Transfusion Medicine and Hemotherapy

Review Article

Transfus Med Hemother
DOI: 10.1159/000530652

Received: February 7, 2023
Accepted: April 3, 2023
Published online: June 16, 2023

Long-Read Sequencing in Blood Group Genetics

Gian Andri Thun Morgan Gueuning Maja P. Mattle-Greminger

Department of Research and Development, Blood Transfusion Service Zurich, Swiss Red Cross, Schlieren, Switzerland

Keywords

Haplotype · Third-generation sequencing · Blood group allele · Long-read sequencing · Structural variation

Abstract

Background: The key advantages of latest third-generation long-read sequencing (TGS) technologies include the ability to resolve long haplotypes and to characterize genomic regions that are challenging to analyze with short-read sequencing. Recent advancements in TGS technologies have significantly improved accuracy, a crucial requirement for the transition from research to diagnostic applications. ***Summary:*** In the field of immunohematology, the adoption of TGS is still in its early stages and published applications are scarce. An undeniable utility of TGS in blood group genomics is the ability to resolve ambiguous genotype-phenotype blood group results. In particular, hybrid genes and other large structural variants, as commonly found in the RHD/CE and MNS blood group systems, cause such discrepant results that can hardly be resolved by conventional methods. Long-read sequencing also greatly aids to generate high-standard reference alleles, establish haplotype sequence databases, or could even serve for high-resolution genotyping of all blood groups in parallel. Additionally, TGS holds the potential to close important knowledge gaps in blood group transcriptomics and epigenetics. ***Key Messages:*** The aims of this review were to examine the prospects of TGS technologies within the field of immunohematology and to highlight practical applications. Furthermore, we present a comprehensive overview of the existing and emerging wet-laboratory strategies for data generation, as well as a summary on bioinformatic data analysis methods. Finally, we provide an outlook on anticipated advancements in the near future.

Introduction

More than 15 years ago, the advent of next-generation sequencing technologies, or retrospectively more accurate second-generation sequencing, had revolutionized DNA sequencing through enabling cheap sequencing of short fragments in a highly parallelized way [1]. This major technological breakthrough also opened up new avenues in the field of immunohematology [2–5]. Using, for instance, target-specific gene amplification or DNA-fragment capturing protocols, short-read sequencing enabled sequencing a multitude of blood group genes in parallel [6–9]. Despite its great benefits, applications of second-generation sequencing in immunohematology suffered from its intrinsic limitation: the short read length. Especially, the highly homologous genomics regions underlying the RHD/CE and MNS blood group systems proved challenging [10]. These complex regions harbor many structural variants (SVs) including hybrid genes that are difficult to analyze with short-read sequencing.

Today, so-called third-generation sequencing (TGS) technologies, also known as long-read sequencing, are again transforming the way of sequencing by producing (ultra-)long, single-stranded sequences. By this, they allow inferring unambiguous haplotypes, i.e., determining the

karger@karger.com
www.karger.com/tmh

Correspondence to:
Maja P. Mattle-Greminger, m.mattle@zhbsd.ch

combination of alleles that an individual inherited from the mother or the father. They also permit identifying SVs or sequencing complex genomic regions. Thanks to the great progress TGS technologies have recently made in terms of sequence accuracy [11, 12], they are now at the doorstep of moving from research applications into diagnostic fields.

In this review, we discuss the power of long-read sequencing technologies for the field of immunohematology and how it already has been applied. We outline current and emerging methodological strategies for wet-laboratory data generation as well as provide insights into bioinformatic data analyses. Finally, we provide perspectives on further applications of TGS in immunohematology and future developments.

TGS Technologies in a Nutshell

There are currently two TGS technologies on the market, one developed by Pacific Biosciences (PacBio) and the other by Oxford Nanopore Technologies (ONT). We briefly summarize both technologies in the following sections and refer to more technical reviews elsewhere [12, 13]. An overview of key specifications of current TGS systems is provided in Table 1.

Sequencing by PacBio

PacBio introduced the first TGS technology in 2011, calling it single-molecule real-time (SMRT) sequencing. Hairpin adapters, ligated to both ends of digested double-stranded DNA, enable the formation of a single-stranded circular DNA template, which is then individually replicated by an anchored DNA-polymerase complex in each nanowell of a flow cell [15]. Similar to some short-read sequencing technologies, the sequence is generated from optical signals by synthesizing a complimentary DNA strand with fluorescently labeled nucleotides. High accuracy is achieved by sequencing both strands multiple times in a circle before computing the circular consensus sequence, which represents the single read [12]. Owing to the finite endurance of the polymerase, this currently works well for ~15 kb average read length. Current systems comprise the well-established sequel II/IIe, outputting ~25 Gb per day, as well as the novel Revio, which can produce up to four human genomes a day (360 Gb, Table 1). The systems are laid out for high-throughput settings, also given the high costs for machine acquisition.

Sequencing by ONT

The first sequencer launched by ONT was the pocket-size MinION in 2014. In this technology, flow cells contain nanopores fixed on a membrane. Single-stranded molecules of DNA or RNA, carrying previously ligated sequence adapters with attached motor proteins, are guided through these nanopores. Ionic currents in the nanopore resulting by a membrane potential are disrupted in distinctive ways depending on the nucleotides (or other molecules) passing the pore. These electrical signals can be translated in real time into a nucleotide sequence (single read). Read length is only limited by the input fragment length and can therefore reach megabase scale. Regarding throughput, nanopore sequencing is highly scalable as ONT offers solutions from inexpensive flow cells (Flongle) for single-sample analysis on portable sequencers up to larger sequencing machines designed to accommodate several flow cells (GridION and PromethION, Table 1) allowing sequencing entire human genomes [16]. Typical output for the mid-size MinION device is up to ~20 Gb per day (Table 1), whereas the high-end PromethION 48 system can produce up to 50 human genomes per day.

Sequencing Accuracy

For all sequencing technologies, two main types of sequencing accuracy need to be differentiated: the single-read accuracy (accuracy of each single read) and the consensus-read accuracy (bioinformatically generated based on all reads covering the same genomic region). High single-read error rates of 10–15% were for a long time the hallmark of TGS technologies [17, 18]. However, both technologies have improved greatly over time, achieving now a high level of both single-read and consensus-read accuracies. Hence, the notion that TGS is currently too error-prone for clinical diagnostics is meanwhile a myth that is, because of the high initial error rates, fading away only slowly.

PacBio's single-read accuracy soared when circular consensus sequencing became feasible for long templates, which resulted in the term high-fidelity (HiFi) read for long reads with Q>20 (99%) [12]. Currently, average HiFi read accuracy is at 99.95% (Q33) (Table 1) [19]; consensus-read accuracy is even higher (Q>50) [19]. In a variant calling benchmarking study [20], HiFi sequencing even outperformed the accuracy of Illumina's short-read technology (Q>30 for most reads), the current gold standard for single-nucleotide variant (SNV) calling with next-generation sequencing technologies.

The increase of ONT's read accuracy is mainly attributed to continuous developments in machine learning algorithms [21] as well as improvements on flow cells and chemistry. With the newest flow cell, single-read accuracy currently reaches 99.35% (Q22) (Table 1) and consensus sequence accuracy Q>45 [22]. Indel calling sensitivity and precision, however, still lack behind (Table 1) [20]. The lower indel accuracy is mainly attributed to long homopolymers (sequence stretches of the same base pair) as the signal in the nanopore is determined from five to nine consecutive nucleotides combined and translocation speed is not constant. Promising developments to close this gap are underway (see *Perspectives*).

