkey
gDNA	Genomic DNA
GT	Genomic Tip
HMW	High molecular weight
NB	Nanobind
PG	Puregene
SOP	Standard Operating Procedure
SV	Structural variant
WGS	Whole genome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11792-7.

Supplementary Material 1: Supplementary Figures.

Supplementary Material 2: Supplementary Table 1.

Supplementary Material 3: Supplementary Table 2.

Supplementary Material 4: Supplementary Table 3.

Supplementary Material 5: Supplementary Table 4.

Supplementary Material 6: Supplementary Methods 1.

Supplementary Material 7: Supplementary Methods 2.

Supplementary Material 8. dMIQE checklist.

Acknowledgements
We would like to thank Miranda Stobbe (CNAG) for assistance with dbGaP submission.

Authors' contributions
AD, SH, JFD, IG, MG and CFo conceived the study. AD, RPAP, DY, EC, MDe, CFu, JGC, IGE, AW and TV developed and implemented extraction/sequencing workflows or dPCR methodology. AD, JM, CJ, RPAP, DY, RT, SH, MDa and SC performed bioinformatic/statistical/data analysis. RPAP, DY, EC, MDe, CFu, JGC and IGE performed experimental analysis. TV provided resources. AD, JM, RT, MDa and MCFL performed data curation. AD, RPAP, LHH and DY wrote the original draft manuscript. AD, JM, CJ, SM, MG and CFo edited the manuscript. AD, JM and CJ visualized data and prepared figures. AD, AW, TV, JFD, IG, MG and CFo performed supervision. AD, SM and CFo performed project administration. JFD, IG, MG and CFo acquired funding. All authors reviewed and approved the final manuscript.

Funding
The work described was conducted under the EASI-Genomics project which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824110.
The work described in this paper performed at the National Measurement Laboratory was funded in part by the UK government Department for Science, Innovation & Technology (DSIT) through the Chemical & Biological Metrology programme. Institutional support to CNAG was provided by the Spanish Government (Ministry of Science, Innovation and Universities) and the Generalitat de Catalunya through the Departament de Recerca i Universitats and the Departament de Salut.
The CEA-CNRGH sequencing platform was supported by the France Génomique National infrastructure, funded as part of the « Investissements d'Avenir » program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

Data availability
FASTQ files are available in dbGaP: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003958.v1.p1.

Declarations

Ethics approval and consent to participate
Not applicable. Cell line sample GM21886 is commercially available and authorisation to perform the research performed in this study was given by the responsible institute, the Coriell Institute (documentation available upon request).

Consent for publication
Not applicable.

Competing interests
QIAGEN was an EASI genomics partner, provided one of the test laboratories and is distributor of two of the tested HMW DNA extraction kits.

Author details
[1]National Measurement Laboratory (hosted at LGC), The Priestley Centre, 10 Priestley Road, Guildford, Surrey GU2 7XY, UK
[2]Centro Nacional de Análisis Genómico (CNAG), Baldiri Reixac 4, Barcelona 08028, Spain
[3]Universitat de Barcelona (UB), Barcelona, Spain
[4]Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), Evry 91057, France
[5]QIAGEN GmbH, QIAGEN Strasse 1, Hilden 40724, Germany

