Nucleic Acids Research, 2025, **53**, gkaf131
https://doi.org/10.1093/nar/gkaf131
Computational Biology

A detailed analysis of second and third-generation sequencing approaches for accurate length determination of short tandem repeats and homopolymers

Sophie I. Jeanjean[1,†], Yimin Shen[2,†], Lise M. Hardy[1], Antoine Daunay[1], Marc Delépine[3], Zuzana Gerber[3], Antonio Alberdi[4], Emmanuel Tubacher[2], Jean-François Deleuze[1,2,3], Alexandre How-Kit[1,*]

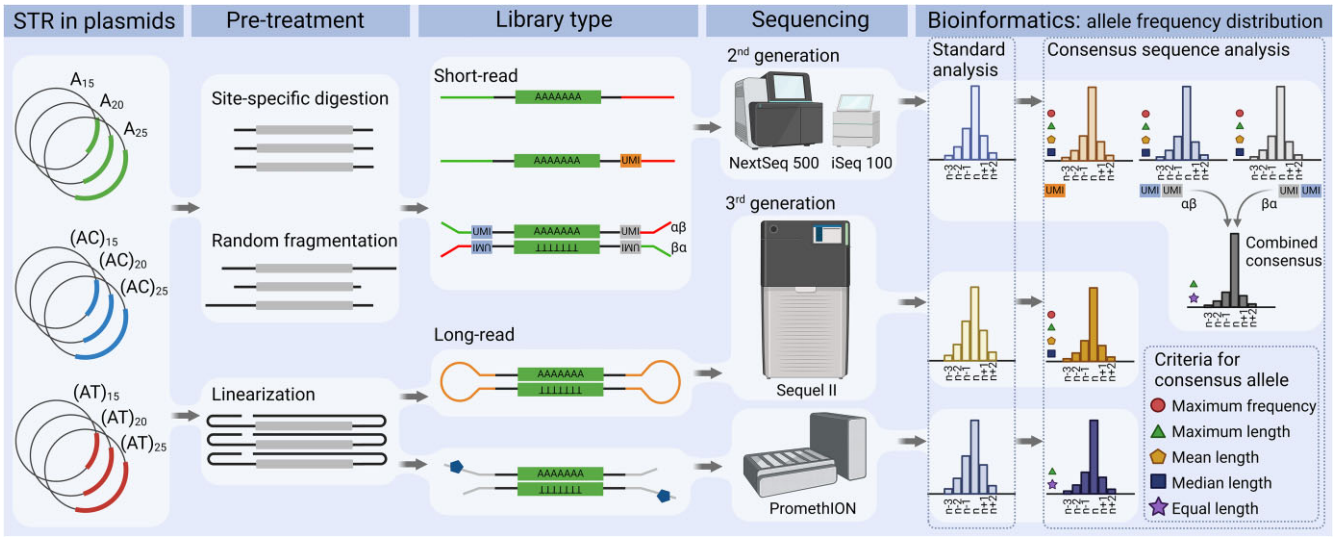
[1]Laboratory for Genomics, Foundation Jean Dausset – CEPH, 75010 Paris, France
[2]Laboratory for Bioinformatics, Foundation Jean Dausset – CEPH, 75010 Paris, France
[3]Centre National de Recherche en Génomique Humaine (CNRGH), CEA, Institut François Jacob, 91000 Evry, France
[4]Technological Platform of Saint-Louis Research Institute (IRSL), Saint-Louis Hospital, University of Paris, 75010 Paris, France

[*]To whom correspondence should be addressed. Tel: +33 1 53725146; Email: alexandre.how-kit@fjd-ceph.org
[†]Equal contribution

Abstract

Microsatellites are short tandem repeats (STRs) of a motif of 1–6 nucleotides that are ubiquitous in almost all genomes and widely used in many biomedical applications. However, despite the development of next-generation sequencing (NGS) over the past two decades with new technologies coming to the market, accurately sequencing and genotyping STRs, particularly homopolymers, remain very challenging today due to several technical limitations. This leads in many cases to erroneous allele calls and difficulty in correctly identifying the genuine allele distribution in a sample. Here, we assessed several second and third-generation sequencing approaches in their capability to correctly determine the length of microsatellites using plasmids containing A/T homopolymers, AC/TG or AT/TA dinucleotide STRs of variable length. Standard polymerase chain reaction (PCR)-free and PCR-containing, single Unique Molecular Indentifier (UMI) and dual UMI 'duplex sequencing' protocols were evaluated using Illumina short-read sequencing, and two PCR-free protocols using PacBio and Oxford Nanopore Technologies long-read sequencing. Several bioinformatics algorithms were developed to correctly identify microsatellite alleles from sequencing data, including four and two modes for generating standard and combined consensus alleles, respectively. We provided a detailed analysis and comparison of these approaches and made several recommendations for the accurate determination of microsatellite allele length.

Graphical abstract



Received: July 12, 2024. **Revised:** January 13, 2025. **Editorial Decision:** February 8, 2025. **Accepted:** February 11, 2025

Introduction

Microsatellite DNA sequences are short tandem repeats (STRs) of a unit of 1 to 6 nucleotides found in almost all genomes of living organisms, also known as homopolymer when the repeated tract is a single nucleotide. They are highly polymorphic due to their high mutation rates, which are primarily caused by polymerase slippage during DNA replication, where the nascent strand denatures from the template strand and re-hybridizes forward or backward on the same strand, resulting in the deletion or insertion of one or more repeat units, respectively [1]. These mutations are corrected *in vivo* by the mismatch repair system (MMR), which can reduce mutation rates by tens to thousands of times [2, 3], but is deficient in many types of cancers known as MMR-deficient (MMRD) or microsatellite instability (MSI) cancers [4, 5]. The rates of microsatellite mutations have been shown to be influenced by several intrinsic factors, among which the decreasing number of nucleotides in the repeated unit and the increasing number of repeats act positively [3, 6–10]. This has also been described *in vitro*, notably during the polymerase chain reaction (PCR) amplification, where microsatellite mutations, known as stutter peaks/bands, are also introduced by polymerase slippage [1, 11, 12]. These artifactual frameshift products can sometimes mask true microsatellite alleles and mutations present in a DNA sample, interfering with many downstream applications.

Noteworthy, microsatellites have been widely used as genetic markers in several biomedical applications due to their high mutation rates and polymorphism, including gene mapping [13], evolutionary genetics [9, 14], conservation and population genetics [15, 16], marker-assisted selection [17], kinship analysis [18], forensic DNA fingerprinting for identification of individuals and parentage [19, 20], and diagnosis of repeat-expansion diseases [21, 22] and MSI cancers [23]. While most of these applications only required microsatellite genotypes, this could sometimes be difficult to obtain due to the limitations of the genotyping technology used. This latter mostly relies on PCR-based approaches that are combined either with capillary electrophoresis fragment analysis/first-generation sequencing, or more recently with next-generation sequencing (NGS) that allowed a higher number of microsatellites analyzed [23, 24]. High background noise generated by stutter artifacts makes accurate microsatellite genotyping difficult, which required the development of computational algorithms correcting stutter errors, notably for NGS data [25–31]. Moreover, NGS data also contain other issues directly linked to the sequencing technology used, which must be considered when calling microsatellite alleles and genotypes. These include sequencing errors induced by STRs and especially homopolymers, for both second- and third-generation sequencing, as well as limitations in the length of the microsatellites analyzed for short-read sequencing, leading to difficulties for correct STR and indel calling [26, 31–42].

In addition to the need for genotyping, other investigations and applications instead require accurate profiling of microsatellite allele length distributions derived from NGS rather than solely identifying microsatellite genotypes. Thus, NGS-based repeat length distributions were used in some fundamental studies to describe the landscape of microsatellite instability and decipher the underlying mechanism of their mutability in different tissue and MMR deficient/proficient contexts, including human cancers [5, 43–45]. Moreover, these length distributions could also be very useful for detecting MSI in many sample types from cancer patients [23, 46, 47], or for deconvolving alleles and genotypes from complex DNA mixtures in forensics [48]. However, the characterization of the true distribution of microsatellite alleles in a sample could also be very difficult due to the same issues described above, i.e. stutter artifacts and sequencing errors.

In the present study we evaluated and compared various NGS methods and instruments for accurate length estimation of different microsatellites in order to identify the best approach minimizing repeat length errors and that could reflect their genuine allele distribution present in a sample. Using several plasmids containing mono- (A/T) and di- (AC/TG and AT/TA)nucleotide microsatellites from 15 to 25 repeats, we generated several NGS libraries based on Illumina 2nd, PacBio and Oxford Nanopore Technologies (ONTs) third-generation sequencing technologies, including some that are specifically dedicated to sequencing errors reduction and rare mutation detection. Thus, with Illumina instruments (iSeq100 and/or NextSeq 500), we evaluated a PCR-free protocol [49], PCR-containing (1–20 cycles) protocols, and single- and dual-UMI (duplex sequencing) PCR-containing (12–20 cycles) protocols that allow PCR and sequencing error correction [50, 51], while we assessed two PCR-free approaches with PacBio Sequel II and ONT PromethION sequencers [38, 52]. Due to the specificities of microsatellites and homopolymers, several bioinformatics approaches have been developed for the analysis of microsatellite sequencing data. Our work provides an in-depth comparison of different NGS platforms and protocols, along with different bioinformatics strategies for accurately determining the allele length distribution of different microsatellite types. This may assist researchers in the selection of an NGS protocol and/or bioinformatics strategy that fits their experimental constraints and scientific objectives for the study of these sequences.

Material and methods

Microsatellite DNA templates

Synthetic $A_{15/20/25}$ mononucleotide, $(AC)_{15/20/25}$ and $(AT)_{15/20/25}$ dinucleotide repeat microsatellites were included in a DNA sequence (153 and 146 nt for mono- and di-nucleotide repeat microsatellites, respectively), including a common sequence of 136 nucleotides and a unique barcode of 4 nucleotides, allowing original sequence identification after NGS (Supplementary Fig. S1A). These synthetic DNA sequences were cloned into pUC57-mini Vector using the EcoRV restriction site and amplified using TOP10 *Escherichia coli* strain. All (sub-)cloning and plasmid preparations were performed by GenScript Biotech (the Netherlands) with industrial-grade standards and quality controlled by Sanger sequencing (Supplementary Fig. S1B).

Illumina library preparation and short-read sequencing

Libraries were prepared for Illumina short-read sequencing using QIAseq 1-Step Amplicon Library Kit (Qiagen), QIAseq FX DNA Library Kit (Qiagen), and xGen Prism DNA Library Prep Kit (Integrated DNA Technologies).

For QIAseq 1-Step Amplicon Library Kit (Qiagen) preparation, 5 μg of an equimolar pool of plasmids underwent site-specific digestion using three type II restriction enzymes (20 U

of HpyCH4V, 20 U of HpyCH4III and 100 U of MspI) in a 100 µl reaction including also 1X rCutSmart™ Buffer (New England Biolabs) for 1 h at 37°C followed by an A-tailing step at 37°C for 30 min by addition of 75 U of Klenow Fragment (3′→5′ exo-) (New England Biolabs) in the reaction mixture. After digestion, the microsatellites of interest were contained in the largest DNA fragments (361–389 pb) that were purified using 1.5X SPRIselect beads (BeckmanCoulter). 100 ng of purified DNA were then used for library preparation using either Combined Dual Index (CDI) adapters (Qiagen) or xGen™ UDI (Unique Dual Indexes)-UMI Unique Molecular Indentifiers (UMI) adapters (Integrated DNA Technologies) according to the manufacturer's instructions. 0 (PCR-free), 1, 2, 5, and 10 PCR cycles were used with libraries including CDI adapters, while 12, 16, and 20 PCR cycles were used with libraries including xGen™ UDI-UMI adapters. For xGen™ UDI-UMI libraries, approximately 20 000 DNA molecules per microsatellite type were used as templates for PCR.

For QIAseq FX DNA Library Kit (Qiagen) preparation, 1 µg of equimolar pools of plasmids was used for enzymatic shearing using the FX Enzyme Mix during 4 min of incubation at 32°C. The other steps were performed according to the manufacturer's instructions and xGen UDI-UMI adapters (Integrated DNA Technologies) were used for ligation. Pre-amplified libraries were sized (550–950 bp) using a 2% gel electrophoresis cassette on a BluePippin Size Selection System (Sage Science) according to the manufacturer's instructions. 0 (PCR-free), 12 and 16 PCR cycles were used for the amplification of the libraries using approximatively 20 000 DNA molecules per microsatellite type as templates for PCR.

For xGen™ Prism DNA Library Prep Kit (Integrated DNA Technologies) preparation, DNA samples were purified either from site-specific restriction digestion or random mechanical fragmentation. Site-specific digested DNA fragments were obtained as described for QIAseq 1-Step Amplicon Library Kit. For random fragmentation, 9 µg of equimolar pools of plasmids were sheared on a Bioruptor Sonication system (Diagenode) using an 8-cycle low position 30 s ON/90 s OFF program, followed by 2X SPRI bead purification and DNA sizing (300–600 bp) using a 2% gel electrophoresis cassette on a BluePippin Size Selection System (Sage Science) according to the manufacturer's instructions. All other steps of library preparation were performed according to manufacturer's instructions. Sixteen PCR cycles were used for library amplification using ∼20 000 DNA molecules per microsatellite type as templates for PCR using xGen™ UDI Primers (Integrated DNA Technologies). The final xGen Prism DNA libraries included two UDI for sample identification and two 8-nt UMI located on either side of the insert for duplex sequencing error correction.