Table 1. Specifications of currently available TGS platforms. The table depicts the technical aspects of both PacBio's and ONT's currently available platforms, as well as output, cost, and possible applications for each platform

Manufacturer	System	Typical read length[1]	Single-read accuracy[2], %	Variant calling (F1 score)[3]			Methylation accuracy (5mC)[4], %	Direct RNA single-read accuracy, %
				SNV, %	Indel, %	SV, %		
PacBio	Sequel IIe	~15–25 kb	99.95	99.95	99.41	95.19	~90	n/a
	Revio	~15–18 kb			99.44	95.59		
ONT	Flongle	~20–50 kb; ultra long: >100 kb	99.35; duplex: 99.9	99.90	89.40	96	99.5	~90
	MinION							
	GridION							
	PromethION 2s/24/48							

Manufacturer	System	Maximal throughput[5]		Run time	Costs in USD[6]			Typical application for single sample (ss) or multiple samples (ms)
		flow cell	device		system access	flow cell	per Gb	
PacBio	Sequel IIe	30 Gb	30 Gb	30 h	~500k	1,300	43	WGS (ss); targeted-seq (ms)
	Revio	90 Gb	360 Gb	24 h	~780k	995	11	WGS (ms)
ONT	Flongle	2.8 Gb	2.8 Gb	Up to 16 h	1.6k	90	32	Targeted-seq (ss)
	MinION	50 Gb	50 Gb	Up to 72 h	1.0k	675	14	Targeted-seq (ss/ms)
	GridION	50 Gb	250 Gb		50k	675	14	WGS (ss); targeted-seq (ms)
	PromethION 2s/24/48	290 Gb	0.6/7/14 Tb		10k/225k/310k	980	3	WGS (ms)

SNV, single-nucleotide variant; Indel, insertion/deletion; SV, structural variant; WGS, whole-genome sequencing. [1]Refers to WGS. Maximal read length for PacBio's HiFi reads is ~70 kb and for ONT >4 Mb. [2]Accuracies advertised by manufacturers; PacBio accuracy is after circular consensus sequencing (i.e., for processed HiFi reads); ONT accuracy is based on R10.4.1 flow cell, kit 14, "super accuracy" mode. [3]Variant calling advertised by manufacturers; based on 30x coverage for PacBio and 60x for ONT. The F1 score is the harmonic mean of precision (i.e., true positives/total predicted positives) and recall (i.e., true positives/total actual positives). [4]Accuracy for PacBio is based on Tse et al. [14] (not directly provided by manufacturer). Accuracy for ONT was advertised by the manufacturer. [5]Throughput for PacBio is given after circular consensus sequencing (i.e., for processed HiFi reads). [6]Prices advertised by manufacturers but will vary according to number of flow cells ordered, countries, applications, library preparation, and throughput. The system access encompasses all costs for acquiring the platforms.

Why Do We Need Long-Read Sequencing in Immunohematology?

The main strength of long-read sequencing is that it finally enables determining long haplotypes and characterizing genomic regions that are difficult to resolve with short reads. This includes, for instance, low complexity regions as well as sequences spread to different locations in the genome by duplication events as is the case of paralogous genes (e.g., *RHD*/*CE*, *GYPA*/*B*/*E*). Long-read sequencing allows to unambiguously identify SVs, i.e., variation that affects more than 50 bp encompassing long insertions, deletions, duplications, inversions, as well as translocations.

These key advantages of TGS promise essential progress for the accurate molecular characterization of blood group antigens. Genetic diversity of blood group systems is very high and identification of novel alleles occurs at steadily increasing pace. Currently, 44 blood group systems with over 350 antigens and thousands of blood group alleles underlying these antigens are recognized by the International Society of Blood Transfusion (ISBT) [23–25]. Long-read sequencing offers great power in particular for (1) uncovering the genetic basis of phenotypes which cannot be resolved with conventional methodology, (2) resolving complex genotype-phenotype discrepancies, which are often associated with novel alleles, (3) sequencing of novel blood group alleles as full-length haplotypes to obtain high-standard reference alleles, (4) establishing haplotype sequence databases for all blood group systems. In the following, we illustrate in more detail how the aforementioned assets of TGS can serve in immunohematology.

Excursus: Concept of Resolving Sequences as Haplotypes

Until the advent of TGS, resolving haplotype information has been a major challenge. Since humans are diploid, our cells carry a maternal and a paternal copy of each chromosome. Such a maternally or paternally derived sequence is called a haplotype. The haplotype concept is illustrated in Figure 1, showing an example of a genomic region with three heterozygous genetic variants. With this trivial example, there are already eight different combination possibilities for the maternal and paternal haplotype, respectively. Resolving haplotypes requires phasing of the genetic variants, i.e., determining which allele of the variants lies on the same chromosome (in cis) and thus constitutes one haplotype and which lies in trans on two different haplotypes.

Classical Sanger sequencing does not allow phasing due to the overlapping signals in the chromatogram (Fig. 1), except in the case of laborious allele-specific polymerase chain reactions (PCRs) [26]. Haplotype reconstruction from second-generation sequencing data works only marginally better as the distance between two or more heterozygous variants must not exceed the read length, which is maximally 400 bases depending on the short-read technology. In contrast, phasing with read lengths >10 kb, as provided by TGS, works well as the average distance between heterozygous positions is usually much smaller. In this way, haplotypes for entire genes can be generated and even much larger haplotype blocks of several megabases can be assembled, which built the scaffold for completely phased telomere to telomere reference chromosomes [27, 28].

Reference Sequences and Haplotype Sequence Collections

Being able to resolve blood group gene alleles as full-length haplotype sequences offers great potential for many aspects of blood group genetics. So far, such sequences remained scarce [29]. For the vast majority of blood group alleles currently curated by the ISBT, sequence data are limited to exonic gene regions and are lacking phase information. With TGS technologies, sequencing (new) blood group alleles as complete gene haplotypes could become the emerging standard [30, 31].

Also, the availability of entire haplotype sequence collections of blood group gene alleles has gained importance [29]. Comprehensive collections of population-specific haplotypes build, for example, the base of statistical phasing. This is required for inferring haplotypes from genotyping data based on probability as well as for imputing unmeasured genetic variants from genome-wide array data [32]. Such array data are getting increasingly produced in blood donor screening programs [33, 34]. Comprehensive haplotype sequence collections would also greatly aid in designing and validating blood group genotyping assays to reduce risks of unnoticed allelic dropout [35, 36]. Furthermore, genetic diversity patterns identified with haplotype collections can greatly help in resolving complex genotype-phenotype discrepancies in routine diagnostics (e.g., defining breakpoints of hybrid genes) [29]. Finally, such haplotype collections would lower error rates in bioinformatic analyses of reference-based read mapping (see Bioinformatic Analyses of Sequencing Data) [37].

Improved Phenotype Prediction

In blood group genetics, direct phasing of novel variants (e.g., variants causing null alleles) to the respective allelic background enables improved blood group phenotype prediction [38, 39]. Indeed, haplotype-resolved sequences are often crucial for interpreting functional consequences of detected variants [40–42]. For instance, variants causing deleterious gene expression or affecting protein function may silence one haplotype while the other remains intact. Such variants, if located in trans, could even affect both haplotypes, resulting in compound heterozygosity, which often alters phenotypes much more seriously [40–42].

