Published online 13 July 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 3 **1**
https://doi.org/10.1093/nargab/lqac051

High-throughput method for the hybridisation-based targeted enrichment of long genomic fragments for PacBio third-generation sequencing

Tim Alexander Steiert[1,†], Janina Fuß[1,†], Simonas Juzenas[1,2], Michael Wittig[1], Marc Patrick Hoeppner[1], Melanie Vollstedt[1], Greta Varkalaite[3], Hesham ElAbd[1], Christian Brockmann[4], Siegfried Görg[4], Christoph Gassner[5], Michael Forster[1] and Andre Franke[1,*]

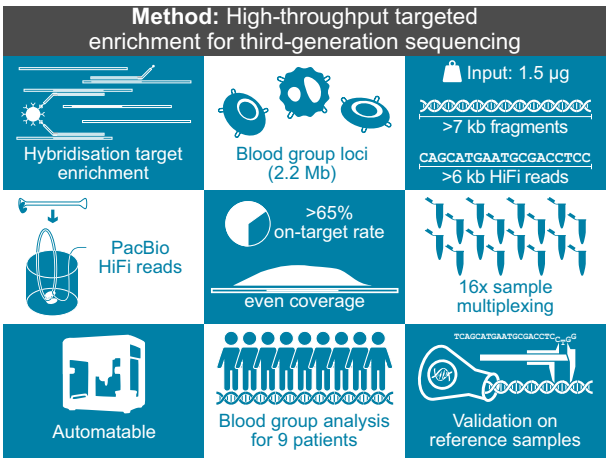
[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel 24105, Germany, [2]Institute of Biotechnology, Life Science Centre, Vilnius University, Vilnius 02241, Lithuania, [3]Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas 44307, Lithuania, [4]Institute of Transfusion Medicine, University Hospital of Schleswig-Holstein, Kiel 24105, Germany and [5]Institute of Translational Medicine, Private University in the Principality of Liechtenstein, Triesen 9495, Liechtenstein

Received March 28, 2022; Revised June 08, 2022; Editorial Decision June 20, 2022; Accepted June 29, 2022

ABSTRACT

Hybridisation-based targeted enrichment is a widely used and well-established technique in high-throughput second-generation short-read sequencing. Despite the high potential to genetically resolve highly repetitive and variable genomic sequences by, for example PacBio third-generation sequencing, targeted enrichment for long fragments has not yet established the same high-throughput due to currently existing complex workflows and technological dependencies. We here describe a scalable targeted enrichment protocol for fragment sizes of >7 kb. For demonstration purposes we developed a custom blood group panel of challenging loci. Test results achieved > 65% on-target rate, good coverage (142.7×) and sufficient coverage evenness for both non-paralogous and paralogous targets, and sufficient non-duplicate read counts (83.5%) per sample for a highly multiplexed enrichment pool of 16 samples. We genotyped the blood groups of nine patients employing highly accurate phased assemblies at an allelic resolution that match reference blood group allele calls determined by SNP array and NGS genotyping. Seven Genome-in-a-Bottle reference samples achieved high recall (96%) and precision (99%) rates. Mendelian error rates were 0.04% and 0.13% for the included Ashkenazim and Han Chinese trios, respectively. In summary, we provide a protocol and first example for accurate targeted long-read sequencing that can be used in a high-throughput fashion.

GRAPHICAL ABSTRACT



INTRODUCTION

Next- or second-generation sequencing (NGS) (1) has revolutionised genomics, transcriptomics and our molecular understanding of disease. However, in some applications NGS techniques are limited by their relatively short read length. For example, with NGS short-reads of a few hundred base pairs in size (up to 600 bp with Illumina paired-end NGS), it is very difficult or impossible to elucidate and resolve difficult-to-map genomic regions such as long repeats or

*To whom correspondence should be addressed. Tel: +49 431 500 15110; Fax: +49 431 500 15168; Email: a.franke@ikmb.uni-kiel.de
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

paralogous sequences that account for up to 50% (2,3) of mammalian genomes (4,5). Detection of clinically relevant long indels, fusion genes or structural variants is also often challenging for short-read NGS approaches. In addition, phasing of haplotypes (6) or *de novo* genomic assemblies (4) accomplished with short-read data are either not feasible for an extensive portion of the genome or require costly sequencing methods (7). Phase analysis to distinguish between in cis and in trans heterozygous mutations is important for clinical diagnosis because in trans compound heterozygous mutations cause autosomal recessive disease (8). Despite these technical limitations of NGS, its applications in high-throughput has transformed almost every discipline of biomedical sciences (9). More recently, third-generation sequencing (TGS) has helped to overcome these limitations by producing significantly longer read lengths of > 10 kb (10,11). Despite the continuous reduction in cost in the last years, utilising NGS and TGS for larger cohorts remains a costly matter. For complex genomes, such as the human genome, the cost per genome is still in the range of one thousand USD (12). Therefore, targeted enrichment is used to enrich genomic fragments of interest rather than sequencing random genomic fragments. With targeted enrichment many barcoded samples can be pooled and be enriched in sample batches for specific loci, so that a high number of samples can be sequenced for the focused target loci in a single sequencing run. Combined with target enrichment, TGS has the benefit that high per base coverages can be achieved, which is relevant, for example, to achieve sufficient diagnostic accuracy and certainty. Sometimes, whole genome sequencing or whole exome sequencing is legally also not possible as the ethical consent of patients or study probands may not allow untargeted analysis that may also yield incidental findings. Therefore, improving current TGS methods is a meaningful undertaking, especially when it comes to increasing sequencing output and throughput lagging behind NGS systems (13).

In-solution hybridisation-based capture (14) and enrichment is currently the most widely used technique to enrich genomic fragments of interest in various research contexts (15–21). Hybridisation-based panels using biotinylated DNA/RNA oligo baits and their capture by streptavidin coated magnetic beads, have the advantage of a flexible, scalable, and customisable targeting strategy over other enrichment techniques (22). To achieve high genotyping accuracy with NGS a coverage depth of 15× is recommended (23). Deep, even, and ideally complete target coverage is also essential for clinical and scientific sequencing applications (24,25). To ensure this, a well-balanced oligo bait panel is required, that captures sufficient and diverse target fragments, for a comprehensive fraction of the target bases. For NGS applications, such oligo bait panels can reach on-target rates ranging from 48% (26) to 95% (27) of all reads, depending on several factors including target loci features, panel size (total covered genomic region), input DNA amount, and capture chemistry.

After the success of high-throughput targeted enrichment in NGS, we here describe a protocol that overcomes the procedural bottlenecks of current targeted enrichment protocols for long fragments. The herein presented protocol overcomes challenges in size selection, capacity limitations for multiplexing, and the requirement of a second PCR of existing long fragment capture protocols, enabling targeted high-throughput sample processing in context of TGS. We here demonstrate that with our approach multiplexing of 16 samples for blood group system loci of 2.2 Mb is possible in contrast to lower sample multiplexing in other TGS protocols. We demonstrate use of the blood group system loci for our targeted TGS panel, in order to show the advantages compared to conventional targeted NGS. Beyond the generally known clinical necessity to match the blood groups of donor and recipient of blood transfusions, blood groups are furthermore implicated in the susceptibility to disease (28). We show that data obtained from our targeted TGS preparation is of high accuracy, and that allele phasing can be achieved, even for challenging paralogous loci. In addition, using the presented TGS-based approach, we show that structural variants can be identified, which often remain concealed using the more established NGS protocols.