Received: 7 August 2024 / Accepted: 5 June 2025
Published online: 28 July 2025

References

1. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018;34(9):666–81.
2. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21(10):597–614.
3. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.
4. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517(7536):608–11.
5. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Briere C, Owens GL, Carrere S, Mayjonade B, et al. The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. Nature. 2017;546(7656):148–52.
6. Miller ME, Zhang Y, Omidvar V, Sperschneider J, Schwessinger B, Raley C, Palmer JM, Garnica D, Upadhyaya N, Rathjen J, et al. De Novo Assembly and Phasing of Dikaryotic Genomes from Two Isolates of Puccinia coronata f. sp. avenae, the Causal Agent of Oat Crown Rust. mBio 2018;9(1):e01650-17. https://doi.org/10.1128/mBio.01650-17.
7. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592(7856):737–46.
8. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
9. Marx V. Method of the year: long-read sequencing. Nat Methods. 2023;20(1):6–11.
10. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.
11. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2019;35(13):2193–8.
12. PacBio Technical Note. Preparing DNA for PacBio HiFi sequencing — Extraction and Quality Control. P/N: 102-193-651 REV02 13OCT2022. https://www.pacb.com/wp-content/uploads/Technical-Note-Preparing-DNA-for-PacBio-HiFi-Sequencing-Extraction-and-Quality-Control.pdf. Accessed 30 Nov 2023.
13. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62.
14. Oxford Nanopore Technologies. High output, long read sequencing from human blood. Version: 8th November, 2019. https://nanoporetech.ent.box.com/s/m94vddle291cmprejro1pb59czjfu540. Accessed 27 June 2024.
15. Pacific Biosciences. Template Preparation and Sequencing Guide. P/N 000-710-821-13. https://www.pacb.com/wp-content/uploads/2015/09/Guide-Pacific-Biosciences-Template-Preparation-and-Sequencing.pdf. Accessed 27 June 2024.
16. Jones A, Torkel C, Stanley D, Nasim J, Borevitz J, Schwessinger B. High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. PLoS ONE. 2021;16(7):e0253830.
17. Goodwin S, Wappel R, McCombie WR. 1D genome sequencing on the Oxford nanopore minion. Curr Protoc Hum Genet 2017;94:18 11 11–18 11 14.
18. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol. 2020;38(6):701–7.
19. Tang Z, Berlin DS, Toji L, Toruner GA, Beiswanger C, Kulkarni S, Martin CL, Emanuel BS, Christman M, Gerry NP. A dynamic database of microarray-characterized cell lines with various cytogenetic and genomic backgrounds. G3 (Bethesda). 2013;3(7):1143–9.
20. Gross AM, Ajay SS, Rajan V, Brown C, Bluske K, Burns NJ, Chawla A, Coffey AJ, Malhotra A, Scocchia A, et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. Genet Med. 2019;21(5):1121–30.
21. Regan JF, Kamitaki N, Legler T, Cooper S, Klitgord N, Karlin-Neumann G, Wong C, Hodges S, Koehler R, Tzonev S, et al. A rapid molecular approach for chromosomal phasing. PLoS ONE. 2015;10(3):e0118270.
22. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. Long-read-based human genomic structural variation detection with CuteSV. Genome Biol. 2020;21(1):189.
23. Jiang T, Cao S, Liu Y, Liu S, Liu B, Wang G, Wang Y. Regenotyping structural variants through an accurate force-calling method. BioRxiv. 2023. https://doi.org/10.1101/2022.1108.1129.505534.
24. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D et al. Detection of mosaic