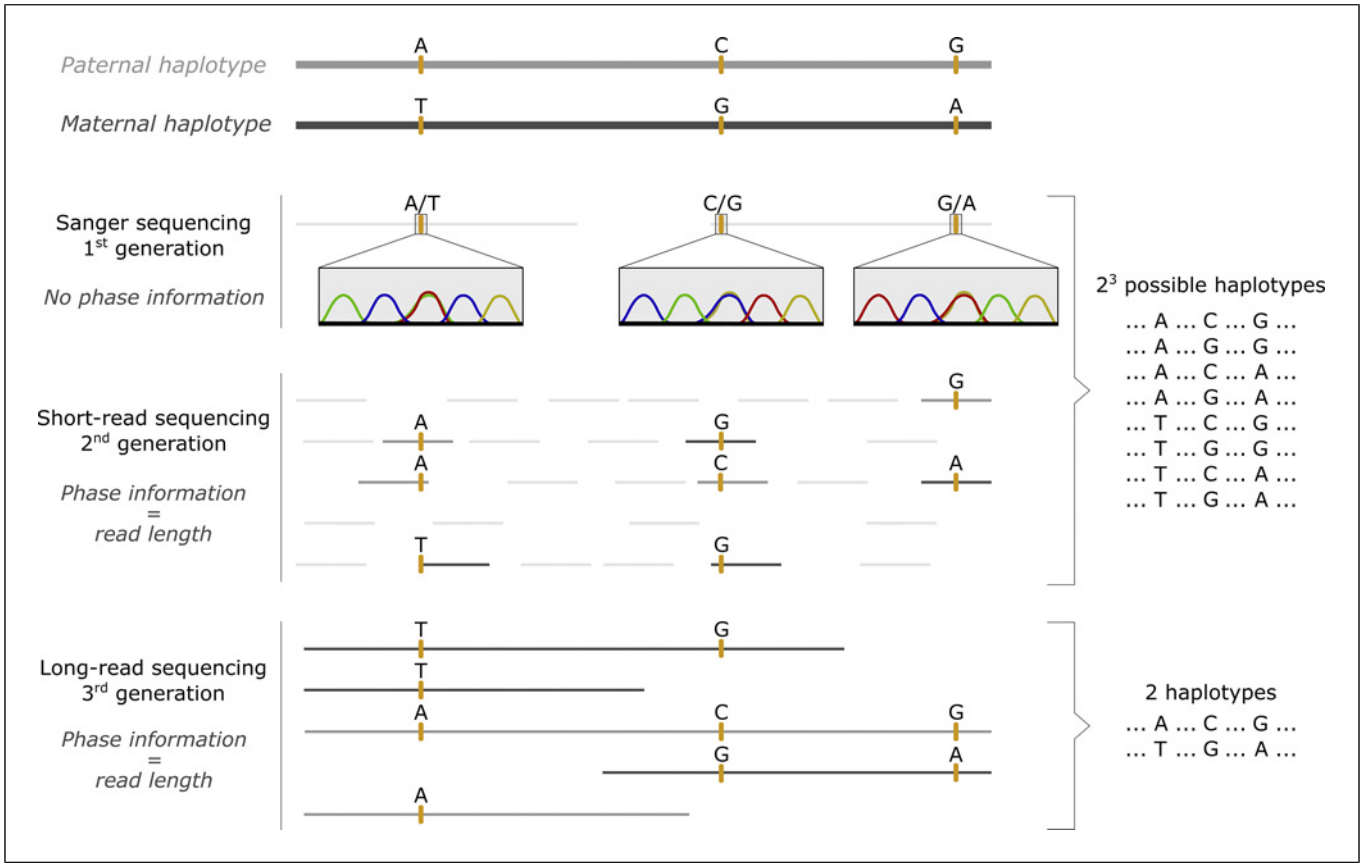

Fig. 1. Graphical illustration of the haplotype phasing concept. Paternal and maternal haplotypes of a genomic region harboring three SNVs are shown at the top. For each generation of sequencing technology, the obtained sequence data for this genomic region are illustrated. With classical Sanger sequencing, the haplotype phase cannot be resolved due to overlapping signals in the chromatogram (gray box). In short-read sequencing, physical phase information is limited by read length. Hence, for both of these technologies, the eight possible haplotype combinations of the three SNVs cannot be resolved. Using TGS, the maternal and paternal haplotype can be reconstructed thanks to overlap between longer reads.

Resolving Challenging Loci: Paralogous Genes and Hybrid Alleles

Long-read sequencing technologies have particularly high potential for resolving challenging loci such as paralogous genes and hybrid alleles, as found in the most complex blood group systems RHD/CE and MNS. The *RHD* and *RHCE* genes are both over 50 kb long and consist of 10 exons each, arranged in an antiparallel way, with complete concordance of exons 8, 10, and, in the case of the common *RHCE*C* allele, exon 2. Their encoded transmembrane proteins RhD and RhCE have 92% sequence similarity [43]. With well over 50 antigens and hundreds of known alleles encoding weak (i.e., reduced expression) and partial (i.e., missing epitopes) phenotypes, the RHD/CE system is highly polymorphic and shows ethnically highly diverse allele frequencies. One of the major difficulties in sequencing *RHD/CE* is the frequent occurrence of SVs such as hybrid alleles that have arisen from translocation events.

The MNS blood group is another notoriously challenging system containing over 50 reported antigens with ethnically distinct allele frequencies. The system is based on the homologous genes *GYPA* and *GYPB*, encoding glycophorin A and B proteins. A third homologous gene, *GYPE*, is usually silent but can be involved in rearrangements resulting in hybrid gene alleles [44]. Those also occur between *GYPA* and *GYPB* [45], which show genetic sequence similarity of 95% across the large homologous part of the genes [46]. Similar to the RHD/CE system, there are fully identical exons between genes (e.g., exon 2 on *GYPA*M* and *GYPE* [47]).

Currently, discrepant results between geno- and phenotyping in these blood group systems cannot unambiguously be resolved with conventional molecular methods. Sanger sequencing, which is state-of-the-art methodology for diagnostic SNV detection, is inappropriate to find structural variation or build haplotypes. Short-read sequencing approaches generally suffer from ambiguous read mapping due to the high sequence similarity of the paralogues genes. Notwithstanding this, they helped characterizing alleles for the RHD/CE system involving great efforts [10, 48–51] and construct *RHD* haplotypes from hemizygous *RHD* samples, where no phasing is required [52].

Resolving Other SVs

Long-read sequencing holds great potential for the identification and characterization of SVs which have remained hidden by traditional Sanger and short-read sequencing [53]. Namely, long reads reduce the risk of missing SVs by spanning over one or both of the breakpoints. A recent example of this is the detection of a large, ~5 kb deletion in the Kidd blood group system by ONT sequencing [38]. Sanger sequencing only amplified the wildtype allele as primers failed in the deleted region of the other allele.

At the genomic scale, TGS has demonstrated that human genomes harbor significantly more SVs than previously estimated [54]. Given that SVs are responsible for an estimated ~30% of rare heterozygous (allele frequency <1%) gene inactivation events per individual [55], it is striking that very few SVs have been reported in blood group genes other than belonging to the RHD/CE and MNS systems. This apparent absence of SVs, apart from rearrangements between homologous genes, is likely at least partly a result of inappropriate methodology. With the advancement of TGS, it is expected that a larger number of SVs will be uncovered in diverse blood group genes.

A Field in Its Infancy: Published Applications of TGS in Blood Group Genetics

Despite the great potential in the field of immunohematology, published applications of TGS are still scarce (Table 2). There is, however, a wide variety of how they make use of the advantages of TGS. For instance, the authors of the first paper ever reporting usage of TGS on blood groups used PacBio sequencing to confirm several novel *ABO* alleles found by short-read sequencing [9]. In this study, Lang and colleagues argue that the use of PacBio sequencing as an orthogonal technology allows for the exclusion of systematic sequencing errors as the error profile is vastly different from the Illumina platform.

Since the introduction of TGS in immunohematology, long-read sequencing has been employed several times to resolve ambiguities not addressable by conventional methods. For instance, Lane and colleagues [47] used a whole-genome sequencing (WGS) approach on ONT's PromethION to resolve a complex case of a large, compound heterozygous deletion in the MNS blood group system of a rare U-individual. Similarly, Montemayor et al. [56] used ONT sequencing to phase variants they detected with short-read sequencing in a genotype-phenotype discrepant case in the Kidd system. Also, Gueuning et al. [38] could resolve complex genotype-phenotype discrepancies in the Kidd system by TGS and Thun et al. [39] in the ABO blood group. These four studies clearly show the benefit of sequencing and phasing blood group gene haplotypes across their entire length (including introns and regulatory elements) using long-read sequencing.

Other studies listed in Table 2 used TGS as a main sequencing tool for characterizing either single or several full-length blood group genes. For instance, two studies used TGS to define reference alleles for the short *ACKR1* gene (Duffy blood group). While Fichou and colleagues [30] used a combination of long-range PCR and PacBio sequencing to define reference alleles for a large number of samples of the *ACKR1* gene, Srivastava et al. [31] used long-range PCR amplicon sequencing on ONT to tackle a similar question. Nanopore sequencing has lately also been used as primary tool for building a large collection of fully resolved *ABO* haplotypes by Gueuning et al. [29]. In a proof-of-principle study, Tounsi et al. [58] performed amplicon-based sequencing of the *RHD* gene locus. Finally, Zhang et al. [59] and Steiert et al. [57] have developed hybridization capture-based protocols for characterizing the complete *RHD/CE* locus and all known blood group genes, respectively.