All libraries were quantified using the Qubit HS dsDNA assay Kit (Life Technologies) on a Qubit 3 Fluorometer and the 2X KAPA Library Quantification Kit on a LightCycler 480 II (Roche), and verified for correct size using either the Agilent High Sensitivity DNA Kit on a 2100 Bioanalyzer Instrument (Agilent Technologies) or the NGS Fragment Kit (1–6000bp) on a Fragment Analyzer (Agilent Technologies), according to the manufacturer's instructions. Pooled libraries were mixed either with 0 (libraries from randomly sheared DNA) or 40–50% (libraries from site-specifically digested DNA) of PhiX Sequencing Control V3 (Illumina) and deposited in the sequencing cartridges according to the manufacturer's instructions. Sequencing reactions were performed either using iSeq100 i1 Reagent v2 (300 cycles) on an iSeq100 Instrument (Illumina) or NSQ 500/550 Mid Output KT v2.5 (150/300 cycles) reagents on a NextSeq 500 Instrument (Illumina). After sequencing, BCL raw and demultiplexed FASTQ data were generated and exported for bioinformatics analysis using Local Run Manager Software (Illumina).

Illumina short-read bioinformatics analyses

Standard microsatellite allele length analysis

In this study, each sample contained a mixture of microsatellites with very similar sequences (common sequences), except for the specific 4-nucleotide barcode (Supplementary Fig. S1). Their identification could be difficult even with a single mutation accepted in the barcodes. Classic mapping methods (bowtie, bwa) or sequence similarity analysis method (blast) were no longer usable. Thus, we developed a microsatellite identification method by recognizing the core part of the sequence on both sides of the microsatellites with specific constraints (Fig. 2A). This sequence contained four consecutive regions, including (i) the 4-nucleotide barcode region; (ii) 5 and 3 constant nucleotides before the microsatellite sequence for mono- and dinucleotide repeat microsatellites, respectively; (iii) the microsatellite with variable lengths; and (iv) a 5-nucleotide sequence following the microsatellite. The microsatellite identification strategy was based on these 4 regions using different criteria per region, i.e. 0, 1, 2, and 1 mutations allowed in the regions (i)–(iv), respectively. When there was uncertainty between a substitution and an insertion/deletion, the substitution was chosen. Following the identification of these four regions, the microsatellite length was calculated and the allele distribution computed for graphical representation. For the standard microsatellite allele length analysis, we processed the FASTQ files from the iSeq100 sequencing system, while the FASTQ files from the NextSeq 500 were extracted by demultiplexing the BCL files by merging the four lanes using bcl2fastq (v2.20).

Our strategy analyzing the four sequence regions has also been incorporated into other bioinformatics workflows, including short-read sequencing with single-UMI (Type II libraries) and dual-UMI error correction (Type III libraries), as well as long-read sequencing with PacBio (Type IV libraries) and ONT (Type V libraries).

PCR and sequencing error correction of repeat lengths using single-UMI

Single UMI allows to group reads originating from the same ssDNA molecule (Type II libraries). To analyze a microsatellite sequence, the four regions described above must be completely contained in the reads. The UMI sequence was not sufficient to identify all the different UMI groups by the classical UMI consensus approach due to multiple length errors in the reads and to the reduced diversity in the reference sequences. Therefore, we developed our own methods to minimize the risk of error in both single-UMI (Fig. 2B) and also dual-UMI data treatments (Fig. 2C).

Since single-UMI data were not included in the inserts (Fig. 1B), we extracted them from the BCL files. For each lane we determined the barcode and performed the demultiplexing including the UMI data using Picard (v2.8.2), and then merged the four lanes by Samtools (v1.14) (Fig. 2B). To reduce the error in UMI group identification, we demultiplexed the sample

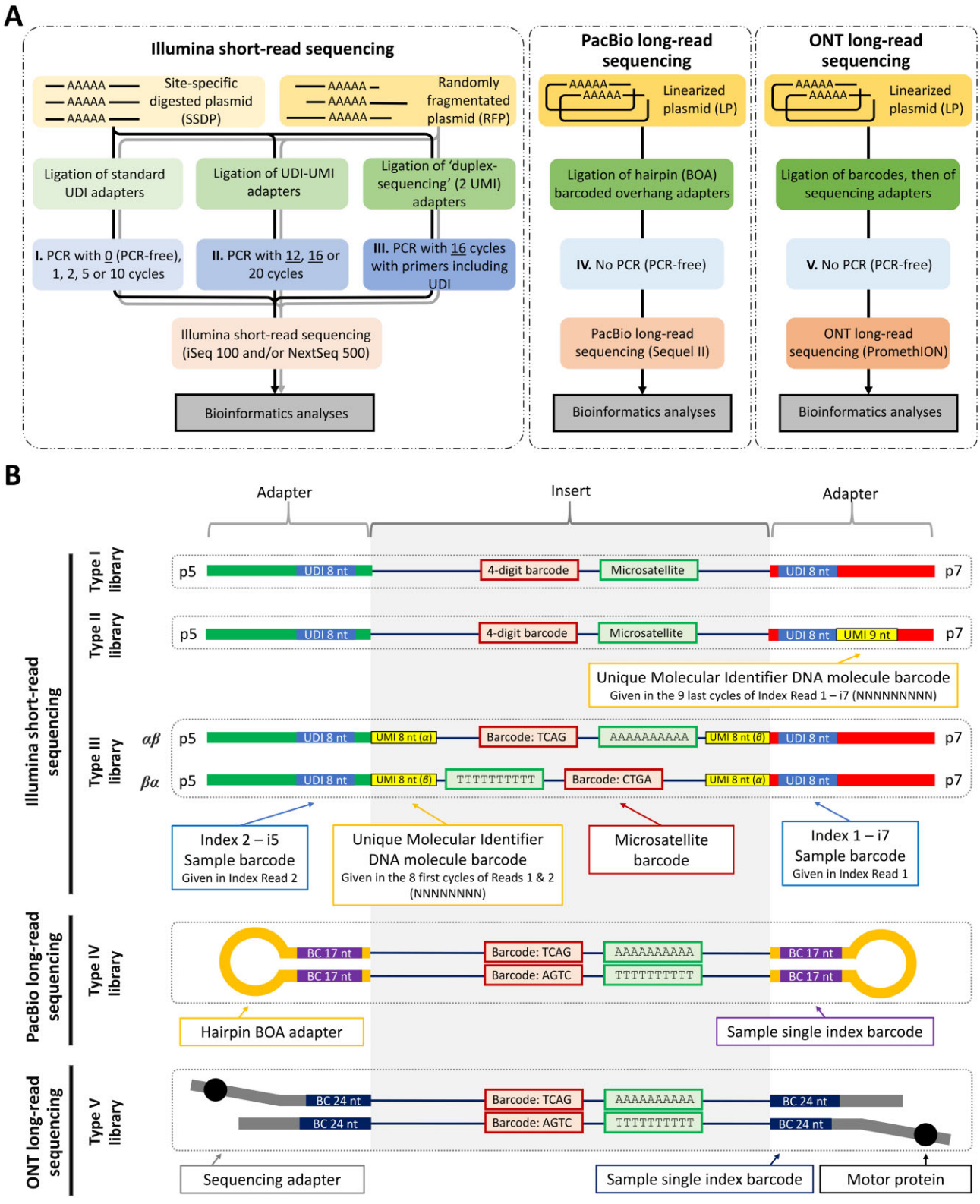


Figure 1. Experimental workflow of NGS experiments performed in our study. (**A**) Detail of the different steps performed in our study for short-read and long-read NGS library preparation using microsatellite-containing plasmids as DNA samples. Underlined PCR conditions correspond to those applied to randomly sheared DNA. (**B**) Schematic representation of the different molecular components of the five types of NGS libraries used in our study. Illumina short-read sequencing libraries are shown single stranded. When considering the double-stranded configuration, type I and II libraries included "Y" structures at each DNA end for PCR-free libraries due to stubby Y-adapters, while being fully complementary in PCR-containing libraries. For "duplex-sequencing" type III libraries, two strands of the same dsDNA molecule were barcoded at each DNA extremity allowing their identification—each strand would present either $\alpha\beta$ or $\beta\alpha$ barcodes combination in reads 1 and 2—and error correction after PCR and sequencing.

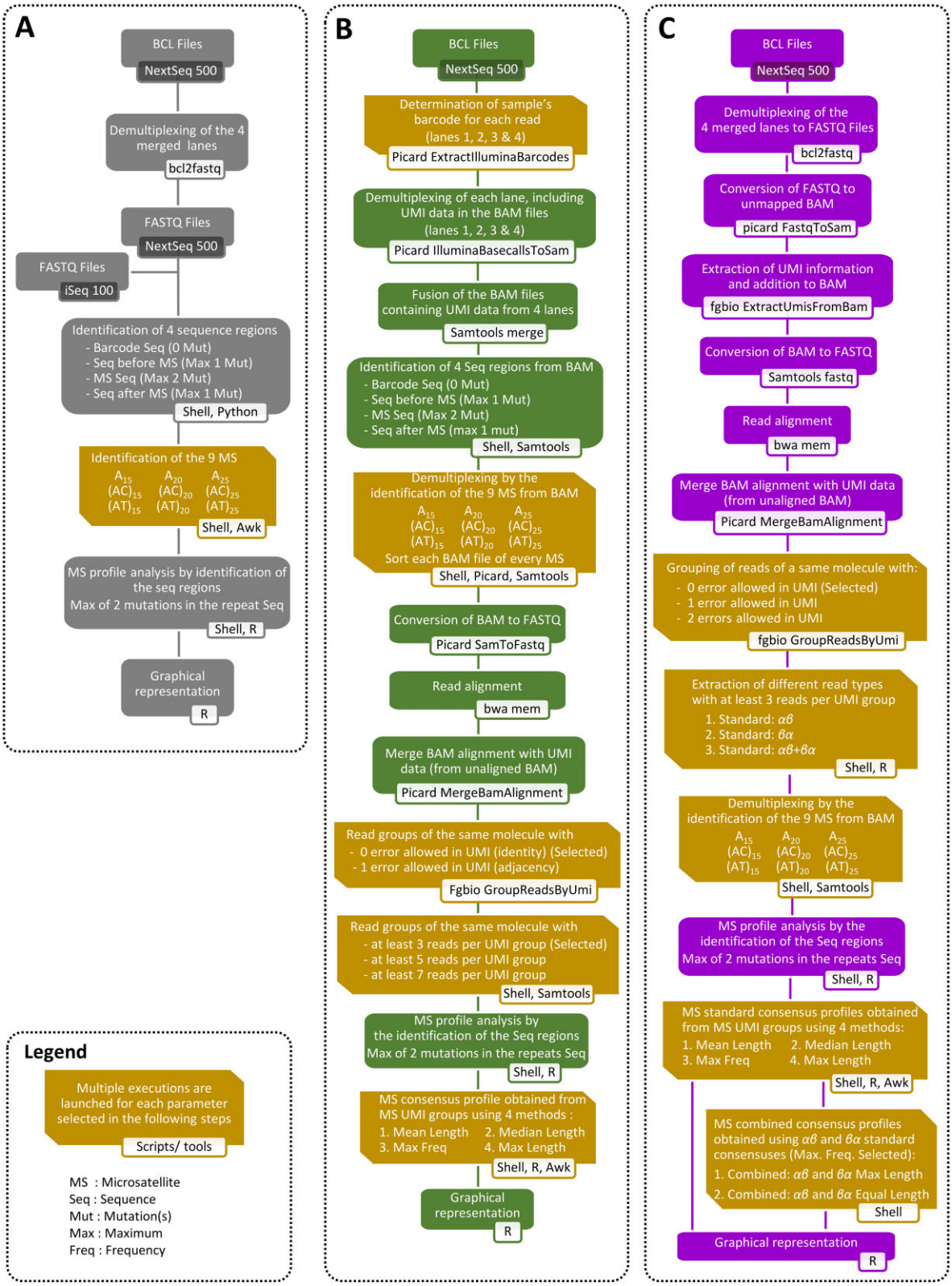


Figure 2. Workflow of the different bioinformatic analyses performed using Illumina short-read sequencing data. (**A**) Standard microsatellite allele length analysis workflow used with type I, II, and III libraries. (**B**) Single-UMI error correction workflow used with type II libraries. (**C**) Duplex sequencing standard and combined error correction workflow used with type III libraries. "(Selected)" indicates the analyses, features or options, whose results are presented in the manuscript.

by microsatellite type before mapping using the microsatellite identification method described above. We then converted the binary alignment/map (BAM) files of each microsatellite to FASTQ and performed the alignment using bwa (v0.7.17). After adding the UMI information in the BAM file mapped by Picard (v2.8.2), we identified the UMI group by fgbio (v2.0.2) with no mutation allowed to reduce errors in the identification of UMI groups. For each group of UMIs containing at least three reads, we identified all the microsatellite lengths and applied the four methods ("Max. Freq.", "Max. Length", "Mean Length", and "Median Length") to obtain a consensus length. Finally, we analyzed the final microsatellite profile based on the distribution of consensus allele lengths.

