GENETICS

The application of Nanopore sequencing for variant calling on the human mitochondrial DNA

Anton Shikov[1,2,3], Viktoriya Tsay[1], Mikhail Fedyakov[1], Yuri Eismont[1], Alena Rudnik[1], Stanislav Urasov[1], Sergey Sherbak[1,3], and Oleg Glotov[1,4]

[1]Genetics Laboratory, City Hospital No. 40, ul. Borisova, 9, Saint Petersburg, 197706, Russian Federation
[2]All-Russia Research Institute for Agricultural Microbiology, Shosse Podbel'skogo, 3, Saint Petersburg, 190608, Russian Federation
[3]Faculty of Medicine, Saint Petersburg State University, 21-ya liniya, 8a, Saint Petersburg, 199106, Russian Federation
[4]Department of Genomic Medicine, D. O. Ott Research Institute of Obstetrics, Gynecology and Reproductology, Mendeleyevskaya liniya, 3, Saint Petersburg, 199034, Russian Federation

Address correspondence and requests for materials to Anton Shikov, antonshikov96@gmail.com

Abstract

The emergence of long-read sequencing technologies has made a revolutionary step in genome biology and medicine. However, long reads are characterized by a relatively high error rate, impairing their usage for variant calling as a part of routine practice. Thus, we here examine different popular variant callers on long-read sequences of the human mitochondrial genome, convenient in terms of small size and easily obtained high coverage. The sequencing of mitochondrial DNA from 8 patients was conducted via Illumina (MiSeq) and the Oxford Nanopore platform (MinION), with the former utilized as a gold standard when evaluating variant calling's accuracy. We used a conventional GATK3-BWA-based pipeline for paired-end reads and Guppy basecaller coupled with minimap2 for MinION data, respectively. We then compared the outputs of Clairvoyante, Nanopolish, GATK3, Longshot, DeepVariant, and Varscan tools applied on long-read alignments by analyzing false-positive and false-negative rates. While for most callers, raw signals represented false positives due to homopolymeric errors, Nanopolish demonstrated both high similarity (Jaccard coefficient of 0.82) and a comparable number of calls with the Illumina data (140 vs. 154) with the best performance according to AUC (area under ROC curve, 0.953) as well. In sum, our results, despite being obtained from a small dataset, provide evidence that sufficient coverage coupled with an optimal pipeline could make long reads of mitochondrial DNA applicable for variant calling.

Keywords: next-generation sequencing, Oxford Nanopore, Illumina, variant calling, mitochondrial DNA

Citation: Shikov, A., Tsay, V., Fedyakov, M., Eismont, Y., Rudnik, A., Urasov, S., Sherbak, S., and Glotov, O. 2021. The application of Nanopore sequencing for variant calling on the human mitochondrial DNA. Bio. Comm. 66(2): 109–123. https://doi.org/10.21638/spbu03.2021.202

Authors' information: Anton Shikov, PhD student, Research Engineer, orcid.org/0000-0001-7084-0177; Viktoriya Tsay, Specialist, orcid.org/0000-0001-6488-8369; Mikhail Fedyakov, Specialist, orcid.org/0000-0002-3291-3811; Yuri Eismont, PhD, Specialist, orcid.org/0000-0002-4828-8053; Alena Rudnik, PhD, Physician, orcid.org/0000-0001-9315-1040; Stanislav Urasov, Physician, orcid.org/0000-0002-5441-2911; Sergey Sherbak, Dr. of Sci. in Medicine, Professor, orcid.org/0000-0001-5036-1259; Oleg Glotov, PhD, Senior Researcher, Head of Laboratory, orcid.org/0000-0002-0091-2224

Manuscript Editor: Pavel Skutschas, Department of Vertebrate Zoology, Faculty of Biology, Saint Petersburg State University, Saint Petersburg, Russia

Received: August 22, 2020;

Revised: March 16, 2021;

Accepted: April 1, 2021.

Funding: No funding information provided.

Ethics statement: The study was carried out following the Declaration of Helsinki (1989) of the World Medical Association and under current ethical guidelines with the approval of the Local Ethics Committee, City Hospital No. 40, Saint Petersburg (Approval # 119 from February 9, 2017).

Competing interests: The authors have declared that no competing interests exist.

Introduction

Next-generation sequencing (NGS) techniques have tremendously facilitated the sequencing of human genomes; consequently, they are now ubiquitously used in genomic medicine (Goodwin et al., 2016). Despite wide implementation in clinical practice (Di Resta and Ferrari, 2018), sequencing-by-synthesis methods, such as bridge amplification in the Illumina platform, have certain limitations due to the short-sized reads obtained, which makes them inapplicable for haplotype analysis (Bansal and Bafna, 2008). A significant step has been made with the occurrence of third-generation NGS approaches like those provided by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms, able to generate fragments up to 30 kb in size (Ardui et al., 2018). Unfortunately, Nano-

pore sequencing is still accompanied by a high error rate, resulting in low per-base quality (Bowden et al., 2019); nevertheless, over the last years, unprecedented progress has been made in the improvement of third-generation sequencing technologies. Not only is the performance of flow cells improving, but also the robustness and the efficacy of base-calling and error-correction algorithms are growing (Edge and Bansal, 2019).

Over the last decades, the extensive study of mitochondrial genetics has provided evidence that impairments in mitochondrial DNA are linked with various pathological conditions (Dashti et al., 2021). Genes presented in mtDNA encode transcripts for independent protein synthesis in mitochondria as well as proteins involved in oxidative phosphorylation (Onyango et al., 2016). Given its role in energetic metabolism, severe alterations in mitochondrial operation could initiate the development of multi-factored diseases. For example, more than 300 variants across the mtGenome (Li et al., 2019) have been demonstrated to increase the risk of Alzheimer's disease (Mannelli et al., 2015), dementia (Tranah et al., 2012), parkinsonism (Coxhead et al., 2016), Huntington's Disease (Banoei et al., 2007), and various neurodegenerative disorders (Simon et al., 1999). This link is usually explained by impairments in respiration, resulting in ROS accumulation and oxidative damage in neurons (Purevsuren et al., 2009). Apart from neurodegenerative conditions, mitochondrial dysfunctions could exhibit comorbidity with diabetes (Naing et al., 2014). Finally, destabilizations in the oxidative phosphorylation complex are related to cancer development (Lee et al., 2004). It was shown that mtDNA tends to accumulate mutations with age, contributing to age-related neurodegenerative disorders, diabetes, and cell malignization (Li et al., 2019).

The information mentioned above clearly indicates the necessity of precise variant calling while analyzing mtDNA for risk predictions and early disease diagnosis. Apart from clinical importance, mitochondrial DNA is convenient for methodological studies when comparing different variant calling tools. These reasons include small size virtually neutralizing the effects related to reference alignment, a considerable number of mtDNA copies providing sufficient coverage, and haploid phasing of variants, which increases recall rate (Alkanaq et al., 2019). Current interest in clinical aspects of mitochondrial genetics accelerates the development of diverse techniques to enhance the sequencing efficacy of mitochondrial DNA, applying capture-based approaches (Zhou et al., 2020) and preprocessing methods (Yao et al., 2019) and control region validation assay (Brandhagen et al., 2020). Nonetheless, accepted recommendations for NGS-driven analysis are focused on Illumina exclusively, and the diagnosing yield is discussed in the context of an adequate panel to be utilized (either target, whole-exome, or whole-genome ones) (Watson et al., 2020). The overwhelming majority of modern clinical studies, either identifying novel pathogenic variants or examining diseases' etiology regarding impairments in the mitochondrial genome, despite applying diverse sequencing approaches, still do not implement Nanopore data (Watson et al., 2020). The reason underlying this phenomenon could be explained by the fact that best practices for variant calling analysis for medical purposes have been developed and widely accepted for short reads only (Koboldt, 2020).