MATERIALS AND METHODS

Samples and ethics agreement

To showcase the TGS method we prepared 16 samples, seven of which were Genome-in-a-Bottle (GIAB) reference samples NA12878, NA24143, NA24149, NA24385, NA24631, NA24694, and NA24695 (all Coriell Biorepository, USA (29)), including the Ashkenazim Jewish and Han Chinese trios. Eight were high molecular weight (HMW) gDNA samples extracted from a research cohort of the Institute of Transfusion Medicine (ITM) of unknown blood groups. One sample of HMW gDNA for which Illumina whole genome sequencing (WGS) data was available was taken from the German Centre for Cardiovascular Research (DZHK) cohort. For all patient samples of the ITM and DZHK cohorts the study aims were approved by the ethics commission of the University of Kiel, Germany in AZ: A 103/14 and AZ: D441/16, respectively. Informed consent was obtained from all the patients.

Custom blood group bait panel design

The custom discovery pool panel (Integrated DNA Technology, USA) used for enrichment covers 35 genomic loci of genes encoding human blood group systems, as listed by the International Society of Blood Transfusion's (ISBT) Working Party for Red Cell Immunogenetics and Blood Group Terminology (v9.0, 2021) (30). The panel further targets the genes of two transcription factors (*GATA1* and *KLF1*). The biotinylated DNA capture probes have a length of 120 bp and the design has a 0.5× tiling (120 bp gaps between probes), targeting 2.2 Mb blood group loci as well as 0.4 Mb flanking regions. Specific probe optimisation for the purpose of this study was kindly provided by Integrated DNA Technology. The final design has 8121 probes. A full list of all genes targeted is available in Supplementary Table S1.

DNA extraction

DNA for the ITM cohort samples was extracted from whole blood using the chemagic DNA Blood Kit special (#CMG-

1710, PerkinElmer, USA). For the sample with matching Illumina WGS data, DNA was extracted with Qiagen MagAttract HMW DNA Kit (#67563, Qiagen, Germany).

High-throughput sample preparation

1.5 μg of genomic DNA were diluted with nuclease-free water to a total volume of 150 μl and fragmented to a target fragment length of 10 kb with a g-tube (#520079, Covaris, USA). DNA pre-fragmentation was done by spinning the g-tube two times for 270 s at 1657 g. Subsequent fragmentation was done by spinning four additional times at 3381 g for 60 s. From the g-tube, 120 μl of fragmented sample was recovered and transferred to a new tube, where it was size-selected using 0.875x volume of size selection beads (washed magnetic bead particles from AmpureXP beads (Beckman Coulter, USA), 0.1 M Tris–HCl, pH 8.5 (#BU-124S-85, Jena Biosciences, Germany), 0.75 M LiCl (#AM9480, Invitrogen, USA), 20% PEG 8000 (#CSS-256, Jena Biosciences, Germany). End-repair, A-tailing, and adapter ligation of previously annealed PacBio multiplexing adapters (custom, IDT, USA) were performed with KAPA HyperPrep Kit (#7962371001, Roche, Germany). Prepared samples ($N = 16$) were cleaned up and pooled using 200 ng of each sample. Pooled samples were then hybridised with a custom blood group biotinylated bait panel, blockers (custom, IDT, USA), and 5 μg human COT DNA (#100285, Twist, USA) for 14 h at 65°C. Biotinylated probes were captured with M-270 Streptavidin beads (#65306, ThermoFisher, Lithuania). Hybridisation and wash were performed with xGen Hybridization and Wash Kit (#1080584, IDT, USA). Captured fragments were amplified with hotstart LA Taq DNA polymerase (#RR042A, Takara, Japan) in a 24-cycle long-range PCR (initial denaturation – 120 s at 95°C; denaturation – 20 s at 95°C; annealing – 15 s at 62°C; extension – 600 s at 68°C, final extension – 300 s at 68°C). The PCR product was size selected using 0.875× volume of previously described size selection beads.

Genotyping array

Genotyping of eight samples of the ITM cohort was conducted employing Illumina's Infinium Global Screening Array 24 v3.0 BeadChip (#20030770, Illumina, USA) following the Illumina Infinium HTS Assay workflow.

PacBio library preparation and SMRT sequencing

The 16-plexed, barcoded and enriched pool was used as the input of the SMRTbell Express Template Prep Kit 2.0 (#102-088-900, PacBio, USA), that was done according to the manufacturer's recommendations. Prepared libraries were sequenced on one SMRT Cell 8M (#101-820-200, PacBio, USA) on the PacBio Sequel II sequencing system with a movie length of 30 h.

Bioinformatic analysis

Subreads were demultiplexed using PacBio *lima* (v2.0.0). Longest subread length versus polymerase read length graph was exported from *SMRT Link Analysis* (v10.1.0.119588). From the PacBio sub read bam-files, circular consensus reads (CCS) with at least four passes (HiFi reads) were generated using PacBio *ccs* tool (v6.0.0). CCS reads were aligned to human hg38 standard chromosomes builds using PacBio *pbmm2* aligner (v1.3.0). NGS dataset was aligned using *bwa* (v0.7.12). Duplicate reads were removed from the alignment using *picard MarkDuplicates* (v2.22.3) Resulting alignment bam-files were analysed with *Alfred QC* (v0.2.1.) for key quality control (QC) metrics. Phasing and variant calling was done with *DeepVariant* (v1.2.0) and *WhatsHap* (v0.18). The blood groups were called using the inhouse software *DeepBlood* (commit a83502d1c95224fcaccf073bd5c24d05205d0d77). Ambiguous calls were manually verified. Mendel checks were conducted with *Bcftools* (v1.9) from vcf-files that were individually filtered for variants with a genotyping quality score (GQS) >20. VCF concordance check according to the PrecisionFDA truth challenge was done using a *nextflow* (v20.10.0) pipeline involving *bedtools* (v2.30.0), *GATK* (v4.2.4.1) and *hap.py* (v0.3.14). NISTv3.3.2 hg38 reference data has been used for concordance checks between reference and query data for all GIAB samples.