PCR and sequencing error correction of repeat lengths using dual-UMI

Dual-UMI reads (Type III libraries) contain 2 UMI sequences (2×8 nt) in both insert ends (Fig. 1B), resulting in more possible UMI combinations, which reduces the errors in identifying UMI groups. In our workflow (Fig. 2C), we first extracted the FASTQ files by demultiplexing the BCL files and merging the lanes by bcl2fastq (v2.20). Then FASTQ files were converted into unmapped BAM files by Picard (v2.8.2) and the information from the UMI was extracted by fgbio (v2.0.2). We then converted the BAM files to FASTQ and performed alignment using bwa (v0.7.17). After the addition of the UMI information in the BAM files mapped by Picard (v2.8.2), we identified the UMI group by fgbio (v2.0.2) by accepting no mutation. We analyzed three types of reads (standard $\alpha\beta$, $\beta\alpha$, and $\alpha\beta$+$\beta\alpha$) containing at least three reads per UMI group and obtained the consensus microsatellite length of each UMI group using the four methods ("Max. Freq.", "Max. Length", "Mean Length", and "Median Length"). We further implemented two additional methods by comparing the consensus ("Max. Freq.") obtained from both $\alpha\beta$ and $\beta\alpha$ reads for each UMI group. These combined error correction methods either considered the longest microsatellite length between $\alpha\beta$ and $\beta\alpha$ consensuses, or selected a length only if $\alpha\beta$ and $\beta\alpha$ consensuses were identical. The final microsatellite profiles based on all of these consensus allele lengths were used for graphical representation (Fig. 2C).

PacBio sequel II library preparation and long-read sequencing

2 μg of each plasmid was linearized (except for A_{15} plasmid where the unique 4-nt barcode creates an additional PciI restriction site) at 37°C for 1 hour using 10 units of PciI restriction enzyme and 1X of NEBbuffer™ r3.1 (New England Biolabs), followed by a heat inactivation step at 80°C for 20 min. The linearized plasmids (LP) were purified using the QIAquick PCR purification kit (Qiagen) and quantified by the Qubit HS dsDNA assay Kit (Life Technologies) according to the manufacturer's instructions. Plasmids were equimolarly pooled by four, and concentrated by ethanol precipitation with sodium acetate.

Libraries were prepared using barcoded overhang adapters (BOA) for multiplexing amplicons (PacificBiosciences) and 200 ng of each DNA pools using SMRTbell prep kit 3.0 (PacificBiosciences) according to manufacturer's instructions. Briefly, PacBio Sequel II library preparation included a DNA damage repair step (30 min at 37°C), an end repair and A-tailing addition combined step (30 min at 20°c and 30 min at 65°C), BOA ligation to fully-repaired linear dsDNA (1 h at 20°C) followed by an enzymatic heat inactivation step (10 min at 65°C). Ligated libraries were purified using 0.6X AMPure PacBio® beads, quantified using the Qubit HS dsDNA assay Kit (Life Technologies) according to the manufacturer's instruction and stored at -20°C until sequencing. The size of the libraries was confirmed using Agilent dsDNA 930 Reagent Kit (75–20 000 bp) on a 5300 Fragment Analyzer system (Agilent Technologies) according to the manufacturer's instructions.

For PacBio Sequel II long-read sequencing reaction, v4 sequencing primers were annealed and DNA polymerase was bound to the multiplexed libraries using SMRTbell binding kit 2.0 (Pacific Biosciences), before purification of the DNA-protein complexes using 0.6X AMPure PacBio beads. Multiplexed amplicon libraries were loaded onto one SMRT cell and sequenced on a PacBio Sequel II instrument using C2 chemistry and a 30-h movie time. After sequencing reaction, sequencing data were demultiplexed by SMRT Link (v9.0.0.92188) and raw subreads as well as circular consensus sequencing (CCS) reads were exported for bioinformatics analyses.

PacBio long-read bioinformatics analyses

Two types of analyses were performed using the CCS BAM data and the raw subread BAM data (Fig. 3 A). The file sizes of these two types of sources were very different: compared to the CCS files, which were ∼400 Mb, the raw subread file data were a thousand times larger (∼400 Gb).

Standard analysis of repeat length distribution

To evaluate the profile performance according to the quality of the CCS reads, we demultiplexed the CCS BAM file according to different Quality Scores: Q_{20}, Q_{30}, Q_{40}, Q_{50}, Q_{60}, Q_{70} and Q_{80} using SMRTLink (v11.0.0.146107). Similarly, to short-read sequencing, our sample contains a mixture of microsatellites close in sequences. Thus, for the same reason, we used the microsatellite identification method by the recognition of the four consecutive regions: (i–iv). We extracted the unmapped CCS BAM sequences after identifying the nine microsatellites and then analyzing their allele length profiles.

In-house subread analysis for repeat length distribution

We also analyzed raw subread data using the BAM subread files. Demultiplexing was performed according to the Barcode Quality Score using SMRTtools (v12.0.0.177059) lima. Since subreads do not have Quality Scores, we used the Barcode Score to pre-assess the subread quality. We tested several Barcode Quality Scores: no filter, Q_{26}, Q_{50}, Q_{60}, Q_{70}, and Q_{80}. Since each demultiplexed file was very large, we split each file into 13 fragment files keeping all subreads from the same group in the same file. We then randomly selected different numbers of subreads per group: 3, 5, 7, 11, 15, 21, 25, 35, and 45. We identified all the microsatellites using the microsatellite identification approach based on the four consecutive regions: (i–iv). Finally, for each subread group, the consensus length was obtained using the four previously described methods ("Max. Freq.", "Max. Length", "Mean Length", and "Median Length") and we analyzed the allele distribution profiles.

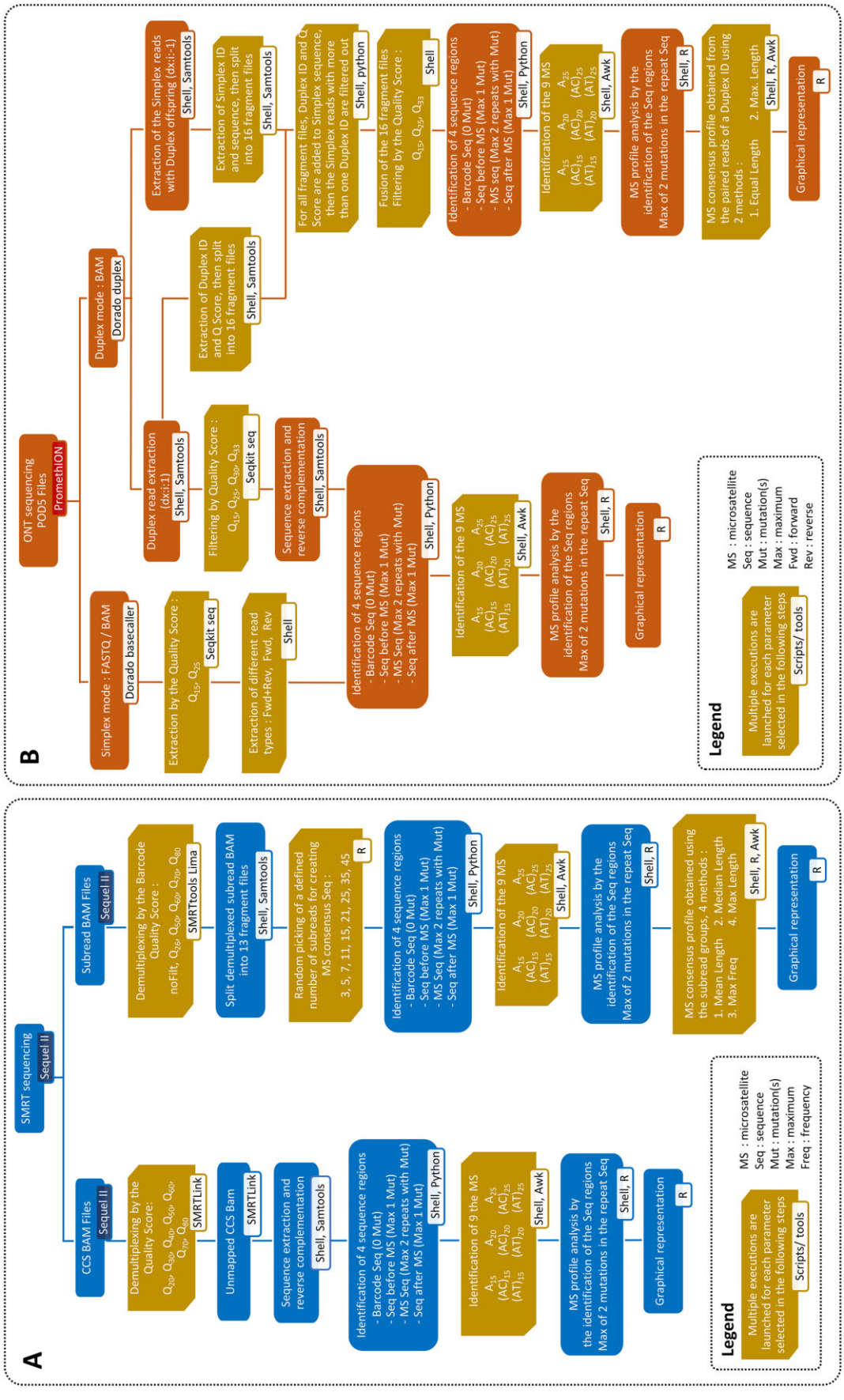


Figure 3. Workflow of the different bioinformatic analyses performed using (**A**) PacBio (type IV libraries) and (**B**) ONT (type V libraries) long-read sequencing data.

ONT PromethION library preparation and long-read sequencing

Around 5 μg of each plasmid was linearized with 100 units of ScaI-HF restriction enzyme and 1X of rCutSmart™ Buffer (New England Biolabs) at 37°C for 1 h, followed by DNA purification with 0.5X SPRI beads (Beckman Coulter), and quantified using the Qubit HS dsDNA assay Kit (Life Technologies) according to the manufacturer's instructions. Plasmids were equimolarly pooled by three (A_{15}/ $(AC)_{20}$/ $(AT)_{25}$, A_{20}/ $(AC)_{25}$/ $(AT)_{15}$ and A_{25}/ $(AC)_{15}$/ $(AT)_{20}$) before library preparation. Libraries were prepared in triplicates from 250 ng of each plasmid pool using Native Barcoding Kit 96 V14 (ONT) together with NEBNext® Ultra II End Repair / dA-tailing Module, NEB Blunt/TA Ligase Master Mix, and NEBNext® Quick Ligation Module (New England Biolabs), according to manufacturer's instructions. Briefly, ONT library preparation included end repair step (5 min at 20°C and 5 min at 65°C), ligation of barcodes (20 min at room temperature), pooling of barcoded samples in equal volumes, purification using 0.4X AMPure XP Beads, ligation of sequencing adapters (20 min at room temperature), and purification using 0.4X AMPure XP Beads with Short Fragment Buffer. Final ONT libraries were quantified using the Qubit HS dsDNA assay Kit (Thermo Fisher) according to the manufacturer's instructions and equimolarly pooled, loaded together with a WGS ONT library (Promega human genomic DNA) at a 1:1 ratio on an R10.4.1 flow cell (ONT) and sequenced on PromethION 24 sequencing device (ONT) for 36 hours, with basecalling SUP (model dna_r10.4.1_e8.2_400bps_5khz_sup.cfg) and automatic demultiplexing in MinKNOW (version 5.8.6).

ONT long-read bioinformatics analyses

Dorado (v0.8.2) was used with the nanopore signal (POD5) files for simplex and duplex basecalling to obtain simplex and duplex nucleotide sequences (FASTQ/BAM) respectively. Three different methods were then developed to analyze the repeat length distribution, including a standard simplex method, a standard duplex method and an in-house duplex-based consensus method (Fig. 3B).

Standard simplex analysis

ONT simplex FASTQ files generated from 'Dorado basecaller' were filtered for minimum read quality (Q_{15} and Q_{25}) using Seqkit (v2.8.2). Then, microsatellite profiles were analyzed using three different read types: forward, reverse, and all (forward + reverse). For the analysis of allele lengths, we used the same microsatellite identification method described above based on the recognition of the four consecutive regions: (i–iv).

Standard duplex analysis

Duplex reads were extracted with the dx tag (dx:i:1) using the ONT duplex BAM files obtained from 'dorado duplex'. We used Seqkit (v2.8.2) for minimum read quality filtering using different thresholds (Q_{15}, Q_{25}, Q_{30} and Q_{33}). The allele length distribution profiles of all reads were then analyzed using the same microsatellite identification method based on the recognition of the 4 consecutive regions: (i–iv).

In-house duplex-based consensus analysis

Duplex reads were first extracted with the dx tag (dx:i:1) using the ONT duplex BAM files obtained from "dorado duplex." Then, simplex reads with duplex offspring were extracted with the dx tag (dx:i:-1) and tagged with a duplex ID and a duplex quality score derived from duplex reads. We filtered out simplex reads with more than one duplex ID and then extracted the simplex reads by the minimum duplex read quality score (Q_{15}, Q_{25}, and Q_{33}). All microsatellites were identified using the microsatellite identification approach based on the four consecutive regions: (i–iv). Finally, for each pair of simplex reads based on one duplex read, the consensus length was calculated using the "Max. Length" and "Equal Length" methods, enabling allele length distribution profile analysis.