Methodological Strategies for Sequencing Data Generation

There are two conceptually alternative ways to produce TGS genomic data for blood group genes (Fig. 2), namely, WGS and targeted-gene sequencing, and their application will depend on the specific question, budget, and resources. In the following sections, we outline different wet-laboratory strategies, their strengths and weaknesses, as well as their utility for transfusion medicine.

Whole-Genome Sequencing

In the WGS strategy, the entire genome is sequenced at equal coverage without targeting specific loci (e.g., blood group genes). This strategy is the most straightforward for detecting variation without prior knowledge on its location and, when carried out with TGS, particularly powerful to resolve SVs and large repetitive elements. Compared to targeted approaches, sequencing library preparation for WGS is less demanding. Specific DNA extraction protocols are available with emphasis on securing long fragments from shearing forces, which is particularly valuable for nanopore sequencing where read length is not limited by technology but only by DNA template integrity.

One key limitation of WGS is the legal restrictions imposed in many countries due to ethical concerns regarding person's privacy and confidentiality [61, 62]. Furthermore, 30× long-read WGS remains costly (>1,000 Euro/genome) and ideally requires access to high-throughput sequencing platforms (Table 1). Given that for blood group diagnostic purposes only a tiny fraction of the genome is of interest (the entire "blood group genome" is ~2.5 Mb), costs per targeted base pair is particularly high. Finally, the large

Table 2. Overview of published articles using TGS on blood group genes. This list encompasses to the best of our knowledge all current articles

Reference	Publication title	ISBT blood group system	Targeted region by TGS	Samples[a]	TGS platform	Wet-laboratory strategy	Purpose of TGS
Lang et al. [9] (2016)	*ABO* allele level frequency estimation based on population-scale genotyping by NGS	ABO	5.8 kb	37	PacBio (RS II)	Long-range PCR	Confirmation of novel alleles found by short-read sequencing
Fichou et al. [30] (2020)	Defining blood group gene reference alleles by long-read sequencing: proof of concept in the *ACKR1* gene encoding the Duffy antigens	FY	3 kb	81	PacBio (sequel I)	Long-range PCR	Full-gene reference sequences/haplotypes
Lane et al. [47] (2020)	Multiple *GYPB* gene deletions associated with the *U*-phenotype in those of African ancestry	MNS	625 kb	1	ONT (PromethION)	WGS	Resolve a sample with genotype-phenotype discrepancy
Srivastava et al. [31] (2020)	*ACKR1* alleles at 5.6 kb in a well-characterized renewable US Food and Drug Administration (FDA) reference panel for standardization of blood group genotyping	FY	5.6 kb	53	ONT (GridION)	Long-range PCR	Full-gene reference sequences/haplotypes
Montemayor et al. [56] (2021)	An open-source python library for detection of known and novel Kell, Duffy, and Kidd variants from exome sequencing	JK	9.6 kb with 2 amplicons [5.1 kb; 7.1 kb]	1	ONT (MinION)	Long-range PCR	Phasing of a sample with genotype-phenotype discrepancy
Gueuning et al. [29] (2022)	Haplotype sequence collection of *ABO* blood group alleles by long-read sequencing reveals putative *A1*-diagnostic variants	ABO	23.6 kb with 2 amplicons [16.9 kb; 13.2 kb]	77	ONT (MinION)	Long-range PCR	Generation of full-length haplotype sequence collection
Steiert et al. [57] (2022)	High-throughput method for the hybridization-based targeted enrichment of long genomic fragments for PacBio THS	35 blood groups +2 transcription factors	2.2 Mb [5.9–6.8 kb][b]	16	PacBio (sequel II)	Hybridization capture	Targeted TGS of all blood group genes
Tounsi et al. [58] (2022)	Rh blood group D antigen genotyping using a portable nanopore-based sequencing device: proof of principle	RHD	58.5 kb with 6 amplicons [9.9–13.7 kb]	13	ONT (MinION)	Long-range PCR	Variant analysis of *RHD*
Zhang et al. [59] (2022)	Accurate long-read sequencing allows assembly of the duplicated *RHD* and *RHCE* genes harboring variants relevant to blood transfusion	RHD/CE	166.4 kb [2.1–2.9 kb][b]	11	PacBio (sequel I)	Hybridization capture	Assembly of *RHD/CE* locus to resolve SVs

Table 2 (continued)

Reference	Publication title	ISBT blood group system	Targeted region by TGS	Samples[a]	TGS platform	Wet-laboratory strategy	Purpose of TGS
Gueuning et al. [38] (2023)	Resolving genotype-phenotype discrepancies of the Kidd blood group using nanopore sequencing	JK	24.5 kb with 2 amplicons [13 kb; 13.1 kb]	10	ONT (MinION)	Long-range PCR	Resolve samples with genotype-phenotype discrepancy
Thun et al. [39] (2023)	Novel regulatory variant in *ABO* intronic RUNX1 binding site inducing A3 phenotype	ABO	23.6 kb with 2 amplicons [16.9 kb; 13.2 kb]	3	ONT (MinION)	Long-range PCR	Resolve a sample with genotype-phenotype discrepancy
Ji et al. [60] (2023)	Patients with Asian-type DEL can safely be transfused using RhD-positive blood	RHD	1.5 kb	8	ONT (MinION)	Long-range PCR	Characterization of transcript repertoire in DEL phenotype

[a]Number of samples sequenced with TGS (not always full-length gene sequences or resolved haplotypes). [b]Average length of HiFi consensus sequences.

amount of sequencing data produced renders analyses lengthy and computationally intense and additionally includes hidden costs of data storage.

Targeted-Gene Sequencing

Targeted-gene sequencing strategies focus on the enrichment of specific genomic regions either by PCR amplification, capturing, or other isolation and enrichment methods (Fig. 2). By targeting a small fraction of the genome, they are more cost-effective than WGS and scalable for routine purposes. In principle, both PacBio and ONT allow pooling hundreds of samples per flow cell (multiplexing) using unique molecular identifiers (barcodes).

We describe here four different targeted-gene TGS strategies for the field of immunohematology, three of which are currently emerging approaches. Key differences among methods include the integration of PCR amplification, maximal read length, achieved coverage, cost-efficiency by pooled sequencing of several samples, and preservation of epigenetic base modifications. Targeted-exome sequencing approaches are not considered here as there is limited benefit of using TGS solely for exons, which are predominantly shorter than 200 base pairs.

Amplicon Sequencing

Sequencing long-range PCR amplicons of the target region is currently the most straightforward TGS strategy in transfusion medicine (Fig. 2). Besides low input requirement, this approach is flexible and cost-effective, especially if several samples are pooled by multiplexing on one flow cell. Sequencing can be performed at high depth of coverage, which increases sensitivity to detect variants with unequal allele distribution. The maximum amplicon length achieved by long-range PCR is highly dependent on PCR components and conditions, DNA integrity, and template complexity. So far, blood group gene amplicon sizes ranging from 3 to 17 kb were generated (Table 2). For most blood group genes, full-length sequencing requires several overlapping PCR amplicons (~60% of the blood group genes are >20 kb) [23, 25, 63]. For instance, complete sequencing of the *RHD* gene (~58 kb) requires around six overlapping long-range PCR amplicons [52]. Shorter genes like *ABO* (~25 kb) need two fragments [29]. Importantly, PCR amplicons need to overlap to enable resolving full-length haplotypes. Required size of the overlapping region depends on its genetic variability as heterozygous variation is necessary for phase reconstruction. Because of the number of amplicons required, sequencing large genes in many samples quickly gets unpractical. Hence, amplicon sequencing is most suitable for single gene analyses. Another caveat of amplicon sequencing is that amplification of alleles harboring SNVs located in the PCR primer annealing sites can lead to allelic biases or in worst case allelic dropout [35].