At the same time, long reads have found application in revealing structural variants (Aganezov et al., 2020), phasing analysis with haplotype reconstruction (Maestri et al., 2020; Popitsch et al., 2020), and the nanopore-based assay for analyzing leukemic samples (Orsini et al., 2018; Cumbo et al., 2019). Nevertheless, reports on long reads' usage with regard to mitochondrial genetics are still scarce. It has been shown, however, that the ONT platform could provide a relevant sequencing result for the full mtGenome (Zascavage et al., 2019) and serve as a source for genotyping single nucleotide variants (Alkanaq et al., 2019) as well as large deletions (Wood et al., 2019) with accuracy relevantly comparable to short reads. Even so, in these studies, only two instruments were applied, namely, PacBio's SMRTtools (Pacific Biosciences Inc., Menlo Park, CA, USA) and Varscan (Koboldt et al., 2012), respectively, while there are available tools specifically devised for variant calling on long reads, such as Longshot (Edge and Bansal, 2019), Clairvoyante (Luo et al., 2019) or Nanopolish (Loman et al., 2015), and no detailed comparative analysis on mitochondrial data have been carried out yet. Moreover, to the best of our knowledge, no recommendations in choosing an optimal pipeline to identify single nucleotide polymorphisms in mtGenome have been made. Thus, our study aimed to compare several variant calling algorithms using the Illumina platform and the Oxford Nanopore techniques, attempting to dissect the most significant factors contributing to obtaining either false-positive or false-negative calls.

Materials and methods

Patients' DNA extraction

All patients (n = 8) were referred to our department under observation for suspected mitochondrial diseases (clinical and instrumental signs of myopathy, neuropathy, oculopathy, cardiomyopathy, high blood levels of lactate, pyruvate, lactate/pyruvate ratio). Notably, previous clinical exome sequencing had not shown any genetic diseases for these patients. Informed consent from all the participants was obtained before including patients' data in the analysis.

Library preparation for the Illumina data

The samples' DNA was extracted from WBC (white blood cells) via the automated MagNA Pure Compact System (Roche Life Science) based on magnetic-beads technology, and all sample preparation procedures were conducted in accordance with the standard Illumina protocol (Human mtDNA Genome). We obtained two long fragments when performing two PCR reactions for further sequencing with the target size of each PCR amplicon in the range of 7800–10400 bp for MTL1 primer pairs and 9500–12500 bp for MTL2, respectively. Subsequently, the quantity of each sample was assessed with Quantus Fluorometer TM (Promega Corp., USA). The NGS itself comprised four stages described in Kapa Hyper Prep Kit and Illumina platform (MiSeq, USA) protocols, namely, fragmentation, end-repair, A-tailing, and adaptor ligation. We first pooled two PCR-products according to the volumes presented in Kapa Hyper Prep Kit and utilized Covaris S220 to obtain fragments with a size range of 200–250 bp. Following fragmentation, each sample was quantified via the Agilent 4200 TapeStation system, which was also applied to assess the samples' quality coupled with Quantus Fluorometer. All libraries displayed a fragment size of ~350bp and yield ~1μg.

Oxford Nanopore library preparation and sequencing

Similar to short reads, amplicons from the respective samples were pooled. Subsequently, library preparation was carried out according to SQK-LSK109 (Oxford Nanopore Technologies, Oxford, UK) protocol with the following steps: (1) end prep, (2) ligation of barcode adapters, (3) PCR barcoding, (4) DNA repair and end-prep. The final product was cleaned up and quantified via Quantus Fluorometer. Prepared sample libraries were loaded to FLO-MINSP6 flowcell of the MinION platform with 24 hours of active flowcell sequencing.

Illumina data analysis

We constructed the variant calling pipeline for short reads according to Genome Analysis Toolkit (GATK) best practices recommendations (Van der Auwera et al., 2013). Reads, quality-checked via FastQC toolkit, were then mapped to reference hg19 human genome with BWA (Burrows-Wheeler Algorithm) aligner v.0.7.12-r1044 (Li et al., 2009). We applied the Picard MarkDuplicates utility for deduplication and realigned reads on target intervals (MT chromosome, 1-16569) with IndelRealigner following RealignerTargetCreator, respectively. After that, realigned BAM files were exposed to the base recalibration (BQSR) procedure via GATK BaseRecalibrator. Variants were identified using GATK HaplotypeCaller in the emit reference confidence (ERC) GVCF mode. All samples included in the dataset were jointly genotyped with the GATK GenotypeGVCFs tool. Variant filtration was done using Variant Quality Score Recalibration (VQSR) with stringent filtering criteria (SNP and insertion and deletions (INDEL) sensitivity 90.0) (DePristo et al., 2011). Finally, VCF files were annotated with SnpEff and SnpSift (Cingolani et al., 2012) using dbSNP v.151 and ClinVar v. 2019-01-10 databases.

ONT data analysis

The pipeline scheme for processing Nanopore and Illumina data is presented in Suppl. 1. Raw ONT sequencing files in FAST5 format obtained with MinKNOW software were processed with Guppy v3.2.1 (guppy_basecaller). We then cleaned individual FASTQ files from insignificant technical sequences applying guppy_barcoder in the "*trim_barcodes*" mode. Only those reads that passed filters were retained for further analysis. Selected reads were then aligned on the reference mitochondrial genome (NC_012920) with the minimap2 v2.17 aligner (Li, 2018), and the BAM files obtained were sorted and indexed. The sequencing depth was evaluated via SAMtools depth (Li and Durbin, 2010) (Table S1, Table S2, Supplementary Data Set[1]). We then launched several callers to identify variants: Clairvoyante v1.02 (Luo et al., 2019) with a pre-trained model for Minion R9.4 samples (NA12878), Nanopolish v0.13.1 (Loman et al., 2015) with reads previously indexed via Nanopolish index, Longshot v0.4.3 (Edge and Bansal, 2019), and DeepVariant v1.1.0 (Poplin et al., 2018) utilities were launched in defaults modes. GATK was applied identically to what was described above, and Varscan v2.4.0 was launched at the SAMtools mpileup output with strict settings (–min-coverage 100, min-reads2 50) due to high sequencing depth (Koboldt et al., 2012).

Statistical analysis

Results from VCF files were aggregated into a single table (Table S3) using a custom Python script, gathering the data referred to variants' annotation, allele frequency (AF), GC-content, per-site depth, calling type (homozygous/heterozygous). GC-content was evaluated in accordance with the allele frequency for each site, multiplying the number of purines/pyrimidines by AF rates for alternative and reference alleles and averaging the results. Apart from the summarized table, the same characteristics were reported for each instrument separately (Table S7). If the data was missed, as in the case of false-negative calls, the average values were attributed. The variants reported were checked for consistency with respective Illumina-based results and characterized ac-

[1] Supplemental material to the article is available at https://biocomm.spbu.ru/article/view/8649.

cordingly (namely, as true positives, false negatives, and false positives).