Results

Description of the NGS approaches used in our comparative study

The present study aimed to evaluate and compare the accuracy of second and third-generation sequencing approaches for identifying the length of microsatellites, including mono- (A/T) and di-nucleotide (AC/TG and AT/TA) microsatellites with 15–25 repetitions that were inserted into plasmids, each containing a specific four-digit barcode (Supplementary Fig. S1A). Each plasmid preparation was assumed to be pure, containing only the sequence with the desired number of repetitions, as confirmed by Sanger sequencing (Supplementary Fig. S1B). We focused on mono- and di-nucleotide STRs, the most common types of microsatellite sequences in the human genome, and selected a repeat number of at least 15, as the accuracy of microsatellite length determination from NGS data has been shown to become increasingly difficult with an increasing number of repetitions [25, 26, 37, 38, 53]. STRs with more than 25 repetitions were not evaluated in the study to limit the influence of short-read sequencing constraints on bioinformatics analyses, and to allow comparison with long-read sequencing. For short-read sequencing approaches, microsatellites excised from plasmids by type II restriction enzyme digestion (SSDP) or randomly sheared plasmids (RFP) were used as DNA samples (Fig. 1A). These DNA were used for library preparations using 0, 1 or 2 UMIs-containing adapters and including either no PCR (PCR-free) and 1–20 cycles of PCR (Fig. 1A), resulting in three types (I, II, and III) of libraries (Fig. 1B). For long-read sequencing, two types of PCR-free libraries (IV and V) were prepared from linearized plasmids (≈2 kb), which were ligated directly to standard BOA circularization and ONT sequencing adapters prior to PacBio and ONT sequencing, respectively (Fig. 1). Subsequently, several bioinformatics approaches were applied to analyze short-read and long-read sequencing data, allowing the identification of microsatellite length (see methods section and Figs 2 and 3).

Impact of PCR cycles in Illumina short-read library preparation on the estimation of microsatellite allele length

PCR is a convenient and widespread DNA amplification technique included at various stages in the preparation of Illumina short-read library, whether for targeted gene (TGS), whole exome (WES), or whole genome sequencing (WGS), while PCR-free protocols are less versatile due to a higher

amount of DNA requirement and are usually restricted to WES and WGS [49, 54]. To evaluate the impact of PCR amplification on the accuracy of microsatellite length determination, we prepared libraries using SSDP and applied 0–20 cycles of PCR amplification before sequencing (Fig. 1A). This type of DNA insert enabled microsatellite sequencing to begin within the same sequencing cycle—34 and 36 cycles for each di- and mononucleotide repeat microsatellite, respectively—which should avoid biases due to sequencing position. Sequencing error rates were calculated from PCR-free data and revealed a drastic increase in indel rates in the microsatellite sequence and in substitution rates after the microsatellite sequence, which is more pronounced for longer repeats (Supplementary Fig. 2). These high substitution error rates in post-STR sequences led us to consider the length of microsatellite alleles using our developed approach (see Methods) rather than the exact sequence obtained from the reads for all downstream analyses.

For each microsatellite type, our results showed the highest proportion of genuine alleles in PCR-free libraries (90% on average, with $(AT)_{20}$ and $(AT)_{25}$ microsatellites presenting the lowest percentage of original alleles, i.e. 89% and 52%, respectively), which decreased with the increasing number of PCR cycles (Fig. 4, Supplementary Fig. S3). A lower proportion of the original allele was observed for microsatellites with a higher number of repetitions in all library types, while the effect of PCR on the reduction of the original allele affected more A/T and AT/TA than AC/TG microsatellites (Fig. 4). The stutter background was increasingly biased toward deletions, as the number of PCR cycles increased (Fig. 4, Supplementary Fig. S3), which was expected. Of note, the stutter background introduced by polymerase slippage was visible from the first PCR cycle with an average decrease of 5% (min = 1.2% with $(AC)_{15}$ and max = 8.9% with $(AT)_{25}$) of the original allele compared to the PCR-free protocol (Fig. 4). When considering iSeq100 and NextSeq 500 instruments, the results were comparable with both technologies, except for $(AC)_{25}$ and $(AT)_{25}$ that showed higher percentage of the original alleles with iSeq100 and NextSeq 500, respectively (Supplementary Fig. S4). Finally, the comparison of PCR-free library data from SSDP and RFP showed worse results for the latter, notably with the longest microsatellites (Supplementary Fig. S5). This was probably due, for a large proportion of the reads, to the increase in the sequencing cycle from which the microsatellites are sequenced, thereby reducing their quality. Taken together, these results indicated that PCR-free protocols outperformed PCR-containing protocols when aiming to preserve the original microsatellite allele frequency distribution from a sample, and thereby allowed measurement of microsatellite length with the highest accuracy.

Accuracy of microsatellite allele length determination by UMI error correction from single-UMI libraries

Unique molecular identifiers (UMI) can be used from single- or dual-UMI protocols to tag DNA templates prior to library amplification and NGS (Fig. 1). The identification of multiple reads bearing the same UMI (i.e. sequencing duplicates) thereby allows for reaching a consensus sequence that reduces PCR and sequencing errors [50, 55, 56]. This strategy could potentially improve microsatellite length identification beyond that obtained with a PCR-free protocol.

We first evaluated this strategy for correct length estimation of microsatellites using a unique nine-nucleotide UMI located in the adapter sequence (Type II library) that barcoded single-stranded DNA molecules (Fig. 1). For the identification of read groups sharing the same UMI, no mutation in the UMI sequence was allowed. This was done to reduce the risk of aggregating multiple reads from more than one original DNA molecule, especially since the specificity of the original DNA molecule identification could not be improved by sequencing read coordinates for SSDP-based data. We considered at least three reads per UMI group for UMI error correction, and developed four modes of microsatellite length calculation, named "Max. Freq.," "Max Length," "Mean Length," and "Median Length" (Supplementary Fig. S6A). Among them, the "Max. Freq." mode was the most performant at maintaining the length of the original allele under most conditions, and was therefore selected for further detailed analyses (Supplementary Fig. S6B).

Our results showed that UMI error correction increased the original allele fraction for each type of repeat sequence analyzed in both SSDP and RFP conditions, although not to the levels observed with PCR-free libraries (Fig. 5A and Supplementary Fig. S7). The reduction in length error rates tended to increase with the number of cycles, and it was generally modest (no more than 2-fold) for microsatellites with at least 20 repeats (Fig. 5B). However, an almost 8-fold reduction in length error was achieved for those with 15 repeats (Fig. 5B), suggesting that UMI error correction is more efficient for smaller STRs. We also attempted to improve the accuracy of allele identification by increasing the minimum number of reads per UMI group from 3 to 7. Slight improvements in length estimation were observed by increasing this threshold for 15- and 20-repeat microsatellites, but not for those with 25 repeats (Supplementary Fig. S8). Thus, although the results obtained for 15-repeat microsatellites were close, UMI error correction did not present higher levels of original alleles compared to PCR-free data under our experimental conditions (Fig. 5A), indicating that the latter approach is better than the former for preserving and identifying the original length of microsatellites with 15–25 repetitions.

Improvement of microsatellite allele length estimation through combined error correction from duplex sequencing data

We next assessed the ability of duplex sequencing to improve microsatellite length determination. Duplex sequencing is based on dual UMI barcoding of each strand of a double-stranded DNA molecule (Type III library, Fig. 1), which enables their identification after NGS. It allows the construction of a combined consensus sequence from two read families ($\alpha\beta$ and $\beta\alpha$), which can reduce PCR and sequencing errors by an additional 10 fold compared to single UMI methods, especially when considering point mutations [50].

In a first step, standard UMI error correction was performed using the same parameters and stringency as for Type II libraries, i.e. no mutations allowed in UMI barcodes and at least three reads per UMI group, followed by a second combined error correction step using $\alpha\beta$ and $\beta\alpha$ consensuses. Consensus microsatellite lengths generated from standard UMI error correction were based on the four modes described previously (Supplementary Fig. S6A), by considering either $\alpha\beta$, $\beta\alpha$ or $\alpha\beta$+$\beta\alpha$ reads (for $\alpha\beta$+$\beta\alpha$, a consensus length was deter-

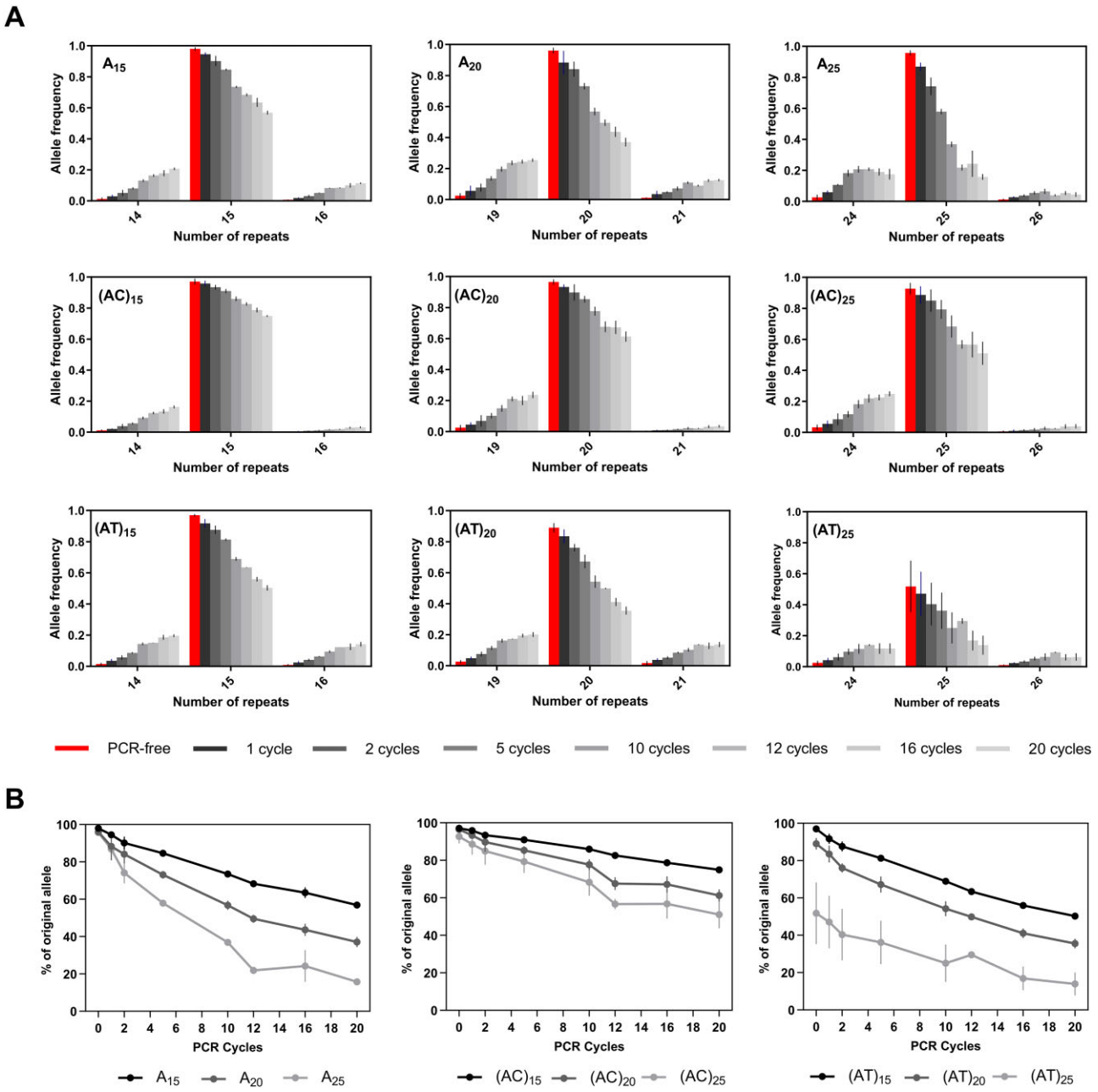


Figure 4. Effect of PCR cycles included in library preparation on microsatellite allele length obtained after Illumina short-read sequencing using SSDP. (**A**). $n-1$, n, and $n+1$ microsatellite allele frequencies obtained with PCR-free and PCR-containing (1–20 cycles) libraries and Illumina short-read sequencing using SSDP as templates. (**B**). Evolution of the percentage of original alleles of the nine studied microsatellites according to the number of PCR cycles. The figure only presents type I and II library data and each point includes at least five replicates. Data were generated on Illumina iSeq100 and NextSeq 500 instruments.

mined from at least three reads sharing the same $\alpha\beta$ or $\beta\alpha$ barcodes). The "Max. Freq." mode was again selected as it presented a higher percentage of the original alleles in most conditions (Supplementary Fig. S6C). Similar to Type II library, uncorrected Type III library data presented a lower percentage of original alleles compared to PCR-free experiments, due to stutter artifacts introduced during PCR that were biased toward deletions (Fig. 6A). After standard UMI error correction, the proportions of original alleles increased for all microsatellites with comparable values between $\alpha\beta$, $\beta\alpha$, and $\alpha\beta$+$\beta\alpha$ consensuses, indicating no major differences in error correction according to read family type (Fig. 6A and Supplementary Fig. S9). In terms of error rate reduction, standard UMI error correction was stronger for small repeats than for large repeats, and in SSDP condition than in RFP condition (Fig. 6B). However, the percentage of original alleles after error correction could not reach levels observed from PCR-free experiments (Fig. 6A).