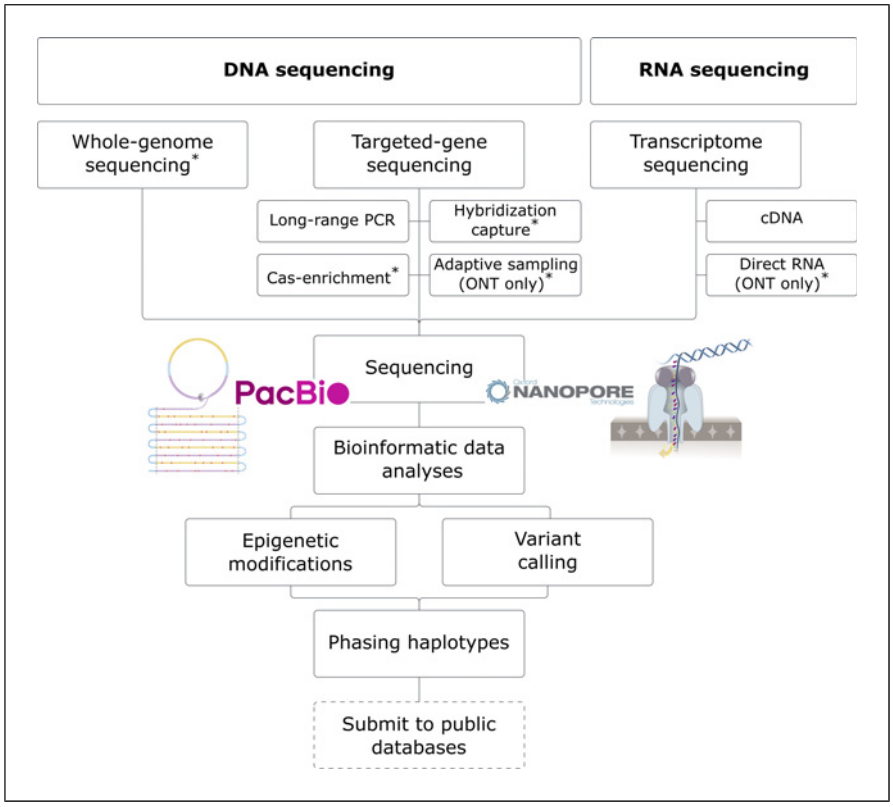

Fig. 2. Flowchart showing different sequencing strategies and workflows when working with TGS. Methods for which epigenetic information is not lost during library preparation are marked by asterisk (*). The "submit to public databases" step is encircled by dashes to illustrate that it is not imperative for most applications, nonetheless highly recommended and beneficial to the community.

Similarly, SVs can be problematic if PCR primers fall within such a region and – without noticing – only the wildtype allele is amplified [38].

Hybridization Capture

Hybridization capture approaches have been widely used in combination with short-read sequencing across fields and are currently adapted for TGS applications. The principle of hybridization capture relies on annealing and capturing of single-stranded DNA or RNA oligonucleotides – also named "probes" or "baits" – specifically designed for the regions of interest. Probes are usually biotinylated, which allows to pulldown targeted DNA fragments using streptavidin-coated magnetic beads. The main advantage of this approach is that several megabases of sequence can be targeted using thousands or even millions of probes in a single reaction [64]. Since no specific PCR primers are required, it is also less prone to allelic dropouts and generally less affected by template complexity than amplicon sequencing. A main challenge of hybridization capture approaches, however, is enrichment efficiency. On-target rates and coverage uniformity across targeted regions vary greatly, depending on factors such as DNA quantity and quality, or number of probes and their overall genome coverage.

Adapting hybridization capture approaches to TGS requires overcoming several challenges, mainly related to optimal fragment and probe sizes, as well as post-capture PCR amplification. Currently, only a few studies have adapted hybridization protocols to PacBio [65, 66] or ONT [67, 68] sequencing across research fields. In transfusion medicine, a first study was recently published by Steiert et al. [57], describing the development of a PacBio-compatible protocol for enriching genes relevant to 35 blood groups and two transcription factors using over 8,000 probes (Table 2). While more optimization is still required, this first assessment was promising [57].

Cas-Enrichment

Another novel promising strategy to enrich genes for TGS – which has not yet been applied to blood group genes – relies on the CRISPR/Cas9 molecular complex and is available for both ONT and PacBio platforms [69, 70]. It is a completely amplification-free method, allowing targeting multiple genomic loci as long fragments. Enrichment is based on the specific endonuclease activity of the Cas-protein (often Cas9, but also possible with other Cas-proteins depending on PAM motif [71]), cleaving genomic DNA at exact sites with respect to custom-designed CRISPR-RNAs (crRNAs). In a single reaction, it is possible to pool up to 100 crRNAs [70, 72, 73]. Only fragments cut by the Cas-protein can serve as templates during TGS, which allows enriched targeted sequencing. Since the approach is PCR free, even epigenetic modifications are maintained and can be analyzed downstream. Challenges include low on-target rates, partly

due to incomplete cutting efficiency [70, 74], and coverage heterogeneity when long fragments are targeted [75]. Finally, protocols are laborious to establish and per sample costs are relatively high.

Adaptive Sampling with ONT

Another promising amplification-free approach for targeted-gene sequencing is the emerging so-called "adaptive sampling" method offered by ONT only. During adaptive sampling, regions of interest are enriched by selectively ejecting reads from off-target regions in real time during actual sequencing. Strands passing nanopores are instantly mapped against reference sequences of the regions of interest. This can either happen directly from the electrical signal [76] or during the live-basecalling process [77]. In case the sequence does not map to a target region, the strand in the nanopore is ejected by reversing the ionic current, liberating the pore for a new stand. To make this approach efficient, reads need to be classified as on or off target within the first few hundred base pairs of sequencing. A key advantage of this method is that no previous laborious and costly DNA manipulation is required. As downside, constantly reversing the ionic current appears to prematurely obstruct pores, hence reduces overall sequencing output and thus depth of coverage.

Bioinformatic Analyses of Sequencing Data

We focus here on the workflow in a variant calling pipeline as this application is most frequently needed to analyze blood group genes. We first outline general requirements for bioinformatic data analysis before describing the workflow in more detail.

General Requirements

PacBio as well as ONT offers basic standard pipelines provided as graphical interfaces in their analysis software platforms *SMRT Link* and *MinKNOW*, respectively. While *SMRT Link* for PacBio's sequel IIe requires additional computing capacity of a local network, ONT's GridION and PromethION devices do not need additional computing power and run *MinKNOW* completely on instrument. ONT's pocket-sized MinION sequencer can in principle be operated using a laptop; however, it best runs on GPU workstations in particular when computationally demanding basecalling models are used. Both PacBio and ONT also offer cloud-based analysis options, though data privacy issues may need to be considered in medical fields.

More sophisticated data analysis with flexible and optimized data processing requires building modified pipelines, for which familiarity with the command-line environment is needed. Crucial in this regard is also the ability to evaluate as well as integrate novel and updated analysis tools into pipelines. Bioinformatic data analysis also requires critical revision and, where appropriate, manual curation of results of intermediate steps. Such tasks require skilled knowledge in both bioinformatics and molecular genetics. Advanced bioinformatic knowledge with profound programming skills, however, is only needed for the development of own applications.

Workflow

Current instruments from PacBio carry out the circular consensus sequencing workflow on the instrument itself and directly output HiFi reads in BAM format (unaligned), including kinetic information that can be used to call base modifications (see *Perspectives*). ONT uses a specific FAST5 file format to store the electrical signals and translates them into nucleotide sequences (FASTQ file format) during the basecalling process. After de-multiplexing (if samples were pooled), trimming of adapter/barcode sequences, and filtering out low quality reads, sequencing reads are mapped with long-read aligners to either the human reference genome or the specific reference sequences of the respective blood group genes. Alternatively, sequencing reads may be assembled de novo, i.e., reference free, for each sample. In such an iterative workflow, an assembled and polished sample-specific consensus sequence serves as individual reference sequence [29]. This approach can be highly beneficial to overcome allelic biases from standard reference-based mapping in case of high genetic difference between the reference sequence and the sequenced allele [37, 78]. In blood group systems with high genetic diversity, reference sequences may occasionally strongly deviate from the sequenced allele, for example, for intronic regions of *ABO*O2* or *RHCE*C* alleles [29, 79], or in case of hybrid alleles of the RHD/CE and MNS systems. After mapping sequencing reads, specific tools are used to call, filter, and annotate genetic variants (stored in VCF file format). If possible, genetic variants are subsequently phased, allowing to build final haplotype sequences (FASTA file format). Large SVs are often called in a separate pipeline as bioinformatic tools differ and evolve rapidly [80, 81]. Most tools are command-line based and not platform specific, i.e., can be used interchangeably between PacBio and ONT data.