Statistical analysis was performed using R v.3.6.2 language. We applied the nonparametric Wilcox test for group comparisons and adjusted resultant p-values if multiple comparisons occurred using FDR (false discovery rate) adjustment. Several approaches were obtained to test variant callers' efficacy. To characterize the similarity of the outputs, we calculated the pairwise Jaccard coefficient as a ratio of common variants to their sum (Suppl. 3). The data were binarized by assigning 1 and 0 if the variants were reported/absent in the output. Next, we used these binarized vectors to build characteristics ROC (receiver operating characteristic) curves which were compared with DeLong's test followed by FDR adjustment of p-values using the pROC v1.17.0.1 package (Robin et al., 2011). To determine the most significant factors affecting long reads-based variant calling, we built a generalized linear mixed model with tools interpreted as groups to account for mixed effects utilizing the glmer function from the lmer4 v1.1-26 package (Bates et al., 2015). The model selection process was carried out according to the algorithm described by Peng and Lu (2012). Data visualization was performed with the ggplot2 v.3.3.0 (Wickham, 2016) and ggsci v2.9 packages.

Results

Sequencing output

In our study, the library comprising 8 samples with individual barcodes was analyzed. After 24h MinION flowcell sequencing, a total of 399 FAST5 files (62 Gb) were generated. The FAST5-to-FASTQ conversion dramatically reduced the data amount to 8.9 Gb of files containing 1,593,000 collected reads. The amount of analyzable reads gradually decreased when considering passing filters (1,463,255), inferring those with recognizable barcodes (932,558), and finally, selecting successfully mapped ones (651,610). No such dramatic loss (almost 2.5 times) was noticed for the Illumina data. Raw signals were transformed to 16 paired FASTQ files with a total volume of 569 Mb via Illumina bcl2fastq2 conversion software, and the demultiplexed files obtained comprised 7,291,354 reads, of which 4,406,087 were aligned on the reference mtDNA sequence. Although the resultant loss for short reads was lower (1.6 times), long reads were more robust to alignment-based loss (1.4 times instead).

Mean coverage for aligned reads was 2872x (ranging from 21x to 7863x) for the Illumina platform and 3867x (with a range from 14x to 9478x) for Nanopore, respectively (Fig. 1a, Table S1, Table S2). Despite the presence of such poorly covered regions, they were noticed only at single-nucleotide sites (Suppl. 2); hence, it did not affect the distribution of variants' detection across the genome on the whole (Fig. 1b). However, the patterns of per-site coverage for both platforms were distinct. While Illumina provided relatively uniform coverage close to mean values with peaks in flanking intervals, a conspicuous drop to an average of 847x at the 5 kb interval was noticed for long reads (p-value < 2.2e-16 with Wilcox test). Remarkably, a significantly high difference in this region was also shown for short reads (2160x, p-value < 2.2e-16 with Wilcox test), even when excluding flanking peaks, which resulted in the interval characterized by 2528x coverage. Thus, for both NGS technologies, less covered regions occurred due to chemistry features presumably at the amplicons' borders; however, in terms of absolute values, this more profoundly affected Nanopore reads. Such effects should be taken into consideration; nevertheless, in this study, it did not exert an effect on the variant calling procedure. Nanopore data is commonly associated with homopolymer errors, resulting in a high error rate in alignment files. On that account, we calculated the error rate within the samples using the SAMtools stat utility. The mean error rate obtained was 8.2 %, which is quite typical for Nanopore-generated reads, while for short reads, it reached 1.8 % only.

Variant calling results

We then estimated raw outputs from variant calling pipelines (Fig. 1c, Table S3, Table S4). After summarizing the results of different sequencing platforms and callers, 627 variants were identified, of which the minimum number was shown for Longshot (47) and the maximum ones for DeepVariant (404), whereas GATK applied on short reads reported 140 variants (Fig. 1c, Table S4). Of all the callers, the most similar results to the Illumina-based pipeline were observed for Varscan and Nanopolish programs with 129 and 154 variants, respectively. Noteworthy, we found a significant correlation (p-value < 0.0017 using linear regression model) between the sum of detected variants and the number of unique calls. Clairvoyante and DeepVariant detected 109 instrument-restricted variants each, while Longshot output did not contain any unique calls (Fig. 1d, Table S4). Notably, 26 polymorphisms were identified with all sequencing approaches and variant calling tools, 4 of which were also presented in all samples (Table S3), namely, rs2853518 (750A>G), rs2001030 (1438A>G), rs2854128 (2706A>G), and rs193302980/rs527236041 (14766C>T). Such an observation could indicate that some specific regions, detectable for both short and long reads, could be considered as target points to evaluate Nanopore-based variant calling accuracy.

In order to check the consistency of results, we performed paired comparisons between different variant callers and sequencing platforms using the Jaccard

Table 1. Missense variants' detection with different sequencing platforms and variant callers

Patients	Variant	Clairvoyante	DeepVariant	GATK	Nanopolish	Longshot	Varscan
1	3505A>G	+	+	•	+	•	+
5,6,7,8	4216T>C	+	+	+	+	•	+
7	4917A>G	+	+	+	+	•	+
4	4960C>T	+	+	+	+	+	+
1	5046G>A	+	+	+	+	•	+
1	5460G>A	+	+	+	+	•	+
4	8084A>G	+	+	•	+	•	+
4	8472C>T	+	+	•	+	•	+
2	8684C>T	+	+	+	+	•	+
4	8836A>G	+	+	+	+	•	+
3	9055G>A	+	+	•	+	•	+
2	10084T>C	+	+	•	+	•	•
8	10192C>T	+	+	•	+	•	•
3,5,6,8	10398A>G	+	+	+	+	+	+
5,6,8	13708G>A	+	+	+	+	•	+
5,6	13934C>T	+	+	+	+	+	+
1,2,3,4,5,6,7,8	14766C>T	+	+	+	+	+	+
4	14777A>C	+	+	+	+	•	•
3,5,6,7,8	14798T>C	+	+	+	+	+	+
1,2,3,4,5,6,7,8	15326A>G	+	+	+	+	•	+
5,6,7,8	15452C>A	+	+	+	+	•	+
3	15479T>C	+	+	•	+	+	+
1	15884G>C	+	+	+	+	•	+

distance metrics and presented the results as a heatmap (Suppl. 3). As could be inferred from it, the highest similarity score was observed between the Illumina pipeline and the Nanopolish tool (0.82). We then extracted 23 missense variants with a MODERATE effect according to SnpEff annotation, as missense mutations are of great importance for clinical genetics, and thus they should be detected accurately (Table 1, Table S6). Clairvoyante, DeepVariant, and Nanopolish managed to report all of these SNPs, while Longshot demonstrated the worst performance with only 6 variants reported. Notably, 2 variants were found in all samples: previously described rs193302980/rs527236041 and missed by Longshot rs2853508 (15326A>G).

Comparative analysis of variant callers

As the Jaccard coefficient showed similarity superficially without the ability to infer information about tools' performance, we built ROC (receiver operating characteristic) curves on binarized variant calling results using the Illumina data as a comparison reference. Longshot outperformed other tools in terms of the false-positive rate, meaning that almost all the calls represented true positives; however, it also demonstrated one of the smallest AUC (area under ROC curve, 0.668), outstripping GATK only (0.591). In contrast, Nanopolish seemed to show an appropriate trade-off between sensitivity and accuracy, with the highest AUC of 0.953.