The performance of duplex sequencing has further been evaluated through combined error correction. Two modes have been tested, considering either the maximum length or the equal length between $\alpha\beta$ and $\beta\alpha$ consensus alleles of the same read family. Thus, the largest allele is retained as a combined consensus with the "Max. Length" mode, while a com-

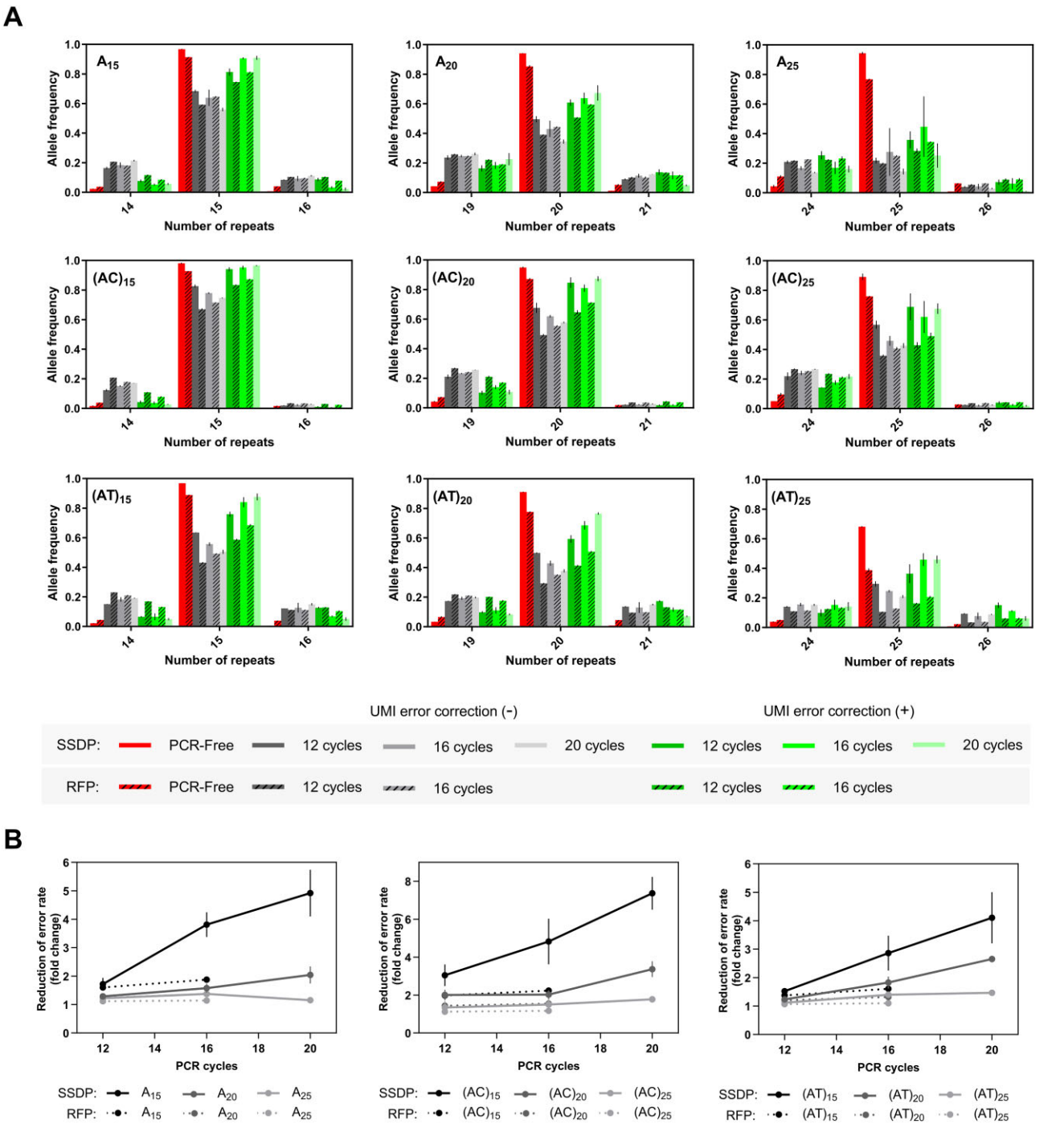


Figure 5. Impact of UMI error correction on Illumina short-read sequencing data for accurate length determination of microsatellites. (**A**) $n-1$, n, and $n+1$ microsatellite allele frequencies obtained before (-) and after (+) UMI error correction (Max. Freq. mode) from type II libraries including 12, 16, and 20 PCR cycles and using SSDP and RFP as templates. (**B**) Reduction of the error rate in the length of microsatellite alleles after UMI error correction, expressed in fold-change. Type I PCR-free data from Illumina short-read sequencing were also presented in panel A for comparison. Each point included at least triplicate experimental data, except for PCR-free – SSDP and $(AC)_{25}$ – 12 cycles – SSDP – UMI error correction conditions (duplicates). All data were generated on an Illumina NextSeq 500.

bined consensus is obtained only when both alleles are of the same length with the "Equal Length" mode. For some microsatellites, combined consensuses could not be generated due to the reduced number of $\alpha\beta$ and $\beta\alpha$ consensus alleles and the inability to assign them to the same family, more in RFP than in SSDP experiments and especially for microsatellites with increasing repeat number (Supplementary Fig. S10). Furthermore, the "Equal Length" approach failed more often than the "Max. Length" approach to achieve a combined consensus, as several read families exhibited discordant $\alpha\beta$ and $\beta\alpha$ consensuses. After combined error correction, the percentage of original alleles strongly increased for each condition where this correction was possible, more so in the "Equal Length" mode than in the "Max. Length" mode,

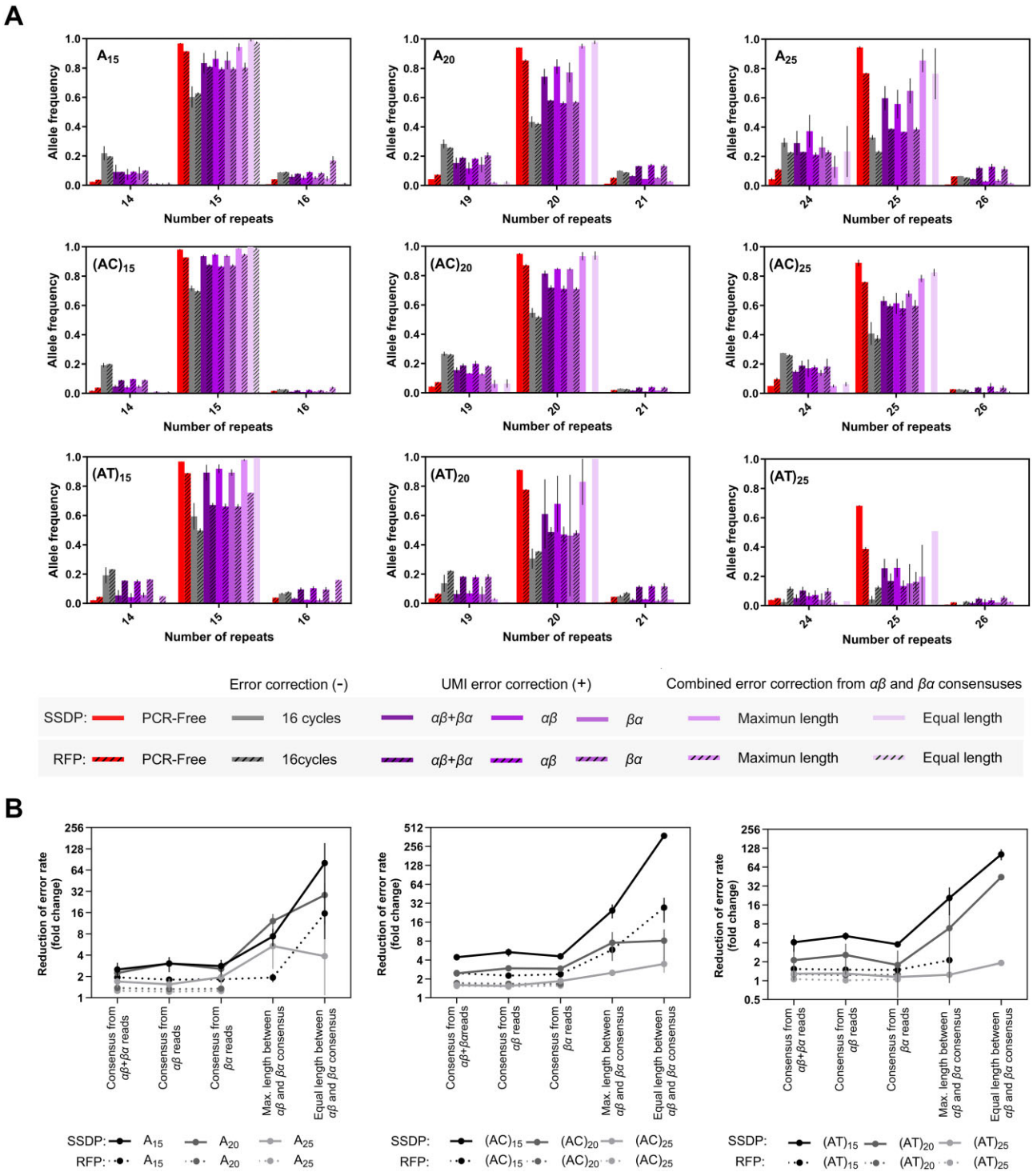


Figure 6. Impact of duplex sequencing dual-UMI error correction on Illumina short-read sequencing data for accurate length determination of microsatellites. (**A**). $n-1$, n, and $n+1$ microsatellite allele frequencies obtained before (-) and after (+) different types of UMI-based error corrections from type III libraries including 16 PCR cycles and using SSDP and RFP as templates. Error corrections included standard UMI error correction (Max. Freq. mode) obtained either from $\alpha\beta$ reads, $\beta\alpha$ reads or both read types ($\alpha\beta$+$\beta\alpha$), and combined error correction obtained from $\alpha\beta$ and $\beta\alpha$ consensus sequences, either based on the maximum (the largest allele is kept as consensus) or equal length (the consensus is reached when the $\alpha\beta$ and $\beta\alpha$ consensus alleles are of the same length). (**B**). Reduction of the error rate in the length of microsatellite alleles after error corrections, expressed in fold-change. Type I PCR-free data from Illumina short-read sequencing were presented in panel A for comparison. Each point originated from duplicate (SSDP) or triplicate (RFP) experiment data, however, for combined error correction, some data points were partially or totally (no bars in the original n alleles) lost. All data were generated on an Illumina NextSeq 500.

and even reached 100% in one SSDP replicate for $(AC)_{15}$ among 1496 combined consensus sequences (Fig. 6A and Supplementary Fig. S9). This increase outperformed standard UMI error correction, and sometimes also PCR-free results for small microsatellites (A_{15}, A_{20} and $(AC)_{15}$ in both SSDP and RFP conditions and $(AT)_{15}$ in SSDP condition), reducing the allele length error rate from less than a dozen to several hundred times depending on the repeat type and number. Thus, duplex sequencing could increase the accuracy of microsatellite length characterization, but it required a large amount of reads to build combined consensus sequences.

Accuracy and improvement of microsatellite allele length identification by PacBio long-read sequencing

Third-generation sequencing is defined by its ability to generate very long reads compared to second-generation sequencing. Single-molecule real-time (SMRT) sequencing developed by Pacific Biosciences is one of the most pioneering and widely used third-generation sequencing technologies based on zero-mode waveguide (ZMW) and single molecule sequencing by synthesis (SBS) [57, 58]. To evaluate the technology's ability to correctly estimate repeat length, template plasmids were linearized to construct libraries without any PCR step (Type IV library) and then sequenced directly on a SMRT cell 8M using a Sequel II. Raw read analysis indicated that each plasmid had more than 1000 ZMWs with at least 10 exploitable subreads with a Barcode Score of 70, which could even increase up to more than 100 subreads depending on the repeat type (Fig. 7A). This could be reflected in the quality of CCS reads that exhibited an important proportion of the maximum Quality Score, i.e. Q_{80} with SMRT Link analysis (Fig. 7B). There were no variations in substitution error rates before and after the microsatellite sequence contrary to short-read sequencing, as the sense of sequencing reverses after each subread with PacBio long-read sequencing (Supplementary Fig. 11). However, high-indel error rates were observed throughout the sequence with this sequencing chemistry, as expected, and also a slight increase in substitution error rate only in AT repeats.