Excursus: Machine Learning in Long-Read Sequencing

Advances in bioinformatic analyses of long-read sequencing data are narrowly intertwined with steep developments in deep learning approaches, in which neural network architectures compute complex features merely by taking the results of preceding operations as input [82]. For instance, the constant increase of ONT's read accuracy is strongly related to improvements of neural networks to

decode the raw electrical signal into a nucleotide sequence. While segmented electrical signal data were transformed using "hidden Markov models" at the beginning, "recurrent" and later "convolutional neural networks" working directly with the raw electrical signal took over [21]. Deep learning models are also central in variant calling pipelines [20]. Unlike for short reads, where statistical methods are still prevalent, variant calling algorithms for long reads are based on deep learning approaches [83, 84]. The success of deep learning methods is promoted by the growing availability of training data to refine model features.

Perspectives

Long-Read Sequencing to Decipher the Blood Group "Regulome"

Long-read sequencing offers great potential for significant advancements in systematic investigation of the cellular regulation of blood group genes, which is still in its early infancy. Both PacBio and ONT provide comprehensive and elegant solutions for investigating epigenomes and transcriptomes across cell types. We focus here on sequencing DNA base modifications and long RNA molecules as these are areas in which the two TGS technologies are forecasted to particularly excel short-read sequencing.

Epigenetics

Epigenetic modifications contribute to phenotypes by regulating gene expression. The most widespread technique to detect cytosine methylation (5mC) across the entire genome is whole-genome bisulfite sequencing [85], which was also adapted to TGS [86]. Specific PacBio and ONT workflows, however, detect a much wider variety of epigenetic DNA modifications without prior chemical treatment. With PacBio, epigenetic modifications are detected by the pulse width and interpulse duration from the fluorescent base signals representing the characteristic time duration in nucleotide incorporation [87]. Machine learning methods were introduced to increase sensitivity and specificity [14]. With ONT, all types of modified bases are detected by their unique electrical signal profiles, which is a hallmark of ONT. Basecalling algorithms allow direct identification of DNA base modifications and accuracy for the most relevant 5mC methylation is already as high as for canonical basecalling (99.5%, Table 1). Future perspectives to measure all base modifications in one run look promising once training data have reached sufficient breadth and depth.

RNA Sequencing

Sequencing of the transcriptome is another area, which has been dominated by short-read sequencing for a long time. However, there are major drawbacks when sequencing long mRNA molecules with short reads. First, mRNAs are reverse transcribed into cDNA, and the need to fragment cDNA renders the computational assembly of transcripts error prone, in particular since transcript processing is a very dynamic process [88, 89]. Second, epigenetic information is lost because of the reverse transcription.

As key advantage, both TGS methodologies enable the identification of full-length isoforms via cDNA sequencing [90, 91]. However, due to the PCR amplification step, transcript quantification remains biased and base modifications are lost. Since ONT does not sequence by synthesizing, native RNA molecules guided through the nanopore can also be sequenced (Table 1) [92] and even distinguished by their base modifications. This direct RNA sequencing (dRNA-seq) opens up the field of epitranscriptomics [93]. Current limitations include the output (RNA nucleotides pass the pore five times slower than those from DNA), the high amount of input material (500 ng polyA RNA), and the still mediocre read accuracy (~90%).

In transfusion medicine, a very recent publication used ONT for cDNA sequencing of *RHD* transcripts in a DEL phenotype finding remarkable splicing variety (Table 2) [60]. In fact, RNA sequencing has the potential to serve as sensitive alternative to the adsorption-elution technique to differentiate null from weak blood group antigen expression. However, a significant challenge in utilizing cDNA sequencing or dRNA-seq for this application is that the expression of transcripts for certain blood group genes is very tissue specific. Furthermore, the transcriptional and translational dynamics may be temporally restricted to different stages of erythrocyte progenitor cells in the bone marrow and peripheral blood [94]. A comprehensive elucidation of the transcriptome and epigenome in progenitor cells as well as erythrocytes is hence paramount to understand the regulation of blood group genes [95].

Continued Technological Evolution

Long-read sequencing is a rapidly evolving area that is expected to see significant progress in the coming years. Technologically, it is forecasted that the accuracy of ONT will soon reach the same level as PacBio. As presented in the latest Nanopore Community Meeting [22], sequencing the complimentary strand in the same pore ("duplex," Table 1) has already reached Q30 (99.9%). Moreover, ONT is currently addressing the issue of homopolymer accuracy by developing longer pores with multiple sensing regions and increasing the stability of transit speed. Additionally, the use of voltage reversion, as implemented in adaptive sampling, has the potential to further increase accuracy by allowing sequencing a molecule several times back and forth, approaching the concept of PacBio's circular consensus sequencing. In dRNA-seq, a newly designed nanopore and motor protein as well as more diverse training sets for the neural network is about to significantly

improve read accuracy and lower input requirements. PacBio, on the other hand, may further increase its insert size (read length) by optimizing polymerase durability and throughput as demonstrated in the recently presented Revio system [19]. Illumina, the most popular short-read sequencing provider, also works on synthetic long reads [96].

It is foreseen that long-read sequencing will be widely adopted in diagnostic laboratories to replace, for instance, alternative technologies for the assessment of samples with discordant or complex serology. With the role model of rapid solutions for HLA characterization, we can anticipate that software for predicting blood group phenotypes will be expanded [97, 98] or newly developed to automatically process TGS data.

Curation and Terminology of Blood Group Allele Haplotypes

With the implementation of short- and long-read sequencing technologies to determine blood group alleles, the detection of novel alleles occurs at steadily increasing pace. This is posing new challenges for the curation and terminology of blood group alleles. In addition, blood group alleles can now be resolved as full-length haplotype sequences. Anticipating that this is an emerging standard, nomenclature guidelines will need to be updated, including adjusting the definition of a unique blood group allele. A collection of alleles solely based on exonic and splice-site variants no longer adequately represents the fast progress of available haplotype-resolved sequences for whole genes. Finally, commonly depositing haplotype sequences, ideally along with ethnicity frequency information and serology, in a dedicated database would be highly valuable.

Conclusions

Long-read sequencing has great potential to address some of the key limitations associated with short-read sequencing. Thanks to recent technological advancements, TGS is becoming a vital tool for medical diagnostic applications. In the field of transfusion medicine, it opens up new avenues to tackle relevant questions in the most complex blood group systems RHD/CE and MNS. It also allows characterizing blood group alleles as complete haplotype sequence, an upcoming standard. Additionally, TGS holds great promise for exploring new domains, such as the transcriptomic and epigenetic analysis of blood group systems.

Statement of Ethics

The authors have no ethical conflicts to disclose.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

No funding was received in the preparation of the manuscript.

Author Contributions

Gian Andri Thun, Morgan Gueuning, and Maja P. Mattle-Greminger contributed ideas and wrote the manuscript.