Next, we pried which factors exhibited the most significant contribution to either missing or misreporting variants. To this end, we first analyzed two variants that were detected with Illumina-based pipeline exclusively. These variants, namely rs878871521 (302A>AC) and rs369786048 (310T>TC), were annotated as INDELs. Notably, there were four INDELs in the short reads data, while the other two, rs879005804 (451A>AT) and rs78907894 (513GCA>G), were detectable for long-reads. While for all these variants, the mean allele frequency was 0.966, indicating the predominance of alter-

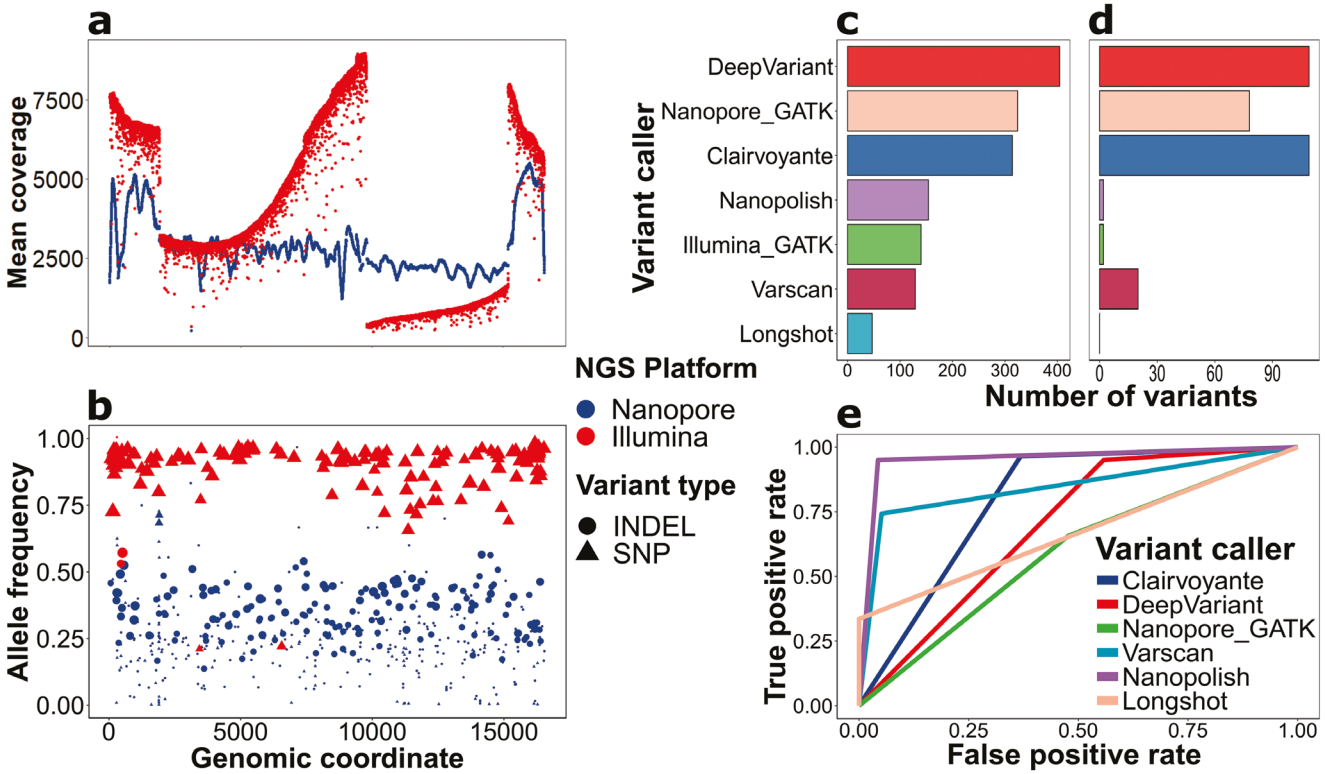


Fig. 1. The comparative analysis of different sequencing platforms and variant calling pipelines. **(a)** Mean per-site sequencing depth across the mitochondrial genome. Red color denotes the Illumina data; blue color refers to the ONT long reads. **(b)** Shown is the distribution of the variants obtained. Color code is identical to **(a)**; circles stand for INDELs, and triangles represent SNPs; the size of the points is proportional to the number of programs reported the variant. **(c)** The sum of raw calls for each pipeline. **(d)** Cumulative numbers of variants uniquely called with each instrument. **(e)** ROC (receiver operating characteristic) curves of false-positive and false-negative rates for callers applied at the Nanopore date comparing to the Illumina-based calls.

native alleles in the reads, the missed ones were characterized by considerably lower coverage (488x) compared to average per-site depth (1836x). It is noteworthy that two identifiable INDELs, despite their apparent homozygotic nature, were reported as heterozygotes by variant callers applied at Nanopore reads due to their mean allele frequency of 0.465.

Considering all the observations, we summarized properties that could possibly affect variant calling, namely, GC-content, alternative allele frequencies (AF), variant type (SNP/INDEL, homozygotes/heterozygotes), and per site-coverage for each calling pipeline (Table S8, Suppl. 4). Differences in GC-content were non-significant, whereas for other quantitative variables, such as coverage and AF, a highly significant difference between tools was observed according to the pairwise Wilcox test followed by FDR adjustment (Table S9). Similarly, a significant result was observed when comparing the sums of INDELs and heterozygous calls (Chi-squared test p-value < 2.2e-16 for both types); however, these sums were proportional to each other, which was confirmed by Cramer's V coefficient (0.721). Notably, the mean per-site depth was strikingly different, ranging from 23 in GATK applied at long reads to 9747 for Nanopolish (Suppl. 4e, Table S8). Such a discrepancy could be explained by the programs' behavior with reads' quality: instruments devised for short reads usually have strict quality cut-offs, while tools designed for long reads mainly operate with nominal coverage.

While the results obtained are informative, they do not link variants' properties with the calling performance. To this end, we characterized each variant with the aforementioned properties specified in terms of consistency with the Illumina data as true positive, false negative, or false positive (Table S10). Next, we built a generalized mixed effects logistic regression model to identify factors affecting variant calling success significantly. After the model selection process, we considered three variables: per-site coverage, variant type (INDEL/SNP), and AF. We built models both with or without Illumina results, which were practically identical and provided almost the same p-values, and we also constructed individual models to delineate the patterns of false positives, false negatives, and all true/false calls (Suppl. 5, Suppl. 6). The latter groups slightly differed in terms of p-values. Allele frequency was the most significant factor (p-value < 2e-16) regardless of the model used; SNP/INDEL attribution was also shown to significantly affect variant calling with p-values of 9.89e-16 1.17e-14, and 6.49e-06 for the combined model, false-positive- and

false-negative–restricted models, respectively. For mean coverage, however, almost no impact was reported with the exception of the combined model, where a p-value on the border of significance (0.0633) was observed. Despite the variables' non-multicollinearity according to VIF (variance inflation factor) function, variant attribution was significantly associated with AF (p-value < 2e-16 using logistic regression), indicating that false-positively called INDELs tend to display smaller alternative allele frequency. The graphical representation of the models corroborated and complemented ROC curves-based results (Suppl. 6). Similar to ROC curves, Longshot was the closest to Illumina-based pipeline in terms of false-positive errors (Suppl. 6b), and Nanopolish showed the best overall performance (Suppl. 6a). Interestingly, Clairvoyante was the most robust to false negatives of all the tools (Suppl. 6c), although this did not compensate a lot of erroneously detected INDELS.