To compare the accuracy of microsatellite length determined by PacBio sequencing, we first used different thresholds of Quality Scores of CCS HiFi reads ranging from Q_{20} to Q_{80}, i.e. a probability of incorrect base call ranging from 10^{-2} to 10^{-8} in each CCS read. Compared to Illumina PCR-free, PacBio CCS read results always presented lower percentages of original alleles, even for the highest quality reads, which decreased with an increasing number of repeats (Fig. 7C). Incorrect allele calls were biased toward insertions for A/T homopolymers and deletions for dinucleotide STRs (Fig. 7C, Supplementary Fig. S12). Increasing the CCS read Quality Score threshold from Q_{20} to Q_{80} did not increase the percentage of original alleles for dinucleotide microsatellites, but slightly increased it for homopolymers. However, this slight difference could possibly be caused by the high amount of Q_{80} CCS reads in our run (Fig. 7B).

We next attempted to improve allele length identification from long-read sequencing by constructing consensus sequences based on the four types of in-house approaches described previously (Supplementary Fig. S6A), and using different numbers of subreads ranging from 3 to 35. "Max Freq.," "Mean Length," and "Median Length" exhibited the best results and all three were presented in Fig. 7C. By increasing the number of subreads, the proportion of original alleles increased with the three modes, as expected. However, these approaches did not exhibit better results than with CCS reads for homopolymers. In contrast, for dinucleotide microsatellites, the three approaches yielded better results than CCS reads above a threshold of subreads per consensus sequence. In the case of AC/TG dinucleotides, we could even increase the original allele fractions above those derived from Illumina short-read PCR-free data, with the highest fraction being 0.994 for $(AC)_{15}$ with "Mean Length" using 35 subreads per consensus sequence. Of note, incorrect allele calls with the three in-house approaches were biased toward insertions for dinucleotide STRs and balanced for homopolymers, differing from CCS reads.

Assessment of ONT sequencing for microsatellite allele length determination

ONT sequencing was evaluated as another third-generation sequencing technology based on single-molecule sequencing, which uses electric signals emitted by a DNA strand passing through a nanopore [59]. To facilitate bioinformatic analysis, only three plasmids were pooled per library and the libraries were sequenced on a PromethION instrument, together with a WGS library prepared from a human Promega DNA. There was no observed variation in error substitution error rates before and after the microsatellite sequence regardless of the forward and reverse orientation of the read, unlike short-read sequencing and similar to PacBio sequencing (Supplementary Fig. S13).

Simplex reads were first analyzed either by separating forward from reverse reads or considering them together, and by using two minimum Quality Score thresholds (Q_{15} and Q_{25}) that gave a sufficient number of reads per microsatellites (Fig. 8A). The results showed more than 70% of erroneous allele lengths for homopolymers and between 49% and 86% of original allele lengths for dinucleotide repeats (Fig. 8C, Supplementary Fig. S14). Increasing the Quality Score from Q_{15} to Q_{25} did not improve the proportion of original allele length, whereas considering the forward and reverse reads separately resulted in different allele length accuracies for A/T and AC/TG microsatellites.

We next generated duplex reads achieving a mean duplex read rate of 23.5% (Fig. 8B) and selected those with minimum Quality Scores of Q_{25} and Q_{33} for microsatellite length analysis (Fig. 8A). There was no improvement in microsatellite length identification for the same minimum Quality Score of Q_{25} (Fig. 8C, Supplementary Fig. S14). However, by using a higher minimum Quality Score ($\geq Q_{33}$), the original allele fraction could be increased beyond that obtained with simplex read analysis for some microsatellites ($A_{15\text{-}20}$, $(AT)_{15\text{–}25}$ and $(AC)_{15}$). Finally, we developed a consensus length approach based on the length of the paired simplex reads identified from the duplex reads, either using the maximum length or considering the length only if equal (Fig. 8C, Supplementary Fig. S14). While the results showed little to no improvement for homopolymers, a marked improvement in the determination of dinucleotide repeat length was observed with the 'Equal Length' method, comparable to that achieved with short-read sequencing.

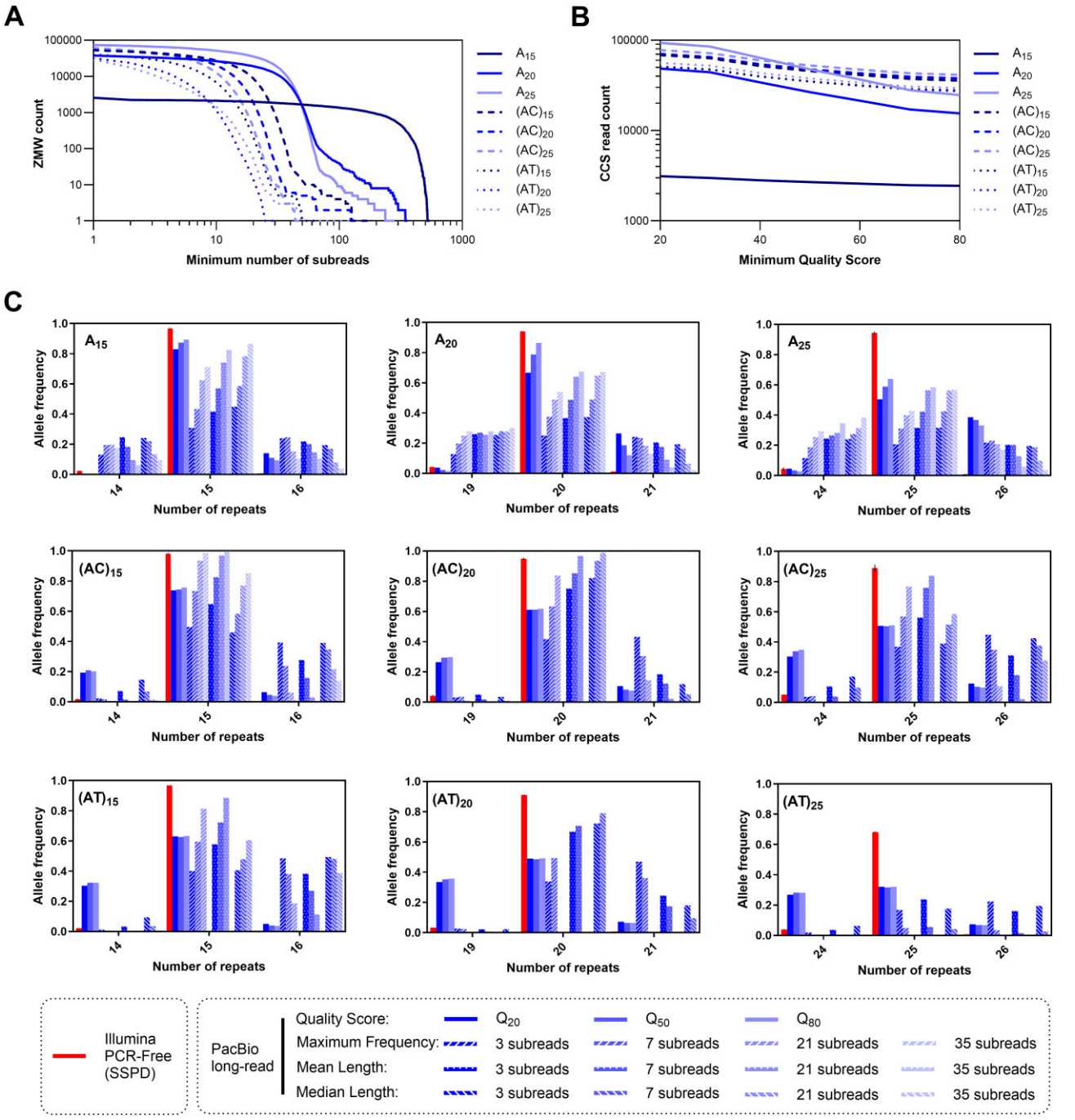


Figure 7. Accurate length determination of microsatellites from Pacbio long-read sequencing (type IV libraries) using different strategies. (**A**). ZMW count per minimum number of subreads. (**B**). CCS read count per minimum Quality Score C. *n* − 1, *n*, and *n* + 1 microsatellite allele frequencies obtained from Pacbio long-read sequencing using different approaches for accurate length determination of microsatellites. Quality score approach is based on different thresholds of CCS high-fidelity read quality. Maximum frequency, mean length and median length are in-house approaches similar to those developed for UMI-error correction (see Supplementary Fig. S6A), based on a defined number of subreads from ZMWs with a Barcode Score of 70 to generate a microsatellite consensus sequence. Conditions with fewer than 100 consensus sequences are not represented on the graph (no bars in the original n alleles). Each point corresponds to a single experiment. Type I PCR-free – SSDP data (two replicates) from Illumina short-read sequencing (NextSeq 500) were also presented in panel B for comparison.

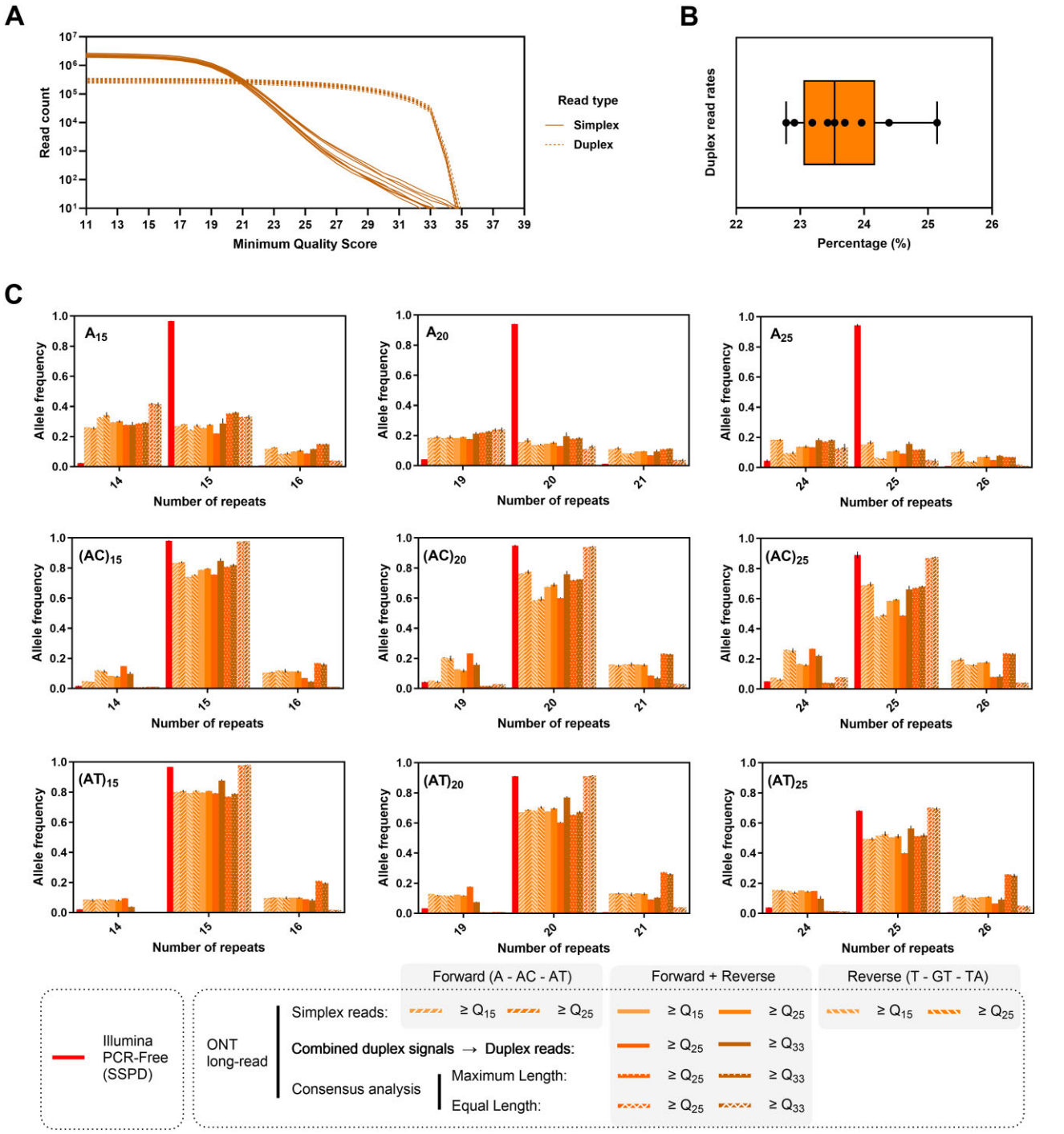


Figure 8. Accurate length determination of microsatellites from ONT long-read sequencing (type V libraries) using different strategies. (**A**). Simplex and duplex reads count per minimum Quality Score. (**B**). Duplex read rate (%) per library. (**C**). $n - 1$, n, and $n + 1$ microsatellite allele frequencies obtained from ONT long-read sequencing using different approaches for accurate length determination of microsatellites. Simplex read analysis considered either the forward, the reverse or all reads using two minimum Quality Score thresholds ($\geq Q_{15}$ and $\geq Q_{25}$). Duplex read analysis was based on two minimum Quality Score thresholds ($\geq Q_{25}$ and $\geq Q_{33}$). Consensus analysis was based on paired simplex reads identified from duplex reads ($\geq Q_{25}$ and $\geq Q_{33}$), using either maximum length or considering the length only when they were equal. Each point represents a triplicate experiment. PCR-free – SSDP data (two replicates) from Illumina short-read sequencing (NextSeq 500) were also presented for comparison.