References

1 Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cel. 2015 May;58(4):586–97.
2 Orzińska A. Next generation sequencing and blood group genotyping: a narrative review. Ann Blood. 2023;8:4.
3 Fürst D, Tsamadou C, Neuchel C, Schrezenmeier H, Mytilineos J, Weinstock C. Next-generation sequencing technologies in blood group typing. Transfus Med Hemother. 2020; 47(1):4–13.
4 Fichou Y, Le Maréchal C, Férec C. Next-generation sequencing for blood group genotyping. ISBT Sci Ser. 2017 Feb;12(1): 184–90.
5 Lane WJ. Recent advances in blood group genotyping. Ann Blood. 2021 Sep;6(5):31.
6 Orzińska A, Guz K, Mikula M, Kulecka M, Kluska A, Balabas A, et al. A preliminary evaluation of next-generation sequencing as a screening tool for targeted genotyping of erythrocyte and platelet antigens in blood donors. Blood Transfus. 2018 May;16(3):285–92.
7 Jakobsen MA, Dellgren C, Sheppard C, Yazer M, Sprogøe U. The use of next-generation sequencing for the determination of rare blood group genotypes. Transfus Med. 2019 Jun;29(3):162–8.
8 Roulis E, Schoeman E, Hobbs M, Jones G, Burton M, Pahn G, et al. Targeted exome sequencing designed for blood group, platelet, and neutrophil antigen investigations: proof-of-principle study for a customized single-test system. Transfusion. 2020 Sep;60(9):2108–20.
9 Lang K, Wagner I, Schöne B, Schöfl G, Birkner K, Hofmann JA, et al. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. BMC Genomics. 2016;17(1):374.
10 Schoeman EM, Lopez GH, McGowan EC, Millard GM, O'Brien H, Roulis EV, et al. Evaluation of targeted exome sequencing for 28 protein-based blood group systems, including the homologous gene systems, for blood group genotyping. Transfusion. 2017 Apr;57(4):1078–88.
11 Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018 Apr;36(4): 338–45.
12 Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019 Oct; 37(10):1155–62.
13 Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.
14 Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, et al. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. Proc Natl Acad Sci. 2021 Feb;118(5):e2019768118.
15 Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278–89.

16 Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17(1):239.
17 Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012;13(1):375.
18 Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol. 2016 May;34(5):518–24.
19 Pacific Biosciences. pacb. [cited 2023 Mar 24]. Available from: https://www.pacb.com/.
20 Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. Cell Genom. 2022 May;2(5):100129.
21 Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018 Dec;19(1):90.
22 Oxford Nanopore Technologies. Nanoporetech. [cited 2023 Mar 24]. Available from: https://nanoporetech.com/.
23 International Society of Blood Transfusion. Red cell Immunogenet blood Gr Terminol. Amsterdam. [cited 2023 Mar 24]. Available from: https://www.isbtweb.org/isbt-working-parties/rcibgt.html.
24 Gassner C, Castilho L, Chen Q, Clausen FB, Denomme GA, Flegel WA, et al. International society of blood transfusion working party on red cell immunogenetics and blood group terminology report of basel and three virtual business meetings: update on blood group systems. Vox Sang. 2022 Nov;117(11):1332–44.
25 Karamatic Crew V, Tilley LA, Satchwell TJ, AlSubhi SA, Jones B, Spring FA, et al. Missense mutations in PIEZO1, which encodes the Piezo1 mechanosensor protein, define Er red blood cell antigens. Blood. 2023 Jan;141(2):135–46.
26 Srivastava K, Lee E, Owens E, Rujirojindakul P, Flegel WA. Full-length nucleotide sequence of ERMAP alleles encoding Scianna (SC) antigens. Transfusion. 2016 Dec;56(12):3047–54.
27 Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585(7823):79–84.
28 Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022 Apr;376(6588):44–53.
29 Gueuning M, Thun GA, Wittig M, Galati A-L, Meyer S, Trost N, et al. Haplotype sequence collection of ABO blood group alleles by long-read sequencing reveals putative A1 -diagnostic variants. Blood Adv. 2023 Mar;7(6):878–92.
30 Fichou Y, Berlivet I, Richard G, Tournamille C, Castilho L, Férec C. Defining blood group gene reference alleles by long-read sequencing: proof of concept in the ACKR1 gene encoding the Duffy antigens. Transfus Med Hemother. 2020;47(1):23–32.
31 Srivastava K, Khil PP, Sippert E, Volkova E, Dekker JP, Rios M, et al. ACKR1 alleles at 5.6 kb in a well-characterized renewable US Food and Drug Administration (FDA) reference panel for standardization of blood group genotyping. J Mol Diagn. 2020 Oct;22(10):1272–9.
32 Hanks SC, Forer L, Schönherr S, LeFaive J, Martins T, Welch R, et al. Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing. Am J Hum Genet. 2022 Sep;109(9):1653–66.
33 Guo Y, Busch MP, Seielstad M, Endres-Dighe S, Westhoff CM, Keating B, et al. Development and evaluation of a transfusion medicine genome wide genotyping array. Transfusion. 2019;59(1):101–11.
34 Gleadall NS, Veldhuisen B, Gollub J, Butterworth AS, Ord J, Penkett CJ, et al. Development and validation of a universal blood donor genotyping platform: a multinational prospective study. Blood Adv. 2020;4(15):3495–506.
35 Shestak AG, Bukaeva AA, Saber S, Zaklyazminskaya EV. Allelic dropout is a common phenomenon that reduces the diagnostic yield of PCR-based sequencing of targeted gene panels. Front Genet. 2021 Feb;12:620337.
36 Lam C, Mak CM. Allele dropout caused by a non-primer-site SNV affecting PCR amplification: a call for next-generation primer design algorithm. Clin Chim Acta. 2013 Jun;421:208–12.
37 Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017 May;27(5):665–76.
38 Gueuning M, Thun GA, Schneider L, Trost N, Sigurdardottir S, Engström C, et al. Resolving genotype-phenotype discrepancies of the Kidd blood group using nanopore sequencing. bioRxiv.
39 Thun GA, Gueuning M, Sigurdardottir S, Meyer E, Gourri E, Schneider L, et al. Novel regulatory variant in ABO intronic RUNX1 binding site inducing A3 phenotype. bioRxiv.
40 Miller DE, Lee L, Galey M, Kandhaya-Pillai R, Tischkowitz M, Amalnath D, et al. Targeted long-read sequencing identifies missing pathogenic variants in unsolved Werner syndrome cases. J Med Genet. 2022 May;59(11):1087–94.
41 Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, et al. Genomic answers for children: dynamic analyses of >1000 pediatric rare disease genomes. Genet Med. 2022 Jun;24(6):1336–48.
42 Watson CM, Dean P, Camm N, Bates J, Carr IM, Gardiner CA, et al. Long-read nanopore sequencing resolves a TMEM231 gene conversion event causing Meckel–Gruber syndrome. Hum Mutat. 2020 Feb;41(2):525–31.
43 Le van Kim C, Mouro I, Cherif-Zahar B, Raynal V, Cherrier C, Cartron JP, et al. Molecular cloning and primary structure of the human blood group RhD polypeptide. Proc Natl Acad Sci U S A. 1992;89(22):10925–9.
44 Willemetz A, Nataf J, Thonier V, Peyrard T, Arnaud L. Gene conversion events between GYPB and GYPE abolish expression of the S and s blood group antigens. Vox Sang. 2015 May;108(4):410–6.
45 Storry JR, Lammers C, Olsson ML, Robb JS. A novel GYP(B-A-B) hybrid in a blood donor identified by a phenotyping discrepancy with different anti-s reagents. Transfusion. 2023 Jan;63(1):3–5.
46 Kudo S, Fukuda M. Structural organization of glycophorin A and B genes: glycophorin B gene evolved by homologous recombination at Alu repeat sequences. Proc Natl Acad Sci U S A. 1989 Jun;86(12):4619–23.
47 Lane WJ, Gleadall NS, Aeschlimann J, Vege S, Sanchis-Juan A, Stephens J, et al. Multiple GYPB gene deletions associated with the U−phenotype in those of African ancestry. Transfusion. 2020 Jun;60(6):1294–307.
48 Chou ST, Flanagan JM, Vege S, Luban NLC, Brown RC, Ware RE, et al. Whole-exome sequencing for RH genotyping and alloimmunization risk in children with sickle cell anemia. Blood Adv. 2017 Aug;1(18):1414–22.
49 Wheeler MM, Lannert KW, Huston H, Fletcher SN, Harris S, Teramura G, et al. Genomic characterization of the RH locus detects complex and novel structural variation in multi-ethnic cohorts. Genet Med. 2019 Feb;21(2):477–86.
50 Stef M, Fennell K, Apraiz I, Arteta D, González C, Nogués N, et al. *RH* genotyping by non-specific quantitative next-generation sequencing. Transfusion. 2020 Nov;60(11):2691–701.
51 Halls JBL, Vege S, Simmons DP, Aeschlimann J, Bujiriri B, Mah HH, et al. Overcoming the challenges of interpreting complex and uncommon RH alleles from whole genomes. Vox Sang. 2020 Nov;115(8):790–801.
52 Tounsi WA, Madgett TE, Avent ND. Complete RHD next-generation sequencing: establishment of reference RHD alleles. Blood Adv. 2018;2(20):2713–23.
53 Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019 Dec;20(1):246.
54 Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet. 2021 Jun;53(6):779–86.
55 Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020 May;581(7809):444–51.
56 Montemayor C, Simone A, Long J, Montemayor O, Delvadia B, Rivera R, et al. An open-source Python library for detection of known and novel Kell, Duffy and Kidd variants from exome sequencing. Vox Sang. 2021 Apr;116(4):451–63.
57 Steiert TA, Fuß J, Juzenas S, Wittig M, Hoeppner MP, Vollstedt M, et al. High-throughput method for the hybridisation-based targeted enrichment of long genomic fragments for PacBio third-generation sequencing. NAR Genom Bioinform. 2022 Jul;4(3):lqac051.
58 Tounsi WA, Lenis VP, Tammi SM, Sainio S, Haimila K, Avent ND, et al. Rh blood group D antigen genotyping using a portable nanopore-based sequencing device: proof of principle. Clin Chem. 2022;68(9):1196–201.
59 Zhang Z, An HH, Vege S, Hu T, Zhang S, Mosbruger T, et al. Accurate long-read sequencing allows assembly of the duplicated RHD and RHCE genes harboring variants relevant to blood transfusion. Am J Hum Genet. 2022 Jan;109(1):180–91.