and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42(10):1571–80. https://doi.org/10.1038/s41587-023-02024-y.
25. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8.
26. Midha MK, Wu M, Chiu KP. Long-read sequencing in Deciphering human genetics to a greater depth. Hum Genet. 2019;138(11–12):1201–15.
27. Wang J, Chen K, Ren Q, Zhang Y, Liu J, Wang G, Liu A, Li Y, Liu G, Luo J, et al. Systematic comparison of the performances of de Novo genome assemblers for Oxford nanopore technology reads from Piroplasm. Front Cell Infect Microbiol. 2021;11:696669.
28. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
29. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20(1):224.
30. Goldrich DY, LaBarge B, Chartrand S, Zhang L, Sadowski HB, Zhang Y, Pham K, Way H, Lai C-YJ, Pang AWC, et al. Identification of somatic structural variants in solid tumors by optical genome mapping. J Pers Med. 2021;11(2):142.
31. Revolugen. HMW DNA Extraction https://revolugen.co.uk/products-technologies/hmw-dna-extraction.html. Accessed 30 Nov 2023.
32. QIAGEN. Gentra® Puregene® Handbook Fourth Edition 12/2014.
33. QIAGEN Genomic DNA Handbook 06/2015.
34. Gand M, Bloemen B, Vanneste K, Roosens NHC, De Keersmaecker SCJ. Comparison of 6 DNA extraction methods for isolation of high yield of high molecular weight DNA suitable for shotgun metagenomics nanopore sequencing to detect bacteria. BMC Genomics. 2023;24(1):438.
35. ISO 15189: 2022 Medical laboratories — Requirements for quality and competence. https://www.iso.org/standard/76677.html.
36. FDA, Considerations for Design. Development, and Analytical Validation of Next Generation Sequencing (NGS) - Based In Vitro Diagnostics (IVDs) Intended to Aid in the Diagnosis of Suspected Germline Diseases (ref. FDA-2016-D-1270) https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-development-and-analytical-validation-next-generation-sequencing-ngs-based. Accessed 08 Jan 2025.
37. CLSI MM09 Human Genetic and Genomic Testing Using Traditional and High-Throughput Nucleic Acid Sequencing Methods, 3rd Edition. 2023. https://clsi.org/shop/standards/mm09/.
38. Scientific Working Group on DNA Analysis Methods (SWGDAM). The Quality Assurance Standards for Forensic DNA Testing Laboratories, Effective July 1. 2020 https://www.swgdam.org/_files/ugd/4344b0_d73afdd0007c4ed6a0e7e2ffbd6c4eb8.pdf. Accessed 8 Jan 2025.
39. Devonshire A, Jones G, Gonzalez AF, Kofanova O, Trouet J, Pinzani P, Gelmini S, Bonin S, Foy C. Interlaboratory evaluation of quality control methods for Circulating cell-free DNA extraction. N Biotechnol. 2023;78:13–21.
40. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, Langlade N, Muños S. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. Biotechniques. 2016;61(4):203–5.
41. CLSI EP05. Evaluation of Precision of Quantitative Measurement Procedures, 3rd Edition. 2014. https://clsi.org/shop/standards/ep05/.
42. ONT Community Documentation, Contaminants. https://community.nanoporetech.com/contaminants. Accessed 2024/07/24.
43. Blocking Unblocking and Flow Cell Output. https://community.nanoporetech.com/posts/blocking-unblocking-and-f. Accessed 08 Jan 2025.
44. Jaudou S, Tran M-L, Vorimore F, Fach P, Delannoy S. Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing Escherichia coli. PLoS ONE. 2022;17(7):e0270751.
45. Gong L, Wong C-H, Idol J, Ngan CY, Wei C-L. Ultra-long read sequencing for whole genomic DNA analysis. JoVE 2019;(145):e58954.
46. Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, Asri M, Alvarez Jerez P, Malik L, Dewan R, et al. Scalable nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. Nat Methods. 2023;20(10):1483–92.
47. Giron'Koetsier PA, Cantor EJ. A simple approach for effective shearing and reliable concentration measurement of ultra-high-molecular-weight DNA. Biotechniques. 2021;71(2):439–44.
48. Klingström T, Bongcam-Rudloff E, Pettersson OV. A comprehensive model of DNA fragmentation for the preservation of High Molecular Weight DNA. bioRxiv. 2018:254276.
49. Gleerup D, Chen Y, Van Snippenberg W, Valcke C, Thas O, Trypsteen W, De Spiegelaere W. Measuring DNA quality by digital PCR using probability calculations. Anal Chim Acta. 2023;1279:341822.
50. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience. 2019;8(5):giz043. https://doi.org/10.1093/gigascience/giz043.
51. Devonshire AS, Whale AS, Gutteridge A, Jones G, Cowen S, Foy CA, Huggett JF. Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification. Anal Bioanal Chem. 2014;406(26):6499–512.
52. Majidian S, Agustinho DP, Chin CS, Sedlazeck FJ, Mahmoud M. Genomic variant benchmark: if you cannot measure it, you cannot improve it. Genome Biol. 2023;24(1):221.
53. LoTempio J, Delot E, Vilain E. Benchmarking long-read genome sequence alignment tools for human genomics applications. PeerJ. 2023;11:e16515.
54. Helal AA, Saad BT, Saad MT, Mosaad GS, Aboshanab KM. Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. Sci Rep. 2024;14(1):6160.
55. Jiang T, Liu S, Cao S, Liu Y, Cui Z, Wang Y, Guo H. Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. BMC Bioinformatics. 2021;22(1):552.
56. Duan X, Pan M, Fan S. Comprehensive evaluation of structural variant genotyping methods based on long-read sequencing data. BMC Genomics. 2022;23(1):324.
57. Luo X, Kang X, Schönhuth A. Phasebook: haplotype-aware de Novo assembly of diploid genomes from long reads. Genome Biol. 2021;22(1):299.
58. Barseghyan H, Tang W, Wang RT, Almalvez M, Segura E, Bramble MS, Lipson A, Douine ED, Lee H, Délot EC, et al. Next-generation mapping: a novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. Genome Med. 2017;9(1):90.
59. Mantere T, Neveling K, Pebrel-Richard C, Benoist M, van der Zande G, Kater-Baats E, Baatout I, van Beek R, Yammine T, Oorsprong M, et al. Optical genome mapping enables constitutional chromosomal aberration detection. Am J Hum Genet. 2021;108(8):1409–22.
60. Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, Zook JM. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet. 2023;24(7):464–83.
61. CEN/TS 17981-1:2023. In vitro diagnostic Next Generation Sequencing (NGS) workflows - Part 1: Human DNA examination. (publication date 29 Nov. 2023). https://standards.iteh.ai/catalog/standards/cen/268bb113-1bc2-41c6-b159-5c25d671dd7c/cen-ts-17981-1-2023?srsltid=AfmBOor114CrUcndsBakXMy-uYEMXEqDXS3ySiF6oi2F1_un-juNfs9o.
62. Dagher G, Becker KF, Bonin S, Foy C, Gelmini S, Kubista M, Kungl P, Oelmueller U, Parkes H, Pinzani P, et al. Pre-analytical processes in medical diagnostics: new regulatory requirements and standards. N Biotechnol. 2019;52:121–5.
63. Oelmueller U. Standardization of generic pre-analytical procedures for in vitro diagnostics for personalized medicine. N Biotechnol. 2021;60:1.
64. ISO 20186-2: 2019. Molecular in vitro diagnostic examinations specifications for pre-examination processes for venous whole blood part 2. Isolated genomic DNA. https://www.iso.org/standard/69799.html.
65. ISO 20184-3. 2021. Molecular in vitro diagnostic examinations Specifications for pre-examination processes for frozen tissue/ Part 3: Isolated DNA. https://www.iso.org/standard/78110.html.
66. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices. https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng.
67. GRCh38 - NCBI FTP site - National Institutes of Health (NIH). http://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz. Accessed 2024/07/24.
68. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
69. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol. 2022;23(1):271.
70. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. http://www.R-project.org/.
71. Pinheiro J, Bates D, Team RC. nlme: Linear and Nonlinear Mixed Effects Models. 2022.

Publisher's Note

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