Discussion

In this study, we provided a detailed comparison of several widely used second- and third-generation sequencing methods as well as various bioinformatics approaches for correctly determining the length of homopolymers and dinucleotide STRs. Accurate calling of microsatellite alleles and genotypes has long been known to be particularly challenging with NGS due to stutter artifacts caused by polymerase slippage during PCR steps and high sequencing error rates, notably at homopolymeric nucleotide runs [12, 25, 26, 38, 53]. Ultimately, our study aimed to identify the best NGS approach to characterize the true distribution of genuine microsatellite alleles present in a sample, minimizing or even suppressing errors introduced during library preparation, sequencing and bioinformatics analysis. This could be of particular interest for samples containing a complex mixture of variable-length alleles, such as in MSI cancers or MMRD cells.

Several plasmids containing mono- (A/T) or di-nucleotide (AC/TG or AT/TA) repeat microsatellites with 15–25 repetitions and assumed to be completely pure were used as templates for library preparation and comparative NGS experiments. We assessed three short-read sequencing approaches with Illumina, based on PCR-free and PCR-containing (1–20 cycles) Type I libraries, standard UMI error correction (Type II libraries), and duplex sequencing combined error correction (Type III libraries) using SSDP and RFP, and two types of PCR-free libraries based on PacBio (Type IV) and ONT (Type V) sequencing using linearized plasmids (Fig. 1). Thus, the SSDP condition was similar to TGS experiments where specific portions of the genome could be captured, amplified and sequenced at specific coordinates [60], while the RFP condition mimicked whole genome and exome sequencing experiments based on random fragmentation of a DNA sample [61]. Two instruments were evaluated for Illumina short-read sequencing, the iSeq 100 and the NextSeq 500. Although the latter is less recommended for amplicon sequencing experiments due to nucleotide diversity and read quality issues, we encountered no such problems with our NextSeq 500 data, which showed similar results to the iSeq 100 data (Supplementary Figs S2 and S4). This is likely due to the high amount of PhiX used and the two 8-nt degenerate barcodes sequenced in the first cycles of each read in the Type III libraries, which were run alongside with the Type I and II libraries, providing for sufficient nucleotide diversity. Another experimental setup to increase diversity could be to spike these libraries in (human) WGS runs to provide a more detailed and relevant overview of microsatellite sequencing in a real experimental context and to reduce wasted sequencing depth with PhiX, especially for high throughput sequencers. Our overall bioinformatics approach relied on the identification of microsatellite flanking regions to correctly estimate their length in each read, a simple approach that has already been used in some studies [62–64]. This strategy has proven to be very effective in obtaining consensus microsatellite alleles from multiple (sub)reads from Illumina single and dual UMIs and PacBio protocols, whereas the standard approach to generate a consensus sequence often failed due to too many indel errors and poor read quality. We developed and evaluated four modes ("Max. Freq.," "Max. Length," "Mean Length," and "Median Length") for consensus allele generation from Illumina UMI reads (Type II and III libraries) and PacBio subreads, as well as two additional modes ("Max. Length" and "Equal Length") for combined consensus and consensus allele lengths from duplex sequencing and ONT sequencing, respectively.

Standard microsatellite length distribution analysis from short-read data showed as expected the increasing effect of stuttering—mainly deletions—as the number of PCR cycles and also of motif repetition increased (Fig. 4). Stutter artifacts were visible since the first cycle of PCR applied to the library amplification, indicating that a single PCR extension step can alter the frequency of the original alleles present in a sample. As a standard approach, short reads generated from PCR-free libraries thereby presented the highest level of original alleles (higher in SSDP than in RFP condition), although not reaching 100% as ideally desired. Our results are consistent with others, which showed a lower proportion of erroneous reads at the microsatellite sequence level in PCR-free compared to PCR-containing WGS data [26].

UMI error correction and duplex sequencing combined error correction are two approaches that improve the sequencing accuracy and enable the detection of rare genetic variations, which we also assessed for their ability to correctly determine the length of microsatellites (Figs 5 and 6). The first limitation of these approaches was that their library preparation mandatorily required several PCR cycles (12, 16, and/or 20 cycles in our study), which introduced many stutter artifacts. The standard UMI error correction used in Type II and Type III libraries based on "Max. Freq." mode increased the original allele percentage (with a maximum 10-fold error rate reduction) but at a lower level than standard PCR-free libraries. Combined error correction of duplex sequencing data further reduced error rates and increased original allele proportions to the same level as or slightly above those obtained with PCR-free short-read data, but this correction was not possible for every microsatellite, notably the longest ones due to poor quality, missing and/or discordant data (Fig. 6). The "Equal Length" combined error correction exhibited higher original allele levels than with the "Max. Length" mode, but reached less frequently a consensus length as it requires two concordant $\alpha\beta$ and $\beta\alpha$ consensus alleles. Although commonly used in numerous studies focusing on read base call accuracy and point mutation detection [50, 60, 65], the impact of standard and combined UMI error correction has been less frequently described specifically in the context of microsatellite length determination, especially in long dinucleotide STRs and homopolymers ($\geq$15 repetitions). Standard UMI approaches have already been assessed in cancer studies investigating rare MSI events, mainly in homopolymers from tumor, blood or plasma DNA [66–70], as well as in forensic studies for the identification of STR-based haplotypes from complex DNA mixtures using mainly tetranucleotide STRs [71, 72]. However, the impact of UMI error correction on different repeat types and numbers has rarely been described in detail. Due to the limited performance of standard UMI error correction, more sophisticated bioinformatics algorithms have also been applied to UMI-based sequencing data to improve the detection of rare alleles [69, 70]. Recently, a protocol derived from duplex sequencing was evaluated on mononucleotide microsatellites with 8–18 repetitions. It presented 0.48% of erroneous alleles compared to 5.5% before combined error correction [73], a result close to ours.

We last assessed the ability of PacBio and ONT long-read sequencing to correctly determine the length of microsatel-

lites. Issues related to the high (indel) error rates and the accuracy of microsatellite allele calling at homopolymers and other STRs are well known and have already been reported in some PacBio and ONT long-read sequencing studies [38–40, 74–77]. Therefore, STR analysis with third-generation sequencing is thus considered to be very challenging to date. From our PacBio sequencing experiments, CCS reads of the highest quality (Q_{80}) still presented a large proportion of erroneous alleles, although slight improvement was observed for homopolymers (Fig. 7). These results were outperformed by Illumina PCR-free short-read sequencing. When making our own consensus from different numbers of subreads, the original allele proportion of dinucleotide STRs increased above the values obtained from Q_{80} CCS reads and sometimes from PCR-free short reads, indicating that subread availability and analysis could be essential for accurate identification of dinucleotide STR length. However, this improvement was not observed for homopolymers. Interestingly, for homopolymers, the CCS reads showed an erroneous allele bias toward insertions, as previously reported [38–40], but the consensus alleles generated from the subreads by our in-house algorithm presented balanced erroneous insertions and deletions. Therefore, we believe that the PacBio CCS processing algorithm, rather than by the SMT sequencing chemistry, may be responsible for this deletion bias. Regarding ONT sequencing, we obtained the poorest results for homopolymers among all sequencing chemistries used in the study (Fig. 8). These results could not be improved by the three bioinformatics approaches developed (simplex read analysis, duplex reads analysis, and consensus length analysis) and showed no more than 40% of the original allele. Conversely, we were able to accurately determine the length of dinucleotide repeats to a level comparable to short-read sequencing using the "Equal Length" consensus approach using paired reads identified from duplex reads, suggesting that ONT sequencing may be suitable for accurate allele calling of dinucleotide repeats when used with the appropriate bioinformatics analysis method.

Finally, we would like to point out some limitations of our study. Although we analyzed three types of microsatellites—among the most common in the human genome—with 15, 20, and 25 repeats, several other motifs, such as poly- (C/G), poly-(TC/AG) as well as tri-, penta- and hexa-nucleotide repeats, were not included. Moreover, our results were obtained using synthetic microsatellite sequences contained in plasmids rather than more complex genomic DNA. Therefore, the results and conclusions presented herein are specific to the analyzed microsatellite type and repeat number and may not be applicable to other microsatellites and/or genomic data.

In addition to the three sequencing technologies and four instruments benchmarked in our study, there are other technologies and instruments that were not evaluated, including some that have only recently become available. These include Illumina's new XLEAP SBS chemistry for its NextSeq 1000/2000 and NovaSeq X sequencers, the PacBio Revio instrument and the PacBio Onso sequencing-by-binding system (SBB), which its manufacturer claims can accurately sequence and resolve homopolymers and other repetitive regions. Other notable technologies include Element Biosciences' AVITI SBB system, which uses rolling circle amplification for polony generation, MGI's DNA nanoball (DNB)-based SBS instruments and Ultima Genomics' SBS technology. We anticipate evaluating these technologies in future studies to improve microsatellite sequencing and length determination.

Conclusion

In summary, our study and comparative results support the systematic use of PCR-free short-read sequencing, notably in the context of SSDP, to characterize the true microsatellite allele length distribution in a sample. The specificities of this approach can be found in STR-seq, a recent CRISPR-Cas9-targeted fragmentation and PCR-free short-read sequencing method that allows the capture and sequencing of thousands of microsatellites [78]. Alternatively, UMI-based methods, notably duplex sequencing, can also be used to detect rare microsatellite alleles. However, they might be more expensive due to the sequencing depth required, especially with duplex sequencing. PacBio long-read sequencing should be avoided for accurate identification of the length of homopolymers and dinucleotide STRs, due to its poor performance compared with short-read sequencing. Although we succeeded in improving consensus alleles of AC/TG STRs, this required 35 subreads per consensus, which might be difficult to achieve with a standard PacBio protocol with a larger library size. ONT sequencing should also be avoided for homopolymer typing, but could be used to accurately identify dinucleotide allele length when used with the "Equal Length" consensus approach developed from both reads of the same duplex read. However, it should be acknowledged that the main benefit of third-generation sequencing for STR analysis concerns long to very long STR tracts, such as those found in triplet expansion diseases, for which it outperforms short-read sequencing [39, 74, 75].

Acknowledgements

We want to acknowledge Prof. François Sigaux (MEARY Center, Saint-Louis Hospital, University of Paris) and Lucie Hernandez (INSERM U944/CNRS UMR7212, Institut de Recherche Saint-Louis, University of Paris, Paris, France) for their help with PacBio Sequel II instrument. We would also like to thank Delphine Bacq (CNRGH) for the quality control analyses of the ONT sequencing data.

Author contributions: All authors contributed significantly to this work. A.H.-K. conceived and supervised the study. S.I.J., L.M.H., A.D., M.D., and A.A. performed the experiments. Y.S. and E.T. performed the bioinformatics analyses. S.I.J., Y.S., Z.G., and A.H.-K. analyzed the data and made the figures and tables. Y.S, L.M.H., and A.H.-K. drafted the first version of the manuscript. S.I.J., Y.S., L.M.H., A.D., A.A., M.D., Z.G., E.T., J-F.D., and A.H.-K. read, edited and approved the final version of the submitted manuscript.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

The study was funded by the institutional budget of Foundation Jean Dausset - CEPH, the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013] and Foundation ARC 2019 [Project Fondation ARC, PJA

20191209442]. Funding to pay theOpen Access publication charges for this article was provided by Institutional budget of Foundation Jean Dausset - CEPH.

Data availability

All data and source code are publicly available in the Harvard Dataverse repository (https://doi.org/10.7910/DVN/0HL1FB) and in the GEO repository (GSE286306, GSE286307, GSE286308, GSE286309 and GSE286310).