60 Ji Y, Luo Y, Wen J, Sun Y, Jia S, Ou C, et al. Patients with Asian-type DEL can safely be transfused using RhD-positive blood. Blood. 2023 Apr;141(17):2141–50.
61 Horn R, Merchant J; UK-FR GENE Consortium. Managing expectations, rights, and duties in large-scale genomics initiatives: a European comparison. Eur J Hum Genet. 2022;31(2):142–7.
62 Thorpe R, Jensen K, Masser B, Raivola V, Kakkos A, von Wielligh K, et al. Donor and non-donor perspectives on receiving information from routine genomic testing of donor blood. Transfusion. 2022 Dec;63(2):331–8.
63 Daniels G. Human blood groups. Oxford, UK: Wiley; 2013.
64 Kozarewa I, Armisen J, Gardner AF, Slatko BE, Hendrickson CL. Overview of target enrichment strategies. Curr Protoc Mol Biol. 2015 Oct;112(1):7.21.1–23.
65 Giolai M, Paajanen P, Verweij W, Percival-Alwyn L, Baker D, Witek K, et al. Targeted capture and sequencing of gene-sized DNA molecules. Biotechniques. 2016 Dec;61(6):315–22.
66 Wang M, Beck CR, English AC, Meng Q, Buhay C, Han Y, et al. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. BMC Genomics. 2015 Dec;16(1):214.
67 Karamitros T, Magiorkinis G. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. Nucleic Acids Res. 2015 Dec;43(22):e152.
68 Eckert SE, Chan JZM, Houniet D, Breuer J, Breuer J, Speight G. Enrichment by hybridisation of long DNA fragments for nanopore sequencing. Microb Genom. 2016 Sep;2(9):e000087.
69 Höijer I, Tsai YC, Clark TA, Kotturi P, Dahl N, Stattin EL, et al. Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. Hum Mutat. 2018;39(9):1262–72.
70 Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020 Apr;38(4):433–8.
71 Grobler L, Suleman E, Thimiri Govinda Raj DB. Patents and technology transfer in CRISPR technology. In: David T, editors. Progress in molecular biology and translational science. 1st ed. Amsterdam: Elsevier Inc.; 2021. p. 153–82.
72 Oxford Nanopore Technologies. Targeted, amplification-free DNA Seq using Crispr/Cas: Infosheet. [cited 2023 Mar 24]. Available from: https://community.nanoporetech.com/info_sheets/targeted-amplification-free-dna-sequencing-using-crispr-cas/v/ECI_S1014_v1_revE_11Dec2018.
73 Höijer I, Johansson J, Gudmundsson S, Chin C-S, Bunikis I, Häggqvist S, et al. Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity. Genome Biol. 2020 Dec;21(1):290.
74 Rubben K, Tilleman L, Deserranno K, Tytgat O, Deforce D, Van Nieuwerburgh F. Cas9 targeted nanopore sequencing with enhanced variant calling improves CYP2D6-CYP2D7 hybrid allele genotyping. PLOS Genet. 2022 Sep;18(9):e1010176.
75 Bruijnesteijn J, van der Wiel M, de Groot NG, Bontrop RE. Rapid characterization of complex killer cell immunoglobulin-like receptor (KIR) regions using Cas9 enrichment and nanopore sequencing. Front Immunol. 2021 Sep;12:722181.
76 Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with uncalled. Nat Biotechnol. 2021 Apr;39(4):431–41.
77 Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021 Apr;39(4):442–50.
78 Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLOS Genet. 2019 Jul;15(7):e1008302.
79 Carritt B, Kemp TJ, Poulter M. Evolution of the human RH (Rhesus) blood group genes: a 50-year-old prediction (partially) fulfilled. Hum Mol Genet. 1997 Jun;6(6):843–50.
80 Luan MW, Zhang XM, Zhu ZB, Chen Y, Xie SQ. Evaluating structural variation detection tools for long-read sequencing datasets in Saccharomyces cerevisiae. Front Genet. 2020 Mar;11:159.
81 Dierckxsens N, Li T, Vermeesch JR, Xie Z. A benchmark of structural variation detection by long reads through a realistic simulated model. Genome Biol. 2021 Dec;22(1):342.
82 Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389–403.
83 Luo R, Sedlazeck FJ, Lam T, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019 Mar;10(1):998.
84 Shafin K, Pesout T, Chang P, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021 Nov;18(11):1322–32.
85 Clark SJ, Statham A, Stirzaker C, Molloy PL, Frommer M. DNA methylation: bisulphite modification and analysis. Nat Protoc. 2006;1(5):2353–64.
86 Yang Y, Scott SA. DNA methylation profiling using long-read single molecule real-time bisulfite sequencing (SMRT-BS). Methods Mol Biol. 2017;1654:125–34.
87 Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010;7(6):461–5.
88 Davies JP, Jeffries AR, Castanho I, Jordan BT, Moore K, Leung SK, et al. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. Cell Rep. 2021;37(7):110022.
89 David JK, Maden SK, Wood MA, Thompson RF, Nellore A. Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads. Genome Biol. 2022;23(1):240.
90 Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31(11):1009–14.
91 Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res. 2017;6:100.
92 Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15(3):201–6.
93 Ramasamy S, Sahayasheela VJ, Yu Z, Hidaka T, Cai L, Sugiyama H, et al. Chemical probe-based nanopore sequencing to selectively assess the RNA modification. SSRN Electron J. 2021.
94 Heshusius S, Heideveld E, Burger P, Thiel-Valkhof M, Sellink E, Varga E, et al. Large-scale in vitro production of red blood cells from human peripheral blood mononuclear cells. Blood Adv. 2019;3(21):3337–50.
95 Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017 Apr;356(6335):eaah4573.
96 Illumina [Internet]. Long-Read Seq Technol. [cited 2023 Mar 24]. Available from: https://www.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html.
97 Lane WJ, Vege S, Mah HH, Lomas-Francis C, Aguad M, Smeland-Wagman R, et al. Automated typing of red blood cell and platelet antigens from whole exome sequences. Transfusion. 2019;59(10):3253–63.
98 Jadhao S, Davison CL, Roulis EV, Schoeman EM, Divate M, Haring M, et al. RBCeq: a robust and scalable algorithm for accurate genetic blood typing. EBioMedicine. 2022 Feb;76:103759.