Discussion

The main advantages of MinION discussed in the literature are easiness in handling, rapid development, expected cost reduction of medical examination in the coming years, and the ability to generate reads up to megabases long, which makes them of great potential to be included in clinical practice (Edge and Bansal, 2019). Nanopore reads find extensive implementation in forensic sequencing (Cornelis et al., 2019), characterizing structural variations and somatic mutations in cancer samples (Orsini et al., 2018; Cumbo et al., 2019; Aganezov et al., 2020), typing STEC O157:H7 Shiga toxin-producing *Escherichia coli* isolates for health monitoring (Greig et al., 2019). The usage of long reads for mitochondrial genome sequencing was reported for vertebrate species identification (Franco-Sierra and Díaz-Nieto, 2020), equine genetics (Dhorne-Pollet et al., 2020), the structural-wise grouping of plan mitochondrial genomes (Masutani et al., 2021), and clinical diagnostics (Wood et al., 2019). Nevertheless, despite the growing number of studies, long reads are still rarely applied as clinical panels (Orsini et al., 2018) due to low base-to-base quality, hampering their application for variant calling in comparison with short reads, for which protocols, standards, and recommendations are being continuously developed (DePristo et al., 2011; Koboldt et al., 2012; Van der Auwera et al., 2013; Koboldt, 2020; Watson et al., 2020).

The popularity of second-generation sequencing techniques could be explained by their accurate (>99.5% accuracy) output that can be applied in clinical practice; however, these approaches are not free from specific errors related to sequencing design (Slatko B. et al., 2018). For instance, the Ion torrent technique is prone to homopolymer errors, while the Illumina platform may be accompanied by substitution errors (Quail et al., 2012). Additionally, being short-sized, these reads could be inadequately mapped to homologous regions of the reference genome (Menzel et al., 2013). For the mitochondrial genome, this effect is even more profound due to the presence of nuclear DNA of mitochondrial origin (nuMTs), thence in clinical/whole-exome or whole-genome panes, false mapping of such reads would substantially impair variant calling. Finally, short reads do not allow detecting large structural variations or performing phasing analysis of mutations (Suzuki et al., 2017). In contrast, long reads are considered to be of great potential to detect links even between distant mutations affecting the same gene, which is extremely important for monitoring chronic leukemia progression (Orsini et al., 2018), and they also could evaluate mitochondrial heteroplasmy (Zascavage et al., 2019). While variant calling on long reads is still scarce, some studies reported at least comparable performance between them and conventional short reads-based pipelines (Orsini et al., 2018; Alkanaq et al., 2019; Greig et al., 2019; Wood et al., 2019; Magdy et al., 2020). Furthermore, new instruments for specific purposes have been devised recently, such as Nanopanel2 for identifying low-frequency somatic polymorphisms (Popitsch et al., 2020) or NanoVar for characterizing structural variations in low-depth sequencing data (Tham et al., 2020). However, in the studies mentioned, a single instrument was used, namely, PacBio's SMRTtools (Alkanaq et al., 2019), Nanopolish (Orsini et al., 2018; Greig et al., 2019; Magdy et al., 2020), or Varscan (Wood et al., 2019); thus, virtually no guidelines for choosing an optimal caller have been proposed.

Considering the issues mentioned above, we characterized six popular algorithms in terms of specificity and accuracy when comparing them with Illumina reads. While it may seem evident that tools for calling on short reads would fail to succeed when applied to Nanopore data, Varscan was reported to be implemented for such purposes (Wood et al., 2019). However, in our study, quite expectedly, these instruments did not show adequate performance, which is most probably explained by non-robust behavior to high error-rate accompanying ONT sequences and stronger requirements for per-base quality gaining low coverage in calling regions due to discarding most of the reads during analysis (Suppl. 4e). Three other programs varied in terms of false-positive and false-negative calls generated. Longshot implies Pair-Hidden Markov Model calculating pairwise probabilities for each site to represent an SNV (Edge and Bansal, 2019). This approach improves detection in duplicated genomic regions; however, it cannot detect INDELs, and despite the highest precision, it lacks specificity (Fig. 1e, Suppl. 6). Clairvoyante, a multitask five-layer convolutional neural network model, was developed for predicting both INDELs and SNPs in er-

ror-prone long reads (Luo et al., 2019). While it demonstrated the highest sensitivity, it was not free from an extensively high rate of false-positive results (Suppl. 6). We propose that this observation could be explained by the fact that a pre-trained model was utilized as our data was insufficient to build a working model. In the current study, Nanopolish outperformed other calling tools, exhibiting the most appropriate trade-off between sensitivity and precision with the highest similarity to the Illumina reference. The program's underlying algorithm is HMM-based as in Longshot; however, instead of analyzing base pairs' distribution patterns, Nanopolish characterizes the intensity of signals and repots variants accordingly (Loman et al., 2015). This approach could, in fact, suffer from false-negative results (Orsini et al., 2018); nonetheless, Nanopolish applied to our data displayed high sensitivity (Fig. 1e). A possible explanation of these results could lie in high coverage. Nanopolish put into analysis the most considerable number of reads, gaining the highest coverage within calling regions (Suppl. 4e) and providing enough data for generating recognizable signals' patterns, which improved its variant calling performance.

When dissecting the generalized logistic regression model, we revealed that the two significant factors hampering precise calling on long reads are variant type and allele frequency (AF), which are, however, interconnected. More specifically, false-reported INDELs tend to have a lower AF as well, indicating that post-calling error-correction algorithms should primarily process these calls. It is also noteworthy that some variants were presented in all samples, and they referred to the European haplogroup (Patel et al., 2019). Interestingly, they have potential clinical significance influencing cisplatin anticancer therapy (rs2853518, rs2001030, rs2854128 (Patel et al., 2019)), increasing the risk of familial breast cancer (rs2853508 (Toncheva et al., 2020)), and possibly mediating neuropathological conditions (rs193302980 (Valentino et al., 2020)). As haplogroups of mitochondrial DNA are comprehensively classified, a possible approach for benchmarking variant calling performance could implement tracking the presence/absence of such haplogroup-related variants.

To sum up, despite a relatively small number of samples presented, our research adds another report to the growing body of evidence supporting the applicability of Nanopore reads in variant calling. Having characterized and compared the efficacy of several variant calling instruments, we recommend applying Nanopolish as finely balanced between precision and recall while detecting variants in the mitochondrial genome. However, our results are different from other studies in which other DNA sources were examined; thus, our recommendations, despite being relevant to mitochondrial genetics, cannot be generalized on a larger scale. This observation stresses the need to develop guidelines and best practices for long reads examination similar to what is formulated for short reads-based clinical analysis.