RESULTS

Approach for high-throughput target enrichment of long fragments

Schematical illustration of the capture-based target enrichment approach is provided in Figure 1. The approach involves two bead-based size selection steps, the conduction of a single long-range PCR, and the multiplexing of up to 16 samples. High-molecular weight (HMW) genomic DNA (Figure 1A) is fragmented to a target peak size of 10 kb (Figure 1B). The bead-based size selection removes small fragments with a hard cut-off at approx. 6 kb and at the same time constitutes a clean-up (Figure 1C). End-repair, A-tailing, adapter ligation, and a clean-up step (Figure 1D) are done before barcoded samples directly get multiplexed for hybridisation and capture step (Figure 1E). Post-capture long-range PCR (Figure 1F) is done for the enriched fragments. The PCR product, instead of a respective clean-up, again is size selected. These fragments can be used as input for a PacBio library preparation, without the need of additional amplification or size selection steps. All bead-based size selection steps, the clean-up, end-repair, and adapter ligation steps can be automatised on a commercial lab robot if throughput needs further to be increased. In comparison to other protocols aiming at fragments >5 kb (31–34), that all require electrophoretic size selection(s), and a second long-range PCR. This saves approx. a third of total reagent and consumable costs. In addition to this, sample multiplexing for capture decreases the cost for capturing chemicals and bait panels by a factor of two (32) to sixteen (31,33,35–38).

General QC metrics of targeted long-read sequencing

General QC metrics provide an overview on distribution of reads between multiplexed samples, the sequencing on-target rate, run-statistics, and coverage depth and evenness.

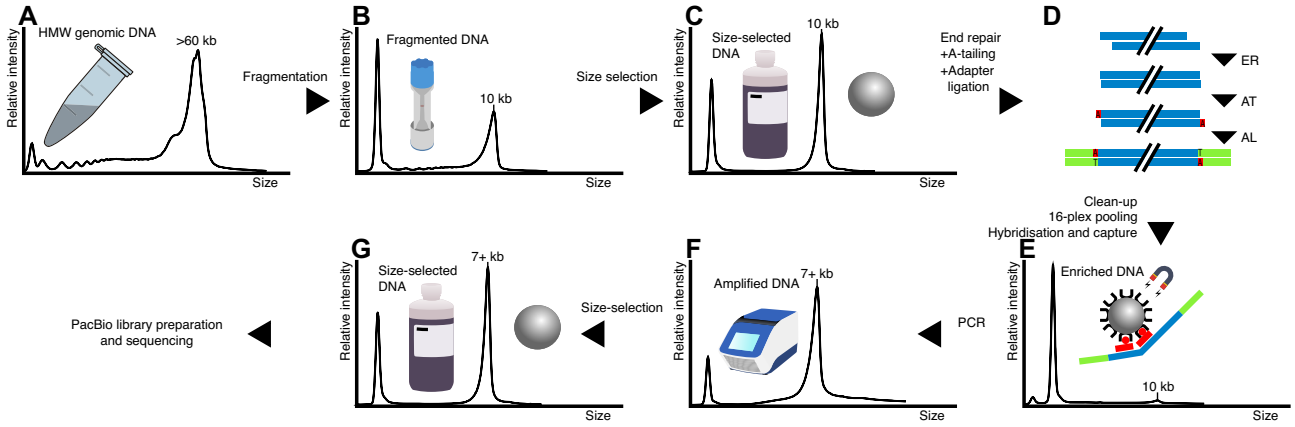


Figure 1. *Schematic workflow of a high-throughput compatible protocol for long fragment targeted enrichment.* (**A**) HMW genomic DNA is physically fragmented to a desired fragment length of 10 kb by repeatedly passing through an orifice of a g-tube (**B**). Here the centrifugation speed is the decisive parameter for the product fragment sizes and can be adjusted accordingly. (**C**) Unwanted smaller fragments are removed in the following step by clean-up with self-prepared size selection beads. Size selection is performed analogously to standard SPRI bead clean-up steps and can be done in 96-well plates on a commercial lab robot. (**D**) The size-selected fragments are end-repaired (ER) A-tailed (AT), and adapter ligated (AL) to an adapter sequence, which can be identified by its molecular barcode in subsequent sample pools. (**E**) After a conventional clean-up barcoded fragments can directly be pooled in equimolar ratios and are dried down for hybridisation, resuspended in hybridisation buffers, and mixed with unspecific blocking sequences (see Materials and Methods). After denaturation of the sample DNA, the targeted enrichment bait panel is added to the reaction and the mix is incubated at 65°C overnight for hybridisation. In the hybridisation step, the baits bind their respective target sequences in the samples. The capturing of the desired target fragments that become immobilised by the binding of DNA-linked biotin to the streptavidin surface of magnetic beads, is schematically depicted. Non-specifically bound fragments are washed away in a series of washing steps at different temperatures and with a variety of different wash buffers. (**F**) The captured and enriched target fragments get amplified in a PCR by the addition of ligated adapter-specific primers. Here, as smaller fragments get amplified preferentially in PCR-based amplification and sequencing steps, fragment sizes decrease from initially 10 to >7 kb (Supplementary Figure S1), and ultimately to >6 kb HiFi reads. (**G**) To avoid a drop in insert size, a second size selection step analogous to the first one is required to increase the average fragment size. This replaces the PCR clean-up step. PacBio libraries can subsequently be prepared from the size-selected, enriched, and barcoded fragments.

425 gigabases (Gb) sequence output resulted from the PacBio TGS run. After demultiplexing, HiFi read generation, read alignment, and duplicate filtering ∼1.6 M reads remained for all 16 samples. Except for one sample, HiFi reads were distributed relatively equally between samples. Mean unique HiFi read number per sample was 97883 (55873–223494, SD = 40837, see Figure 2A). Through the high accuracy of redundantly sequenced reads and their long span of the sequencing reads it is possible to reliably resolve genomic regions of interest with relatively few unique sequences. PacBio zero mode waveguides achieved a good subread size distribution with very few reads below 2 kb in length and half of all polymerase reads N50 of 174 kb (see Figure 2B). This means that an exemplary 10 kb fragment has a median of 17.4 passes which result in a median consensus base quality of >99.98% (39). Shorter reads respectively being of higher accuracy. Resulting HiFi read length averaged at 6287 bp (5909–6819 bp, SD = 285 bp) for all 16 samples. Sequencing on-target rates of aligned reads for all 16 multiplexed samples were similar, averaged at 66.3%, and ranged from 65.2% to 67.4% with SD of 0.72% (Figure 2C). The prepared samples had an average duplicate rate of 16.5% (15.2–17.8%, SD = 0.8%). The blood group panel contains challenging paralogous targets, that short-read sequencing cannot reliably resolve. Therefore, a good coverage of these challenging targets is critical. Figure 2D shows that in the described capture paralogous loci ('par.'), e.g. the NMS blood group system, consisting of the three highly homologous genes *GYPA*, *GYPB* and *GYPE*, are sufficiently covered by HiFi reads. Coverage distribution is roughly comparable to a set of non-paralogous, normal ('nor.') loci of equivalent size. This indicates that despite the paralogous character of these sequences a hybridisation-based enrichment approach is suitable to enrich such loci. Sufficient coverage can be achieved by long-reads spanning from unique anchor sequences, which can be targeted with capture probes, throughout the paralogous or repetitive loci (Supplementary Figure S2). Reads were of sufficient length and coverage was high and uniform enough to phase 93.8% of heterozygous SNPs (see Supplementary Data).