References

1. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 2004;**5**:435–45. https://doi.org/10.1038/nrg1348
2. Strand M, Prolla TA, Liskay RM *et al.* Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* 1993;**365**:274–6. https://doi.org/10.1038/365274a0
3. Sia EA, Kokoska RJ, Dominska M *et al.* Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol Cell Biol* 1997;**17**:2851–8. https://doi.org/10.1128/MCB.17.5.2851
4. Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology* 2010;**138**:2073–87. https://doi.org/10.1053/j.gastro.2009.12.064
5. Hause RJ, Pritchard CC, Shendure J *et al.* Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med* 2016;**22**:1342–50. https://doi.org/10.1038/nm.4191
6. Wierdl M, Dominska M, Petes TD. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 1997;**146**:769–79. https://doi.org/10.1093/genetics/146.3.769
7. Eckert KA, Hile SE. Every microsatellite is different: intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol Carcinog* 2009;**48**:379–88. https://doi.org/10.1002/mc.20499
8. Sun JX, Helgason A, Masson G *et al.* A direct characterization of human mutation based on microsatellites. *Nat Genet* 2012;**44**:1161–5. https://doi.org/10.1038/ng.2398
9. Kelkar YD, Tyekucheva S, Chiaromonte F *et al.* The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* 2008;**18**:30–8. https://doi.org/10.1101/gr.7113408
10. Kelkar YD, Strubczewski N, Hile SE *et al.* What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biol Evolut* 2010;**2**:620–35. https://doi.org/10.1093/gbe/evq046
11. Shinde D, Lai Y, Sun F *et al.* Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res* 2003;**31**:974–80. https://doi.org/10.1093/nar/gkg178
12. Daunay A, Duval A, Baudrin LG *et al.* Low temperature isothermal amplification of microsatellites drastically reduces stutter artifact formation and improves microsatellite instability detection in cancer. *Nucleic Acids Res* 2019;**47**:e141. https://doi.org/10.1093/nar/gkz811
13. Gulcher J. Microsatellite markers for linkage and association studies. *Cold Spring Harb Protoc* 2012;**2012**:e141. https://doi.org/10.1101/pdb.top068510
14. Sainudiin R, Durrett RT, Aquadro CF *et al.* Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* 2004;**168**:383–95. https://doi.org/10.1534/genetics.103.022665
15. Putman AI, Carbone I. Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecol Evol* 2014;**4**:4399–428. https://doi.org/10.1002/ece3.1305
16. Abdul-Muneer PM. Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genet Res Int* 2014;**2014**:691759.
17. Miah G, Rafii MY, Ismail MR *et al.* A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. *Int J Mol Sci* 2013;**14**:22499–528. https://doi.org/10.3390/ijms141122499
18. Stadele V, Vigilant L. Strategies for determining kinship in wild populations using genetic data. *Ecol Evol* 2016;**6**:6107–20. https://doi.org/10.1002/ece3.2346
19. Gettings KB, Aponte RA, Vallone PM *et al.* STR allele sequence variation: current knowledge and future issues. *Forensic Sci Int Genet* 2015;**18**:118–30. https://doi.org/10.1016/j.fsigen.2015.06.005
20. Zhang XR, Meng HT, Shi JF *et al.* Efficiency evaluation of common forensic genetic markers for parentage identification involving close relatives. *Forensic Sci Int* 2023;**345**:111594. https://doi.org/10.1016/j.forsciint.2023.111594
21. Mirkin SM. Expandable DNA repeats and human disease. *Nature* 2007;**447**:932–40. https://doi.org/10.1038/nature05977
22. Lyon E, Laver T, Yu P *et al.* A simple, high-throughput assay for Fragile X expanded alleles using triple repeat primed PCR and capillary electrophoresis. *J Mol Diagn* 2010;**12**:505–11. https://doi.org/10.2353/jmoldx.2010.090229
23. Baudrin LG, Deleuze JF, How-Kit A. Molecular and computational methods for the detection of microsatellite instability in cancer. *Front Oncol* 2018;**8**:621. https://doi.org/10.3389/fonc.2018.00621
24. Vieira ML, Santini L, Diniz AL *et al.* Microsatellite markers: what they mean and why they are so useful. *Genet Mol Biol* 2016;**39**:312–28. https://doi.org/10.1590/1678-4685-GMB-2016-0027
25. Raz O, Biezuner T, Spiro A *et al.* Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Res* 2019;**47**:2436–45. https://doi.org/10.1093/nar/gky1318
26. Fungtammasan A, Ananda G, Hile SE *et al.* Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res* 2015;**25**:736–49. https://doi.org/10.1101/gr.185892.114
27. Highnam G, Franck C, Martin A *et al.* Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* 2013;**41**:e32. https://doi.org/10.1093/nar/gks981
28. Gymrek M, Golan D, Rosset S *et al.* lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res* 2012;**22**:1154–62. https://doi.org/10.1101/gr.135780.111
29. Cheng K, Bright JA, Kelly H *et al.* Developmental validation of STRmix NGS, a probabilistic genotyping tool for the interpretation of autosomal STRs from forensic profiles generated using NGS. *Forensic Sci Int Genet* 2023;**62**:102804. https://doi.org/10.1016/j.fsigen.2022.102804
30. Ganschow S, Silvery J, Kalinowski J *et al.* toaSTR: a web application for forensic STR genotyping by massively parallel sequencing. *Forensic Sci Int Genet* 2018;**37**:21–8. https://doi.org/10.1016/j.fsigen.2018.07.006
31. Tang H, Kirkness EF, Lippert C *et al.* Profiling of short-tandem-repeat disease alleles in 12,632 Human whole genomes. *Am Hum Genet* 2017;**101**:700–15. https://doi.org/10.1016/j.ajhg.2017.09.013
32. Fang H, Wu Y, Narzisi G *et al.* Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* 2014;**6**:89. https://doi.org/10.1186/s13073-014-0089-z
33. Tae H, Kim DY, McCormick J *et al.* Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. *Bioinformatics* 2014;**30**:652–9. https://doi.org/10.1093/bioinformatics/btt595
34. Ivady G, Madar L, Dzsudzsak E *et al.* Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. *Bmc Genomics [Electronic*

Resource] 2018;**19**:158. https://doi.org/10.1186/s12864-018-4544-x
35. Zavodna M, Bagshaw A, Brauning R *et al.* The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. *PLoS One* 2014;**9**:e113862. https://doi.org/10.1371/journal.pone.0113862
36. Rajan-Babu IS, Peng JJ, Chiu R *et al.* Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Med* 2021;**13**:126. https://doi.org/10.1186/s13073-021-00932-9
37. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform* 2016;**17**:154–79. https://doi.org/10.1093/bib/bbv029
38. Ross MG, Russ C, Costello M *et al.* Characterizing and measuring bias in sequence data. *Genome Biol* 2013;**14**:R51. https://doi.org/10.1186/gb-2013-14-5-r51
39. Mitsuhashi S, Frith MC, Mizuguchi T *et al.* Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol* 2019;**20**:58. https://doi.org/10.1186/s13059-019-1667-6
40. Sacristan-Horcajada E, Gonzalez-de la Fuente S, Peiro-Pastor R *et al.* ARAMIS: from systematic errors of NGS long reads to accurate assemblies. *Brief Bioinform* 2021;**22**:bbab170. https://doi.org/10.1093/bib/bbab170
41. Cao MD, Balasubramanian S, Boden M. Sequencing technologies and tools for short tandem repeat variation detection. *Briefings Bioinf* 2015;**16**:193–204. https://doi.org/10.1093/bib/bbu001
42. Fondon JW 3rd, Martin A, Richards S *et al.* Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. *PLoS One* 2012;**7**:e33036. https://doi.org/10.1371/journal.pone.0033036
43. Christopher J, Thorsen AS, Abujudeh S *et al.* Quantifying microsatellite mutation rates from intestinal stem cell dynamics in Msh2-deficient murine epithelium. *Genetics* 2019;**212**:655–65. https://doi.org/10.1534/genetics.119.302268
44. Aska EM, Zagidullin B, Pitkanen E *et al.* Single-cell mononucleotide microsatellite analysis reveals differential insertion-deletion dynamics in mouse T cells. *Front Genet* 2022;**13**:913163. https://doi.org/10.3389/fgene.2022.913163
45. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* 2013;**155**:858–68. https://doi.org/10.1016/j.cell.2013.10.015
46. Renault V, Tubacher E, How-Kit A. Assessment of microsatellite instability from next-generation sequencing data. *Adv Exp Med Biol* 2022;**1361**:75–100. https://doi.org/10.1007/978-3-030-91836-1_5
47. Wang Y, Zhang X, Xiao X *et al.* Accurately estimating the length distributions of genomic micro-satellites by tumor purity deconvolution. *BMC Bioinf* 2020;**21**:82. https://doi.org/10.1186/s12859-020-3349-5
48. Novroski NMM, Wendt FR, Woerner AE *et al.* Expanding beyond the current core STR loci: an exploration of 73 STR markers with increased diversity for enhanced DNA mixture deconvolution. *Forensic Sci Int Genet* 2019;**38**:121–9. https://doi.org/10.1016/j.fsigen.2018.10.013
49. Kozarewa I, Ning Z, Quail MA *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 2009;**6**:291–5. https://doi.org/10.1038/nmeth.1311
50. Schmitt MW, Kennedy SR, Salk JJ *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci USA* 2012;**109**:14508–13. https://doi.org/10.1073/pnas.1208715109
51. Kivioja T, Vaharautio A, Karlsson K *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 2012;**9**:72–4. https://doi.org/10.1038/nmeth.1778
52. Rhoads A, Au KF. PacBio Sequencing and its applications. *Genom Proteom Bioinform* 2015;**13**:278–89. https://doi.org/10.1016/j.gpb.2015.08.002
53. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* 2021;**3**:lqab019. https://doi.org/10.1093/nargab/lqab019
54. Yamaguchi I, Watanabe T, Ohara O *et al.* PCR-free whole exome sequencing: cost-effective and efficient in detecting rare mutations. *PLoS One* 2019;**14**:e0222562. https://doi.org/10.1371/journal.pone.0222562
55. Kinde I, Wu J, Papadopoulos N *et al.* Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 2011;**108**:9530–5. https://doi.org/10.1073/pnas.1105422108
56. Sloan DB, Broz AK, Sharbrough J *et al.* Detecting rare mutations and DNA damage with sequencing-based methods. *Trends Biotechnol* 2018;**36**:729–40. https://doi.org/10.1016/j.tibtech.2018.02.009
57. Iizuka R, Yamazaki H, Uemura S. Zero-mode waveguides and nanopore-based sequencing technologies accelerate single-molecule studies. *Biophysics* 2022;**19**:e190032. https://doi.org/10.2142/biophysico.bppb-v19.0032
58. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;**21**:597–614. https://doi.org/10.1038/s41576-020-0236-x
59. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol* 2016;**34**:518–24. https://doi.org/10.1038/nbt.3423
60. Hiatt JB, Pritchard CC, Salipante SJ *et al.* Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* 2013;**23**:843–54. https://doi.org/10.1101/gr.147686.112
61. Zhao Y, Fang LT, Shen TW *et al.* Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. *Sci Data* 2021;**8**:296. https://doi.org/10.1038/s41597-021-01077-5
62. Gan C, Love C, Beshay V *et al.* Applicability of next generation sequencing technology in microsatellite instability testing. *Genes* 2015;**6**:46–59. https://doi.org/10.3390/genes6010046
63. Warshauer DH, Lin D, Hari K *et al.* STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet* 2013;**7**:409–17. https://doi.org/10.1016/j.fsigen.2013.04.005
64. Herbreteau G, Airaud F, Pierre-Noel E *et al.* MEM: an algorithm for the reliable detection of microsatellite instability (MSI) on a small NGS panel in colorectal cancer. *Cancers* 2021;**13**:4203. https://doi.org/10.3390/cancers13164203
65. Wang TT, Abelson S, Zou J *et al.* High efficiency error suppression for accurate detection of low-frequency variants. *Nucleic Acids Res* 2019;**47**:e87. https://doi.org/10.1093/nar/gkz474
66. Gallon R, Sheth H, Hayes C *et al.* Sequencing-based microsatellite instability testing using as few as six markers for high-throughput clinical diagnostics. *Hum Mutat* 2020;**41**:332–41. https://doi.org/10.1002/humu.23906
67. Gallon R, Muhlegger B, Wenzel SS *et al.* A sensitive and scalable microsatellite instability assay to diagnose constitutional mismatch repair deficiency by sequencing of peripheral blood leukocytes. *Hum Mutat* 2019;**40**:649–55. https://doi.org/10.1002/humu.23721
68. Waalkes A, Smith N, Penewit K *et al.* Accurate pan-cancer molecular diagnosis of microsatellite instability by single-molecule molecular inversion probe capture and high-throughput sequencing. *Clin Chem* 2018;**64**:950–8. https://doi.org/10.1373/clinchem.2017.285981
69. Georgiadis A, Durham JN, Keefer LA *et al.* Noninvasive detection of microsatellite instability and high tumor mutation burden in cancer patients treated with PD-1 blockade. *Clin Cancer Res* 2019;**25**:7024–34. https://doi.org/10.1158/1078-0432.CCR-19-1372
70. Willis J, Lefterova MI, Artyomenko A *et al.* Validation of microsatellite instability detection using a comprehensive plasma-based genotyping panel. *Clin Cancer Res*

2019;**25**:7035–45.
https://doi.org/10.1158/1078-0432.CCR-19-1324
71. Woerner AE, Mandape S, King JL *et al.* Reducing noise and stutter in short tandem repeat loci with unique molecular identifiers. *Forensic Sci Int Genet* 2021;**51**:102459.
https://doi.org/10.1016/j.fsigen.2020.102459
72. Crysup B, Mandape S, King JL *et al.* Using unique molecular identifiers to improve allele calling in low-template mixtures. *Forensic Sci Int Genet* 2023;**63**:102807.
https://doi.org/10.1016/j.fsigen.2022.102807
73. Bae JH, Liu R, Roberts E *et al.* Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. *Nat Genet* 2023;**55**:871–9. https://doi.org/10.1038/s41588-023-01376-0
74. Dolzhenko E, English A, Dashnow H *et al.* Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* 2024;**42**:1606–14. https://doi.org/10.1038/s41587-023-02057-3
75. Chiu R, Rajan-Babu IS, Friedman JM *et al.* Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol* 2021;**22**:1606–14.
https://doi.org/10.1186/s13059-021-02447-3
76. Fang L, Liu Q, Monteys AM *et al.* DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biol* 2022;**23**:108.
https://doi.org/10.1186/s13059-022-02670-6
77. Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. *Bmc Genomics [Electronic Resource]* 2020;**21**:889.
https://doi.org/10.1186/s12864-020-07227-0
78. Shin G, Grimes SM, Lee H *et al.* CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nat Commun* 2017;**8**:14291.
https://doi.org/10.1038/ncomms14291

Received: July 12, 2024. **Revised:** January 13, 2025. **Editorial Decision:** February 8, 2025. **Accepted:** February 11, 2025