References

Aganezov, S., Goodwin, S., Sherman, R. M., Sedlazeck, F. J., Arun, G., Bhatia, S., Lee, I., Kirsche, M., Wappel, R., Kramer, M., Kostroff, K., Spector, D. L., Timp, W., McCombie, W. R., and Schatz, M. C. 2020. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Research* 30(9):1258–1273. https://doi.org/10.1101/gr.260497.119

Alkanaq, A. N., Hamanaka, K., Sekiguchi, F., Taguri, M., Takata, A., Miyake, N., Miyatake, S., Mizuguchi, T., and Matsumoto, N. 2019. Comparison of mitochondrial DNA variants detection using short- and long-read sequencing. *Journal of Human Genetics* 64(11):1107–1116. https://doi.org/10.1038/s10038-019-0654-9

Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research* 46(5):2159–2168. https://doi.org/10.1093/nar/gky066

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* 43(1110):11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Banoei, M. M., Houshmand, M., Panahi, M. S. S., Shariati, P., Rostami, M., Manshadi, M. D., and Majidizadeh, T. 2007. Huntington's disease and mitochondrial DNA deletions: event or regular mechanism for mutant huntingtin protein and CAG repeats expansion?! *Cellular and Molecular Neurobiology* 27(7):867–875. https://doi.org/10.1007/s10571-007-9206-5

Bansal, V. and Bafna, V. 2008. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 24(16):153–159. https://doi.org/10.1093/bioinformatics/btn298

Bates, D., Mächler, M., Bolker, B., and Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). https://doi.org/10.18637/jss.v067.i01

Bowden, R., Davies, R. W., Heger, A., Pagnamenta, A. T., de Cesare, M., Oikkonen, L. E., Parkes, D., Freeman, C., Dhalla, F., Patel, S. Y., Popitsch, N., Ip, C. L. C., Roberts, H. E., Salatino, S., Lockstone, H., Lunter, G., Taylor, J. C., Buck, D., Simpson, M. A., and Donnelly, P. 2019. Sequencing of human genomes with nanopore technology. *Nature Communications* 10(1):1–9. https://doi.org/10.1038/s41467-019-09637-5

Brandhagen, M. D., Just, R. S., and Irwin, J. A. 2020. Validation of NGS for mitochondrial DNA casework at the FBI Laboratory. *Forensic Science International: Genetics* 44:102151. https://doi.org/10.1016/j.fsigen.2019.102151

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6(2):80–92. https://doi.org/10.4161/fly.19695

Cornelis, S., Gansemans, Y., Vander Plaetsen, A. S., Weymaere, J., Willems, S., Deforce, D., and Van Nieuwerburgh, F. 2019. Forensic tri-allelic SNP genotyping

using nanopore sequencing. *Forensic Science International: Genetics* 38:204–210. https://doi.org/10.1016/j.fsigen.2018.11.012

Coxhead, J., Kurzawa-Akanbi, M., Hussain, R., Pyle, A., Chinnery, P., and Hudson, G. 2016. Somatic mtDNA variation is an important component of Parkinson's disease. *Neurobiology of Aging* 38:217.e1–217.e6. https://doi.org/10.1016/j.neurobiolaging.2015.10.036

Cumbo, C., Minervini, C. F., Orsini, P., Anelli, L., Zagaria, A., Minervini, A., Coccaro, N., Impera, L., Tota, G., Parciante, E., Conserva, M. R., Spinelli, O., Rambaldi, A., Specchia, G., and Albano, F. 2019. Nanopore targeted sequencing for rapid gene mutations detection in acute myeloid leukemia. *Genes* 10(12):1026. https://doi.org/10.3390/genes10121026

Dashti, M., Alsaleh, H., Eaaswarkhanth, M., John, S. E., Nizam, R., Melhem, M., Hebbar, P., Sharma, P., Al-Mulla, F., and Thanaraj, T. A. 2021. Delineation of mitochondrial DNA variants from exome sequencing data and association of haplogroups with obesity in Kuwait. *Frontiers in Genetics* 12:626260. https://doi.org/10.3389/fgene.2021.626260

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43(5):491–498. https://doi.org/10.1038/ng.806

Dhorne-Pollet, S., Barrey, E., and Pollet, N. 2020. A new method for long-read sequencing of animal mitochondrial genomes: application to the identification of equine mitochondrial DNA variants. *BMC Genomics* 21(1):785. https://doi.org/10.1186/s12864-020-07183-9

Edge, P. and Bansal, V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications* 10(1):564443. https://doi.org/10.1038/s41467-019-12493-y

Franco-Sierra, N. D. and Díaz-Nieto, J. F. 2020. Rapid mitochondrial genome sequencing based on Oxford Nanopore Sequencing and a proxy for vertebrate species identification. *Ecology and Evolution* 10(7):3544–3560. https://doi.org/https://doi.org/10.1002/ece3.6151

Goodwin, S., McPherson, J. D., and McCombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17(6):333–351. https://doi.org/10.1038/nrg.2016.49

Greig, D. R., Jenkins, C., Gharbia, S., and Dallman, T. J. 2019. Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. *GigaScience* 8(8):1–12. https://doi.org/10.1093/gigascience/giz104

Koboldt, D. C. 2020. Best practices for variant calling in clinical sequencing. *Genome Medicine* 12(1):91. https://doi.org/10.1186/s13073-020-00791-w

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22(3):568–576. https://doi.org/10.1101/gr.129684.111

Lee, H.-C., Li, S.-H., Lin, J.-C., Wu, C.-C., Yeh, D.-C., and Wei, Y.-H. 2004. Somatic mutations in the D-loop and decrease in the copy number of mitochondrial DNA in human hepatocellular carcinoma. *Mutation Research* 547(1–2):71–78. https://doi.org/10.1016/j.mrfmmm.2003.12.011

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H. and Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Slone, J., Fei, L., and Huang, T. 2019. Mitochondrial DNA variants and common diseases: a mathematical model for the diversity of age-related mtDNA mutations. *Cells* 8(6):608. https://doi.org/10.3390/cells8060608

Loman, N. J., Quick, J., and Simpson J. T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* 12(8):733–735. https://doi.org/10.1038/nmeth.3444

Luo, R., Sedlazeck, F. J., Lam, T.-W., and Schatz, M. C. 2019. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nature Communications* 10(1):998. https://doi.org/10.1038/s41467-019-09025-z

Maestri, S., Maturo, M. G., Cosentino, E., Marcolungo, L., Iadarola, B., Fortunati, E., Rossato, M., and Delledonne, M. 2020. A long-read sequencing approach for direct haplotype phasing in clinical settings. *International Journal of Molecular Sciences* 21(23):9177. https://doi.org/10.3390/ijms21239177

Magdy, T., Kuo, H., and Burridge, P. W. 2020. Precise and cost-effective nanopore sequencing for post-GWAS fine-mapping and causal variant identification. *iScience* 23(4):100971. https://doi.org/10.1016/j.isci.2020.100971

Mannelli, M., Rapizzi, E., Fucci, R., Canu, L., Ercolino, T., Luconi, M., and Young, W. F. J. 2015. 15 YEARS OF PARAGANGLIOMA: Metabolism and pheochromocytoma/paraganglioma. *Endocrine-Related Cancer* 22(4):T83–T90. https://doi.org/10.1530/ERC-15-0215

Masutani, B., Arimura, S., and Morishita, S. 2021. Investigating the mitochondrial genomic landscape of *Arabidopsis thaliana* by long-read sequencing. *PLOS Computational Biology* 17(1):e1008597. https://doi.org/10.1371/journal.pcbi.1008597