Variant validation by truth challenge

To evaluate the accuracy of the targeted sequencing approach for the targeted blood group system panel we benchmarked against available reference data of Coriell reference samples. The seven reference samples of the Genome-in-a-Bottle (GIAB) consortium (29), comprising two trios, are among the most extensively sequenced samples in the world. Reference data from these samples are, for example, used in the PrecisionFDA truth challenge to determine the accuracy of variant calls (40). For variants called from TGS data, the mean PrecisionFDA recall and precision rates were 96,02% (95.53–97.47%, SD = 0.61%) and 98.79% (98.28–99.23%, SD = 0.28%), respectively. The number of true positive, false positive, and false negative variant calls of all seven GIAB samples are provided in Figure 3. Overall, a high conformity was achieved for the variants called, despite that the target regions contained challenging loci.

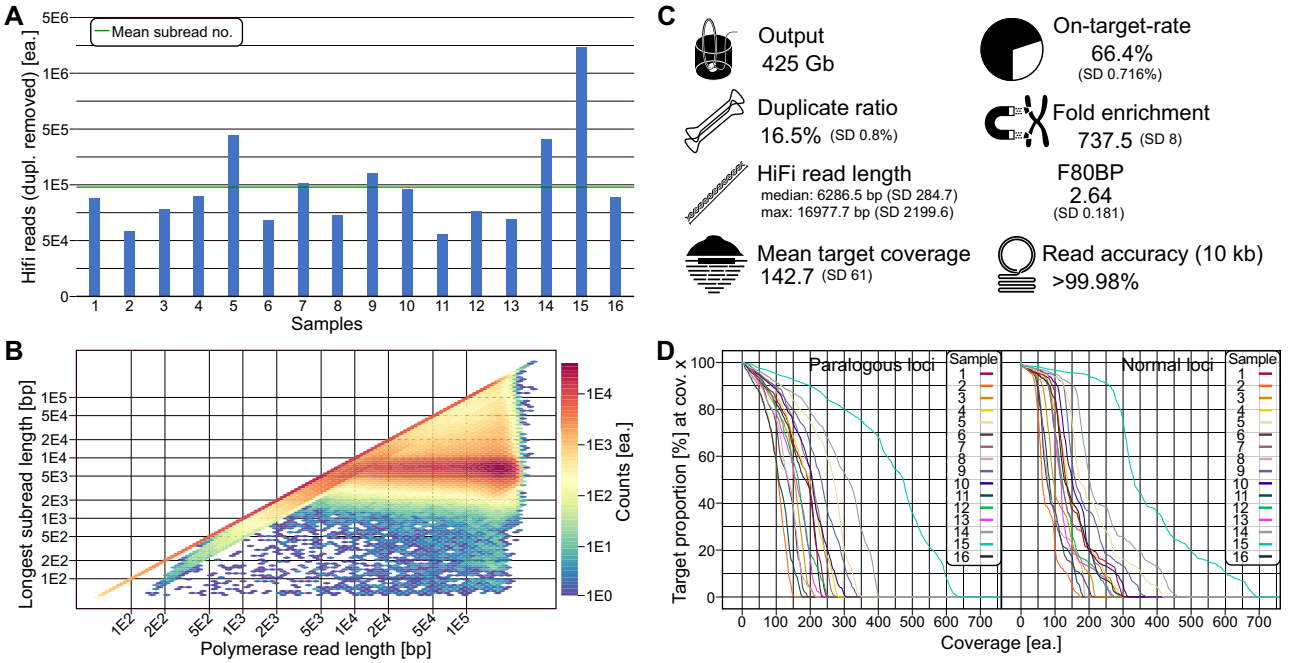


Figure 2. *General QC metrics.* (**A**) Number of non-duplicate HiFi reads per sample for the 16-sample multiplexed capture pool. The green line indicates the mean of all samples. (**B**) Selected run and QC metrics: run sequence output in Gb, duplicate ratio for HiFi reads, mean and maximum read length, mean target coverage of the HiFi read alignments, on-target-rate, fold enrichment, fold 80 base penalty (F80BP), and mean HiFi read accuracy based on an exemplary 10 kb fragment. Except for the read output and accuracy all numbers reflect the mean over all 16 samples and the respective standard deviation (SD) is provided. (**C**) Longest subread length versus polymerase read length. Count for respective subread/polymerase read length is indicated by colour, referring to scale on the right. (**D**) Coverage uniformity plot, showing the percentage of target region covered with x reads. Representative set of loci of equivalent size with paralogous ('par.': *GYPA/GYPB/GYPE*, 81 kb, left panel) and non-paralogous, normal ('nor.': *ACKR1/ABCB6/KEL/AQP1/SEMA7A/SLC4A1*, 79 kb right panel) sequences are depicted. Almost all target bases in both loci sets, paralogous (97%, 88.14–99.97%, SD = 3.47%) and normal (99.14%, 98.84–100%, SD = 0.25%) are covered with HiFi reads.

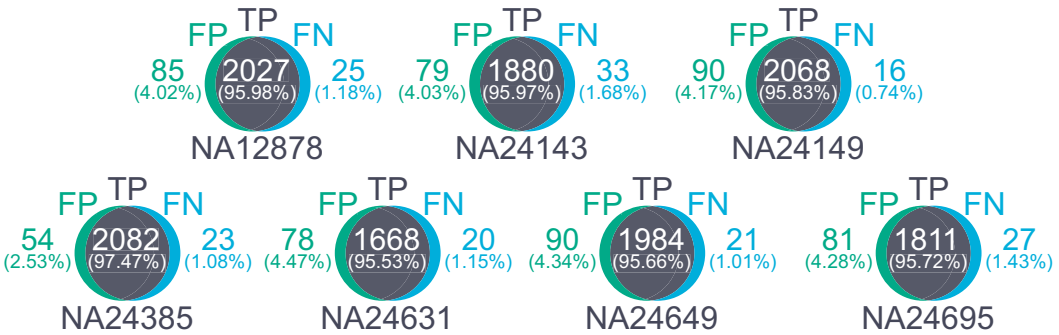


Figure 3. PrecisionFDA benchmarking results. Venn diagrams depict the number of true positive (TP, dark grey) SNP variants that were depicted in both the query and the reference data as well as the number of false negative (FN, light blue) and false positives (FP, green) in the query data. This is provided for the seven GIAB reference samples analysed in the current study. Only high-confidence variants of the GIAB samples were considered in the analysis. This was done by intersecting all files with the respective high-confidence region bed-files available for each GIAB sample. The resulting variant files were processed and benchmarked according to the PrecisionFDA truth challenge.

Variant validation by Mendelian check

In addition to the benchmarking against external reference data mendelian error rates were calculated as an internal control for the Ashkenazim and Han Chinese trios. On-target variant data was filtered for GQS >20 (41), no coverage filter was applied. Mendelian concordance (error) rates were 99.96% (0.04%) for the Ashkenazim Jewish and 99.87% (0.13%) for the Han Chinese trio. This concordance is higher than for identically filtered NGS WGS data (41).