Menzel, P., Frellsen, J., Plass, M., Rasmussen, S. H., and Krogh, A. 2013. On the accuracy of short read mapping. *Methods in Molecular Biology* 1038:39–59. https://doi.org/10.1007/978-1-62703-514-9_3

Naing, A., Kenchaiah, M., Krishnan, B., Mir, F., Charnley, A., Egan, C., and Bano, G. 2014. Maternally inherited diabetes and deafness (MIDD): diagnosis and management. *Journal of Diabetes and its Complications* 28(4):542–546. https://doi.org/10.1016/j.jdiacomp.2014.03.006

Onyango, I. G., Dennis, J., and Khan, S. M. 2016. Mitochondrial dysfunction in Alzheimer's disease and the rationale for bioenergetics based therapies. *Aging and Disease* 7(2):201–214. https://doi.org/10.14336/AD. 2015.1007

Orsini, P., Minervini, C. F., Cumbo, C., Anelli, L., Zagaria, A., Minervini, A., Coccaro, N., Tota, G., Casieri, P., Impera, L., Parciante, E., Brunetti, C., Giordano, A., Specchia, G., and Albano, F. 2018. Design and MinION testing of a nanopore targeted gene sequencing panel for chronic lymphocytic leukemia. *Scientific Reports* 8(1):1–10. https://doi.org/10.1038/s41598-018-30330-y

Patel, T. H., Norman, L., Chang, S., Abedi, S., Liu, C., Chwa, M., Atilano, S. R., Thaker, K., Lu, S., Jazwinski, S. M., Miceli, M. V., Udar, N., Bota, D., and Kenney, M. C. 2019. Eu-

ropean mtDNA variants are associated with differential responses to cisplatin, an anticancer drug: implications for drug resistance and side effects . *Frontiers in Oncology* 9:640. https://doi.org/10.3389/fonc.2019.00640

Peng, H. and Lu, Y. 2012. Model selection in linear mixed effect models. *Journal of Multivariate Analysis* 109:109–129. https://doi.org/10.1016/j.jmva.2012.02.005

Popitsch, N., Preuner, S., and Lion, T. 2020. Nanopanel2 calls phased low-frequency variants in Nanopore panel sequencing data. *bioRxiv* 2020.11.06.370858. https://doi.org/10.1101/2020.11.06.370858

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., and Depristo, M. A. 2018. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology* 36(10):983. https://doi.org/10.1038/nbt.4235

Purevsuren, J., Fukao, T., Hasegawa, Y., Kobayashi, H., Li, H., Mushimoto, Y., Fukuda, S., and Yamaguchi, S. 2009. Clinical and molecular aspects of Japanese patients with mitochondrial trifunctional protein deficiency. *Molecular Genetics and Metabolism* 98(4):372–377. https://doi.org/10.1016/j.ymgme.2009.07.011

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341. https://doi.org/10.1186/1471-2164-13-341

Di Resta, C. and Ferrari, M. 2018. Next generation sequencing: from research area to clinical practice. *EJIFCC* 29(3):215–220.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12(1):77. https://doi.org/10.1186/1471-2105-12-77

Simon, D. K., Pulst, S. M., Sutton, J. P., Browne, S. E., Beal, M. F., and Johns, D. R. 1999. Familial multisystem degeneration with parkinsonism associated with the 11778 mitochondrial DNA mutation. *Neurology* 53(8):1787–1793. https://doi.org/10.1212/wnl.53.8.1787

Slatko, B., Gardner, A., and Ausubel, F. 2018. Overview of next generation sequencing technologies. *Current Protocols in Molecular Biology* 122(1):1–15. https://doi.org/doi:10.1002/cpmb.59

Suzuki, A., Suzuki, M., Mizushima-Sugano, J., Frith, M. C., Makałowski, W., Kohno, T., Sugano, S., Tsuchihara, K., and Suzuki, Y. 2017. Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. *DNA Research* 24(6):585–596. https://doi.org/10.1093/dnares/dsx027

Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T. H., Wang, W., Ng, C. H., Chng, W. J., Thiery, A., Tenen, D. G., and Benoukraf, T. 2020. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biology* 21(1):56. https://doi.org/10.1186/s13059-020-01968-7

Toncheva, D., Serbezov, D., Karachanak-Yankova, S., and Nesheva, D. 2020. Ancient mitochondrial DNA pathogenic variants putatively associated with mitochondrial disease. *PloS ONE* 15(9):e0233666–e0233666. https://doi.org/10.1371/journal.pone.0233666

Tranah, G. J., Nalls, M. A., Katzman, S. M., Yokoyama, J. S., Lam, E. T., Zhao, Y., Mooney, S., Thomas, F., Newman, A. B., Liu, Y., Cummings, S. R., Harris, T. B., and Yaffe, K. 2012. Mitochondrial DNA sequence variation associated with dementia and cognitive function in the elderly. *Journal of Alzheimer's disease* 32(2):357–372. https://doi.org/10.3233/JAD-2012-120466

Valentino, R. R., Tamvaka, N., Heckman, M. G., Johnson, P. W., Soto-Beasley, A. I., Walton, R. L., Koga, S., Uitti, R. J., Wszolek, Z. K., Dickson, D. W., and Ross, O. A. 2020. Associations of mitochondrial genomic variation with corticobasal degeneration, progressive supranuclear palsy, and neuropathological tau measures. *Acta Neuropathologica Communications* 8(1):162. https://doi.org/10.1186/s40478-020-01035-z

Watson, E., Davis, R., and Sue, C. M. 2020. New diagnostic pathways for mitochondrial disease. *Journal of Translational Genetics and Genomics* 4(3):188–202. https://doi.org/10.20517/jtgg.2020.31

Wickham, H. 2009. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. https://doi.org/10.1007/978-0-387-98141-3

Wood, E., Parker, M. D., Dunning, M. J., Hesketh, S., Wang, D., Pink, R., and Fratter, C. 2019. Clinical long-read sequencing of the human mitochondrial genome for mitochondrial disease diagnostics. *bioRxiv* 597187. https://doi.org/10.1101/597187

Yao, Y., Nishimura, M., Murayama, K., Kuranobu, N., Tojo, S., Beppu, M., Ishige, T., Itoga, S., Tsuchida, S., Mori, M., Takayanagi, M., Yokoyama, M., Yamagata, K., Kishita, Y., Okazaki, Y., Nomura, F., Matsushita, K., and Tanaka, T. 2019. A simple method for sequencing the whole human mitochondrial genome directly from samples and its application to genetic testing. *Scientific Reports* 9(1):17411. https://doi.org/10.1038/s41598-019-53449-y