Comparison with genotyping array

For eight unknown samples, the blood groups called from TGS data were compared to genotyping results obtained from genotyping array. For all eight samples and for all blood group systems that were covered by both the array, and the probe panel, we achieved perfect concordance (see Supplementary Table S2 and S3 for more details).

Table 1. Concordance of blood groups alleles between NGS/WGS using Illumina and TGS using PacBio technology. The table contains all blood group system loci typable in the latest version of DeepBlood and included in the bait panel. Differences that could be resolved using phasable TGS data were highlighted in bold letters. The error made in the allele assignment of GYPA is crossed out. This table contains all blood group system loci typable in the latest version of DeepBlood and included in the bait panel

Blood system	Alleles Illumina WGS	Alleles PacBio targeted enrichment
001/*ABO*	*ABO*O.02.01/ABO*O.01.01*	*ABO*O.02.01/ABO*O.01.01*
002/*MNS:GYPA*	*GYPA*01**/GYPA*08***	*GYPA*01*
002/*MNS:GYPB*	*GYPB*04*	*GYPB*04*
004/*RHCE*	*RHCE*01.01*	*RHCE*01.01**/RHCE*04, RHCE*01.02.02***
004/*RHD*	*RHD*01/RHD*01N.01*	*RHD*01/RHD*01N.01*
005/*LU*	*LU*02*	*LU*02*
006/*KEL*	*KEL*02*	*KEL*02*
007/*LE*	secretor	secretor
008/*FY*	*FY*02/FY*01*	*FY*02/FY*01*
009/*JK*	*JK*02/JK*01*	*JK*02/JK*01*
010/*DI*	*DI*02*	*DI*02**.04***
011/*YT*	*YT*01*	*YT*01*
013/*SC*	*SC*01*	*SC*01*
014/*DO*	*DO*02*	*DO*02*
015/*CO*	*CO*01.01*	*CO*01.01*
016/*LW*	*LW*05*	*LW*05*
018/*H*	*FUT1*01*	*FUT1*01*
019/*XK*	*XK*01*	*XK*01*
020/*GE*	*GE*01*	*GE*01*
021/*CROM*	*CROM*01*	*CROM*01*
022/*KN*	*KN*01/KN*01.-05*	*KN*01/KN*01.-05*
023/*IN*	*IN*02*	*IN*02*
024/*OK*	*OK*01.01*	*OK*01.01*
025/*RAPH*	*RAPH*01*	*RAPH*01*
026/*JMH*	*JMH*01*	*JMH*01*
027/*I*	*GCNT2*01*	*GCNT2*01*
028/*GLOB*	*GLOB*01*	*GLOB*01*
029/*GIL*	*GIL*01*	*GIL*01*
030/*RHAG*	*RHAG*01*	*RHAG*01*
031/*FORS*	*GBGT1*01N.01*	*GBGT1*01N.01*
032/*JR*	*ABCG2*01/ABCG2*01N.02.02, ABCG2*01N.06, ABCG2*01N.13*	*ABCG2*01/ABCG2*01N.02.02, ABCG2*01N.06, ABCG2*01N.13*
033/*LAN*	*ABCB6*01*	*ABCB6*01*
034/*VEL*	*VEL*01*	*VEL*01*
035/*CD59*	*CD59*01*	*CD59*01*
036/*AUG*	*AUG*01*	*AUG*01*

Comparison with NGS

For one sample, a 2 × 150 bp, 34× coverage Illumina WGS dataset was available which was used as a reference. The blood groups were called from both, the Illumina WGS and the targeted PacBio TGS data. In general, both datasets achieve a high concordance of 91.4% (32 of 35 blood group systems, see Table 1). However, with the phased TGS data two major improvements can be achieved compared to NGS. Firstly, especially for paralogous loci, such as *RHCE*, the true allelic configuration, in this case a heterozygous *RHCE*01.01/RHCE*04* or *RHCE*01.02.02* can better be elucidated. Structural variation within this locus in the NGS dataset caused misalignments of the short-reads which obscured the presence of four SNPs in exon 2 of *RHCE* (Supplementary Figure S3). Similarly, a 109 bp insertion in intron 2 of *RHCE* that is known to be C-specific (42), could not be detected in the short-read dataset (Supplementary Figure S4). In case of the paralogous MNS:*GYPA* alleles, a false positive *GYPA*08* call could be corrected in the phased TGS data. Secondly, as observed for the Diego blood group system, the allelic specificity for the TGS data is higher than for NGS, thus the reference allele (NG_007498.1) can be specified more accurately.

DISCUSSION

Targeted enrichment protocols are very well established for short-read sequencing technologies that yield reads of a few hundred base pairs in length, but only a few of such protocols have been reported for long-read technologies. The previously reported protocols, however, either aim at low average fragment sizes of <5 kb (35–38), which does not leverage the current full potential of TGS or require one or even two pulse-field electrophoretic size selections (31–34). Electrophoretic size selection is a severe procedural bottleneck, as costs, time, and labour-intensive working steps that require an additional sample-individual polymerase chain reaction (PCR) (31–35) and clean-ups in between steps are included. These bottlenecks limit the protocol's applicability since the sizing is done per sample with a maximum batch size of four (BluePippin™) or two (SageELF™). A higher-throughput system, the PippinHT™, can size-select up to 22 samples per batch, but is limited to a maximum of 1.5 μg DNA input, which is lower than the amount required by comparable protocols (31–34). Despite this elaborate and costly size selection, the reported average subread length in many protocols is still below 5 kb (31–33). Most protocols do not leverage sample multiplexing for enrichment (31,33,35–38) while one protocol uses intermediate sample

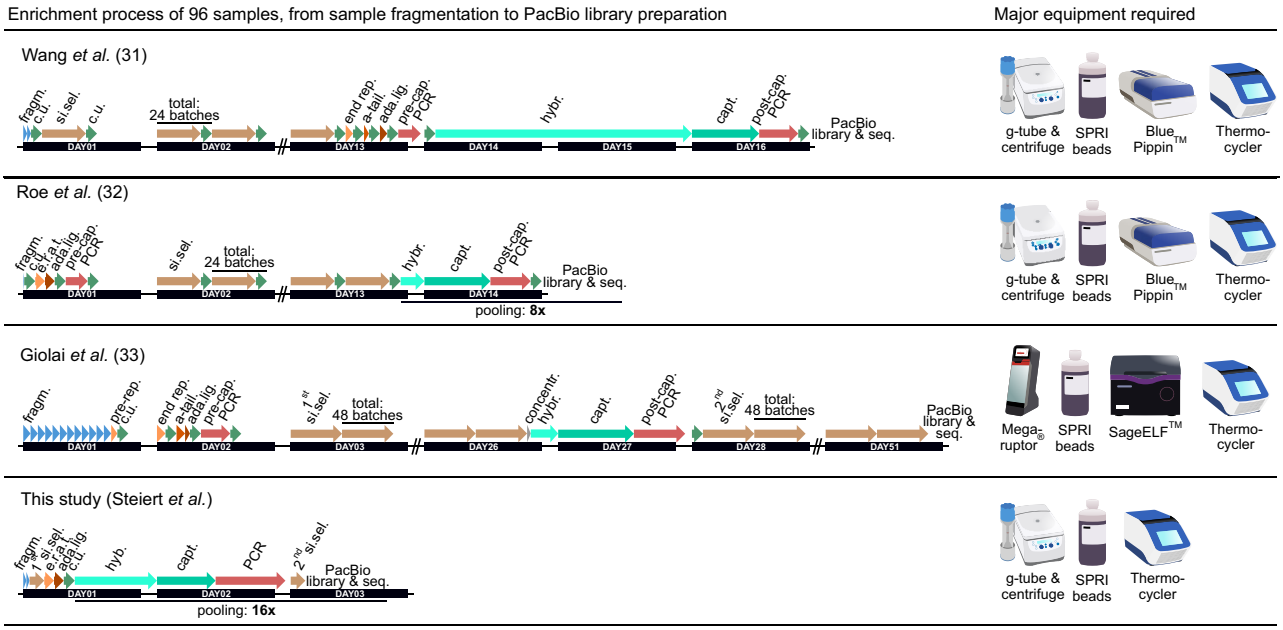


Figure 4. Comparison between protocols for targeted enrichment of long genomic fragments. Workflow for preparation of 96 samples and major equipment required is illustrated for each protocol. Best-case workflow is illustrated from sample fragmentation until enriched fragments that can be subject to PacBio library preparation. Individual process steps for 96 samples are depicted with respect to an eight-hour workday, including hands-on and incubation times. Overnight incubation was considered when possible for the respective step. Enrichment process steps encompass fragmentation (fragm.), bead-based clean-up (c.u.), size selection (si.sel.), end repair (end rep.), A-tailing (a-tail.), combined step of the last two (e.r.a.t.), adapter ligation (ada.lig.), pre-capture PCR (pre-capt. PCR), probe hybridisation (hyb.), streptavidin capture (capt.), post-capture PCR (post-capt. PCR), sample concentration (concentr.) and pre-repair (pre-rep.). Throughput for the introduced protocol is significantly higher compared to all other current comparable approaches reported elsewhere.

multiplexing of eight (32) samples per capture. This additionally drives cost per sample compared to the 16 samples (43) to be pooled in short-read targeted enrichment. For these stated reasons, current protocols are not yet appropriate for high-throughput TGS sequencing.

We here present a fast and cost-effective bead-based size selection (44), that is easily scalable bypasses the electrophoretic size selection bottleneck. This also eliminates the need for investment in electrophoretic size selection devices, making this technique suitable for small-scale laboratories, too. In addition, our protocol eliminates the requirement for a second long-range PCR and associated clean-ups, resulting in a leaner and less biased protocol. Multiplexing of up to 16 samples per capture, which increases capacity, while decreasing cost per sample. The protocol's flexibility allows it to be automated to a large degree (45). Figure 4 shows that for an exemplary throughput of 96 samples significantly less time is needed than for electrophoretic approaches. With the herein presented approach, a batch of 96 samples can be manually enriched within a little more than two workdays. For such a sample batch, beginning with fragmentation until PacBio library preparation, the presented approach reduces the preparation time by approx. 14 (87%), 12 (85%) and 49 workdays (96%), respectively, as compared to other protocols by Wang *et al.* (31), Roe *et al.* (32), and Giolai *et al.* (33). Cost reductions are challenging to quantify; however, in addition to labour cost reduction due to less hands-on time, early sample pooling saves reagent and consumable costs by a factor of 16 (31,33) and two (32) for all subsequent preparation steps. As size selection beads can easily and cost-efficiently be prepared from SPRI beads, high consumable and instrument costs for electrophoretic size selection are not required.

To showcase the applicability and usefulness of our high-throughput method, we applied it to a targeted enrichment of blood group system loci which include difficult-to-genotype targets. Especially *RHD*, *RHCE* and the genetic loci of the NMS blood group system *GYPA*, *GYPB* and *GYPE* are particularly challenging, as paralogous regions and structural variation exist (46,47). Array genotyping, a recently established technique for high-throughput blood group typing, only provides information on specific, non-interconnected blood-group-relevant SNPs from which the phenotypic blood group can be derived. Thus, genotyping arrays have technical problems resolving structural variation and paralogous loci, such as for the RH system (48). Moreover, standardised genotyping chips cannot detect novel variations, nor can they keep up to the evolving field of additional and revised blood group systems as opposed to more flexible spike-in bait panels. On the other hand, blood-group-system-targeted NGS approaches face the challenge that their relatively short reads do not span far enough from probe-targetable unique anchor regions to cover all loci of interest. While whole genome NGS reads cover these regions, however, it is much more expensive, problematic with respect to ethical issues, and even prone to misalignments, coverage drops, and misdetection of structural variation (49). Therefore, there is a high demand for a focussed, fast, and cheap protocol that leverages TGS long-reads to cope with these challenging targets (10,11). Long-reads are advantageous in such contexts since they can bridge regions missing unique bait anchor sequences

or even span the distances between SNPs, which can be used for haplotype phasing. Hybridisation-based targeted enrichment is, however, limited with respect to highly complex or comprehensive, yet unknown allelic variation. For example, to avoid coverage drops the design of bait panels targeting the human leukocyte antigen (HLA) should be optimised with regards to the diversity of alleles reported in related databases and not only towards the human genome reference (50).

We showed that with our TGS-based targeted high-throughput strategy, it is possible to leverage the full potential of TGS PacBio sequencing for a focussed research question. We targeted 35 blood group systems, including systems with challenging genomic loci for 16 patients, of which 9 were previously unknown. The vast majority of SNPs (93.8%) could successfully be phased and not only the phenotypic blood group could be determined, but also the accurate ISBT reference alleles. Average unique on-target HiFi read number (97883), read length (6287 bp) and coverage uniformity (2.64 F80BP) were sufficient. We benchmarked our method using fully elucidated GIAB reference samples, including the Ashkenazim Jewish and the Han Chinese trios. Variant calling concordance was high for GIAB references in terms of recall (96,02%) and precision (98.79%) rates according to the PrecisionFDA Truth Challenge (40). The phenotypic blood group results from genotyping array for eight of the unknown patients perfectly matched the trivialised TGS results. Improvements in typing accuracy were not only achieved in comparison to array genotyping, but also between TGS and whole genome NGS for one patient sample. In this latter patient, a 109 bp insertion in the intron 2 of *RHCE* remained concealed in the NGS data due to short-read misalignments, despite its relatively small genomic span (Supplementary Figure S4). This highlights the technical advantage of TGS as compared to array genotyping and even NGS with respect to the elucidation of structural variants. Analogously, NGS misalignment caused the non-detection of four SNPs in the adjacent exon 2, resulting in allelic mistyping. With respect to the clinical relevance of the *RHCE* locus (47,48) the correct and unbiased elucidation of this region is of particular importance. A higher accuracy and allelic resolution in combination with the possibility to resolve complex structural variation by phased data is of high interest in basic and clinical research (49). With customisable DNA bait panels our approach can easily be adapted to other research questions or clinical settings. With the described protocol significantly reducing the price, hands-on and turn-around time for targeted TGS, we hope to pave the way for a broader adaptation of these technologies in biological research and precision medicine in the future.

DATA AVAILABILITY

Due to the informed consent obtained from the participants the genetic blood group sequencing not all data generated and analysed during the current study can be deposited publicly. Patient data from DHKZ and ITM cohorts are available upon request from P2N biobank. Read Archive, accession no. P2N_XGVGG; http://www.uksh.de/p2n/. P2N is a controlled-access human data repository subject to European data protection laws. Therefore, data access is subject to an application, ethics approval by the applicant's ethics board and a data access agreement. Raw sequencing data of Genome-in-a-Bottle reference samples are publicly available via the European Nucleotide Archive (ENA) with study accession no. PRJEB51372.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Sören Franzenburg, Georg Hemmrich-Stanisak, Teide Boysen and Iacopo Torre for excellent support and maintenance of the bioinformatic infrastructure.
Author contributions: T.A.S. and J.F. designed the study. T.A.S. developed the method based on previous experiments by G.V. and S.J. M.W. performed the majority of bioinformatic analysis and designed the bait panel. T.A.S. performed the targeted enrichment experiment. M.V. performed the PacBio library preparation. T.A.S. analysed the sequencing data. T.A.S. wrote the manuscript and created all figures. All authors reviewed, edited, and approved the final manuscript.

FUNDING

European Union's Horizon 2020 research and innovation program European Advanced infraStructure for Innovative Genomics, **EASI-Genomics** [824110]; German Research Foundation (DFG) as part of the Next Generation Sequencing Competence Network, Competence Centre for Genomic Analysis (**CCGA) Kiel** [423957469].
Conflict of interest statement. The authors have no relevant affiliations or financial involvement with any organisation or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. No writing assistance was utilised in the production of this manuscript.

REFERENCES

1. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
2. International,Human Genome Sequencing Consortium (2001) International human genome sequencing consortium. *Nature*, **409**, 860–921.
3. Waterson,R., Lindblad-Toh,K., Birney,E., Rogers,J. and Abril,J. (2002) Mouse genome sequencing consortium. *Nat. Methods*, **420**, 61–65.
4. Alkan,C., Sajjadian,S. and Eichler,E.E. (2011) Limitations of next-generation genome sequence assembly. *Nature Methods*, **8**, 61.
5. Salzberg,S.L. and Yorke,J.A. (2005) Beware of mis-assembled genomes. *Bioinformatics*, **21**, 4320–4321.
6. Huang,J., Pallotti,S., Zhou,Q., Kleber,M., Xin,X., King,D.A. and Napolioni,V. (2020) PERHAPS: paired-end short Reads-based HAPlotyping from next-generation sequencing data. *Brief. Bioinform.*, **22**, bbaa320.
7. Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.
8. Ura,H., Togi,S. and Niida,Y. (2021) Targeted double-stranded cDNA sequencing-based phase analysis to identify compound heterozygous mutations and differential allelic expression. *Biology*, **10**, 256.

9. Loman,N.J., Misra,R.V., Dallman,T.J., Constantinidou,C., Gharbia,S.E., Wain,J. and Pallen,M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, **30**, 434–439.
10. van Dijk,E.L., Jaszczyszyn,Y., Naquin,D. and Thermes,C. (2018) The third revolution in sequencing technology. *Trends Genet*., **34**, 666–681.
11. Logsdon,G.A., Vollger,M.R. and Eichler,E.E. (2020) Long-read human genome sequencing and its applications. *Nat. Rev. Genet.*, **21**, 597–614.
12. Wetterstrand,K.A. (2021) DNA sequencing costs: data from the NHGRI genome sequencing program (GSP). www.genome.gov/sequencingcostsdata, (10 June 2021, date last accessed).
13. Hu,T., Chitnis,N., Monos,D. and Dinh,A. (2021) Next-generation sequencing technologies: an overview. *Hum. Immunol.*, **82**, 801–811.
14. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S. and Russ,C. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.
15. Shearer,A.E., DeLuca,A.P., Hildebrand,M.S., Taylor,K.R., Gurrola,J., Scherer,S., Scheetz,T.E. and Smith,R.J. (2010) Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21104–21109.
16. Poultney,C.S., Goldberg,A.P., Drapeau,E., Kou,Y., Harony-Nicolas,H., Kajiwara,Y., De Rubeis,S., Durand,S., Stevens,C. and Rehnström,K. (2013) Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *The Am. J. Hum. Genet.*, **93**, 607–619.
17. Calvo,S.E., Compton,A.G., Hershman,S.G., Lim,S.C., Lieber,D.S., Tucker,E.J., Laskowski,A., Garone,C., Liu,S. and Jaffe,D.B. (2012) Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Sci. Transl. Med.*, **4**, 118ra110.
18. Clark,T.A., Chung,J.H., Kennedy,M., Hughes,J.D., Chennagiri,N., Lieber,D.S., Fendler,B., Young,L., Zhao,M. and Coyne,M. (2018) Analytical validation of a hybrid capture–based next-generation sequencing clinical assay for genomic profiling of cell-free circulating tumor DNA. *J. Mol. Diagn.*, **20**, 686–702.
19. Carpenter,M.L., Buenrostro,J.D., Valdiosera,C., Schroeder,H., Allentoft,M.E., Sikora,M., Rasmussen,M., Gravel,S., Guillén,S. and Nekhrizov,G. (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.*, **93**, 852–864.
20. Melnikov,A., Galinsky,K., Rogov,P., Fennell,T., Van Tyne,D., Russ,C., Daniels,R., Barnes,K.G., Bochicchio,J. and Ndiaye,D. (2011) Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.*, **12**, R73.
21. Gaudin,M. and Desnues,C. (2018) Hybrid capture-based next generation sequencing and its application to human infectious diseases. *Front. Microbiol.*, **9**, 2924.
22. Mertes,F., Elsharawy,A., Sauer,S., van Helvoort,J.M., van der Zaag,P.J., Franke,A., Nilsson,M., Lehrach,H. and Brookes,A.J. (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief. Funct. Genomics*, **10**, 374–386.
23. Song,K., Li,L. and Zhang,G. (2016) Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Sci. Rep.*, **6**, 3573.
24. García-García,G., Baux,D., Faugère,V., Moclyn,M., Koenig,M., Claustres,M. and Roux,A.-F. (2016) Assessment of the latest NGS enrichment capture methods in clinical context. *Sci. Rep.*, **6**, 20948.
25. Frampton,G.M., Fichtenholtz,A., Otto,G.A., Wang,K., Downing,S.R., He,J., Schnall-Levin,M., White,J., Sanford,E.M. and An,P. (2013) Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.*, **31**, 1023–1031.
26. Tewhey,R., Nakano,M., Wang,X., Pabón-Peña,C., Novak,B., Giuffre,A., Lin,E., Happe,S., Roberts,D.N. and LeProust,E.M. (2009) Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.*, **10**, R116.
27. Samorodnitsky,E., Jewell,B.M., Hagopian,R., Miya,J., Wing,M.R., Lyon,E., Damodaran,S., Bhatt,D., Reeser,J.W. and Datta,J. (2015) Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Hum. Mutat.*, **36**, 903–914.
28. Severe Covid,G.G., Ellinghaus,D., Degenhardt,F., Bujanda,L., Buti,M., Albillos,A., Invernizzi,P., Fernández,J., Prati,D., Baselli,G. *et al.* (2020) Genomewide association study of severe covid-19 with respiratory failure. *New Eng. J. Med.*, **383**, 1522–1534.
29. Zook,J.M., Catoe,D., McDaniel,J., Vang,L., Spies,N., Sidow,A., Weng,Z., Liu,Y., Mason,C.E. and Alexander,N. (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, **3**, 160025.
30. International Society of Blood Transfusion (2021) Table of blood group systems v. 90 03-FEB-2021. https://www.isbtweb.org/fileadmin/user_upload/Table_of_blood_group_antigens_within_systems_v9.0_03-FEB-2021.pdf, (12 July 2021, date last accessed).
31. Wang,M., Beck,C.R., English,A.C., Meng,Q., Buhay,C., Han,Y., Doddapaneni,H.V., Yu,F., Boerwinkle,E. and Lupski,J.R. (2015) PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genom.*, **16**, 214.
32. Roe,D., Williams,J., Ivery,K., Brouckaert,J., Downey,N., Locklear,C., Kuang,R. and Maiers,M. (2020) Efficient sequencing, assembly, and annotation of human KIR haplotypes. *Front. Immunol.*, **11**, 582927.
33. Giolai,M., Paajanen,P., Verweij,W., Percival-Alwyn,L., Baker,D., Witek,K., Jupe,F., Bryan,G., Hein,I. and Jones,J.D. (2016) Targeted capture and sequencing of gene-sized DNA molecules. *Biotechniques*, **61**, 315–322.
34. Pacific Biosciences (2021) Procedure & checklist – multiplexed genomic DNA target capture using IDT xGen® lockdown® probes. https://www.pacb.com/wp-content/uploads/Procedure-Checklist-%E2%80%93-Multiplex-Genomic-DNA-Target-Capture-Using-IDT-xGen-Lockdown-Probes.pdf, (14 July 2021, date last accessed).
35. Lefoulon,E., Vaisman,N., Frydman,H.M., Sun,L., Voland,L., Foster,J.M. and Slatko,B.E. (2019) Large enriched fragment targeted sequencing (LEFT-SEQ) applied to capture of Wolbachia genomes. *Sci. Rep.*, **9**, 5939.
36. Karamitros,T. and Magiorkinis,G. (2015) A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic Acids Res*., **43**, e152.
37. Eckert,S.E., Chan,J.Z.-M. and Houniet,D. (2016) Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. *Microb. Genom.*, **2**, e000087.
38. Bethune,K., Mariac,C., Couderc,M., Scarcelli,N., Santoni,S., Ardisson,M., Martin,J.F., Montúfar,R., Klein,V. and Sabot,F. (2019) Long-fragment targeted capture for long-read sequencing of plastomes. *Appl. Plant Sci.*, **7**, e1243.
39. Wenger,A.M., Peluso,P., Rowell,W.J., Chang,P.-C., Hall,R.J., Concepcion,G.T., Ebler,J., Fungtammasan,A., Kolesnikov,A. and Olson,N.D. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, **37**, 1155–1162.
40. Olson,N.D., Wagner,J., McDaniel,J., Stephens,S.H., Westreich,S.T., Prasanna,A.G., Johanson,E., Boja,E., Maier,E.J. and Serang,O. (2021) precisionFDA truth challenge V2: calling variants from short-and long-reads in difficult-to-map regions. *Cell Genom.*, **2**, 100129.
41. Patel,Z.H., Kottyan,L.C., Lazaro,S., Williams,M.S., Ledbetter,D.H., Tromp,G., Rupert,A., Kohram,M., Wagner,M. and Husami,A. (2014) The struggle to find reliable results in exome sequencing data: filtering out mendelian errors. *Front. Genet.*, **5**, 16.
42. Carritt,B., Kemp,T. and Poulter,M. (1997) Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled. *Hum. Mol. Genet.*, **6**, 843–850.
43. MacConaill,L.E., Burns,R.T., Nag,A., Coleman,H.A., Slevin,M.K., Giorda,K., Light,M., Lai,K., Jarosz,M. and McNeill,M.S. (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genom*., **19**, 30.
44. Stortchevoi,A., Kamelamela,N. and Levine,S.S. (2020) SPRI Beads-based size selection in the range of 2-10kb. *J. Biomol. Tech.*, **31**, 7.
45. Tegally,H., San,J.E., Giandhari,J. and de Oliveira,T. (2020) Unlocking the efficiency of genomics laboratories with robotic liquid-handling. *BMC Genom*., **21**, 729.

46. Gleadall,N.S., Veldhuisen,B., Gollub,J., Butterworth,A.S., Ord,J., Penkett,C.J., Timmer,T.C., Sauer,C.M., Van Der Bolt,N. and Brown,C. (2020) Development and validation of a universal blood donor genotyping platform: a multinational prospective study. *Blood Adv.*, **4**, 3495–3506.
47. Wheeler,M.M., Lannert,K.W., Huston,H., Fletcher,S.N., Harris,S., Teramura,G., Maki,H.J., Frazar,C., Underwood,J.G. and Shaffer,T. (2019) Genomic characterization of the RH locus detects complex and novel structural variation in multi-ethnic cohorts. *Genet. Med.*, **21**, 477–486.
48. Haer-Wigman,L., Veldhuisen,B., Jonkers,R., Lodén,M., Madgett,T.E., Avent,N.D., de Haas,M. and van der Schoot,C.E. (2013) RHD and RHCE variant and zygosity genotyping via multiplex ligation–dependent probe amplification. *Transfusion*, **53**, 1559–1574.
49. Bizjan,B.J., Katsila,T., Tesovnik,T., Šket,R., Debeljak,M., Matsoukas,M.T. and Kovač,J. (2020) Challenges in identifying large germline structural variants for clinical use by long read sequencing. *Comput. Struct. Biotechnol. J.*, **18**, 83–92.
50. Wittig,M., Anmarkrud,J.A., Kässens,J.C., Koch,S., Forster,M., Ellinghaus,E., Hov,J.R., Sauer,S., Schimmler,M. and Ziemann,M. (2015) Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res.*, **43**, e70.