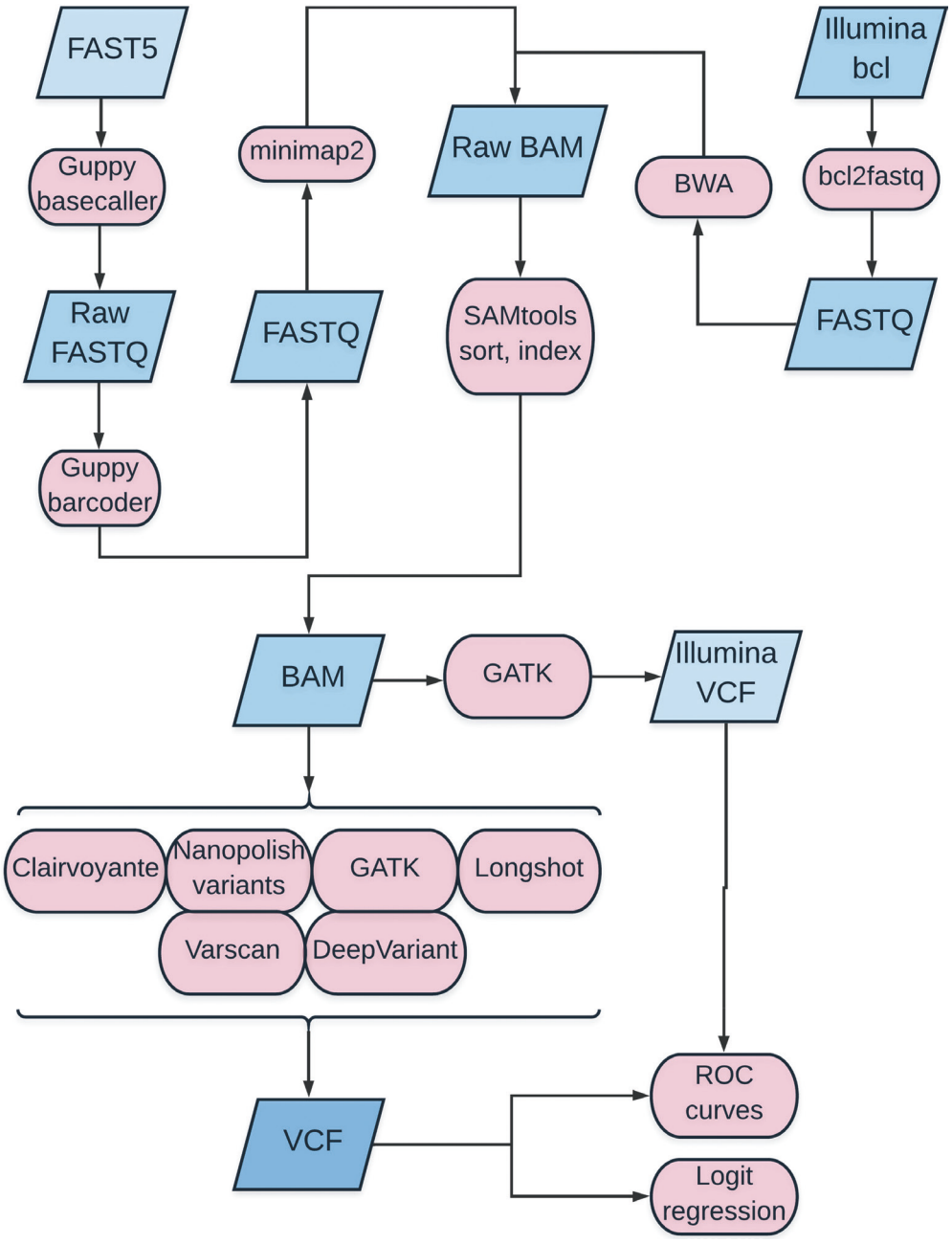
Zascavage, R. R., Thorson, K., and Planz, J. V. 2019. Nanopore sequencing: An enrichment-free alternative to mitochondrial DNA sequencing. *Electrophoresis* 40(2):272–280. https://doi.org/10.1002/elps.201800083

Zhou, K., Mo, Q., Guo, S., Liu, Y., Yin, C., Ji, X., Guo, X., and Xing, J. 2020. A novel next-generation sequencing-based approach for concurrent detection of mitochondrial DNA Copy number and mutation. *The Journal of Molecular Diagnostics* 22(12):1408–1418. https://doi.org/10.1016/j.jmoldx.2020.09.005

SUPPLEMENTS

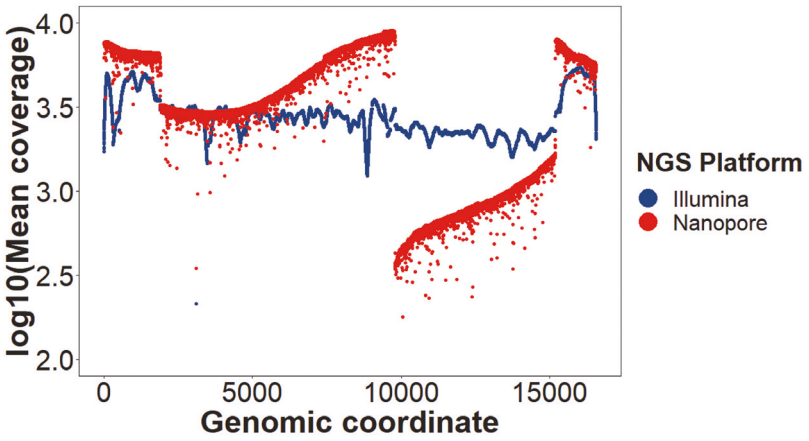
Supplementary 1

The overview of the pipeline for variant calling. Raw ONT FAST5 files were processed with Guppy basecaller followed by Guppy barcoder. After alining reads with minimap2 and sorting BAM files using SAMtools, six variant callers were applied. Raw Illumina signals in bcl format were basecalled using bcl2fastq utility and aligned with BWA, and the variant calling results were obtained with GATK. Illumina- and ONT-generated VCF files were then compared with ROC curves and logistic regression with a mixed linear model applied.



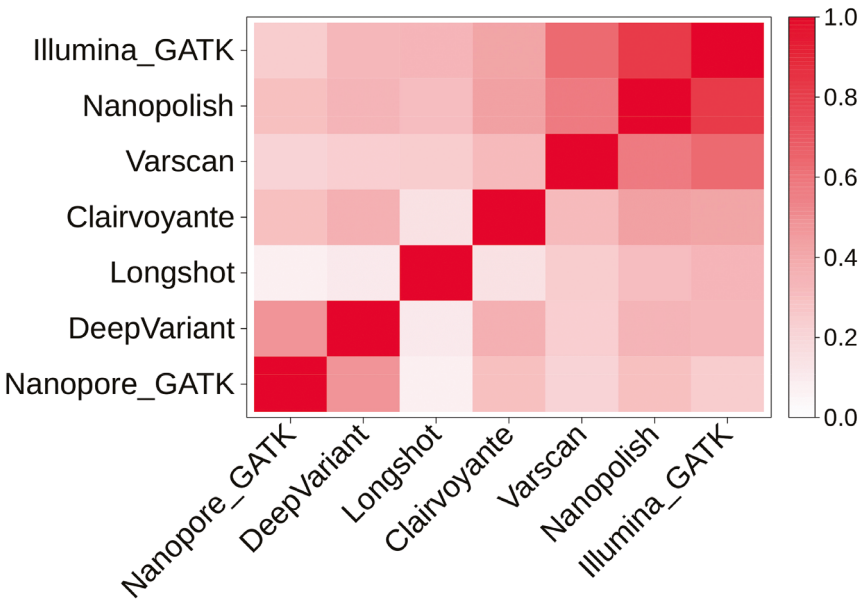
Supplementary 2

The per-site sequencing depth distribution at the logarithmic scale. The red color denoteы the ONT platform, while the blue color stands for the Illumina reads.



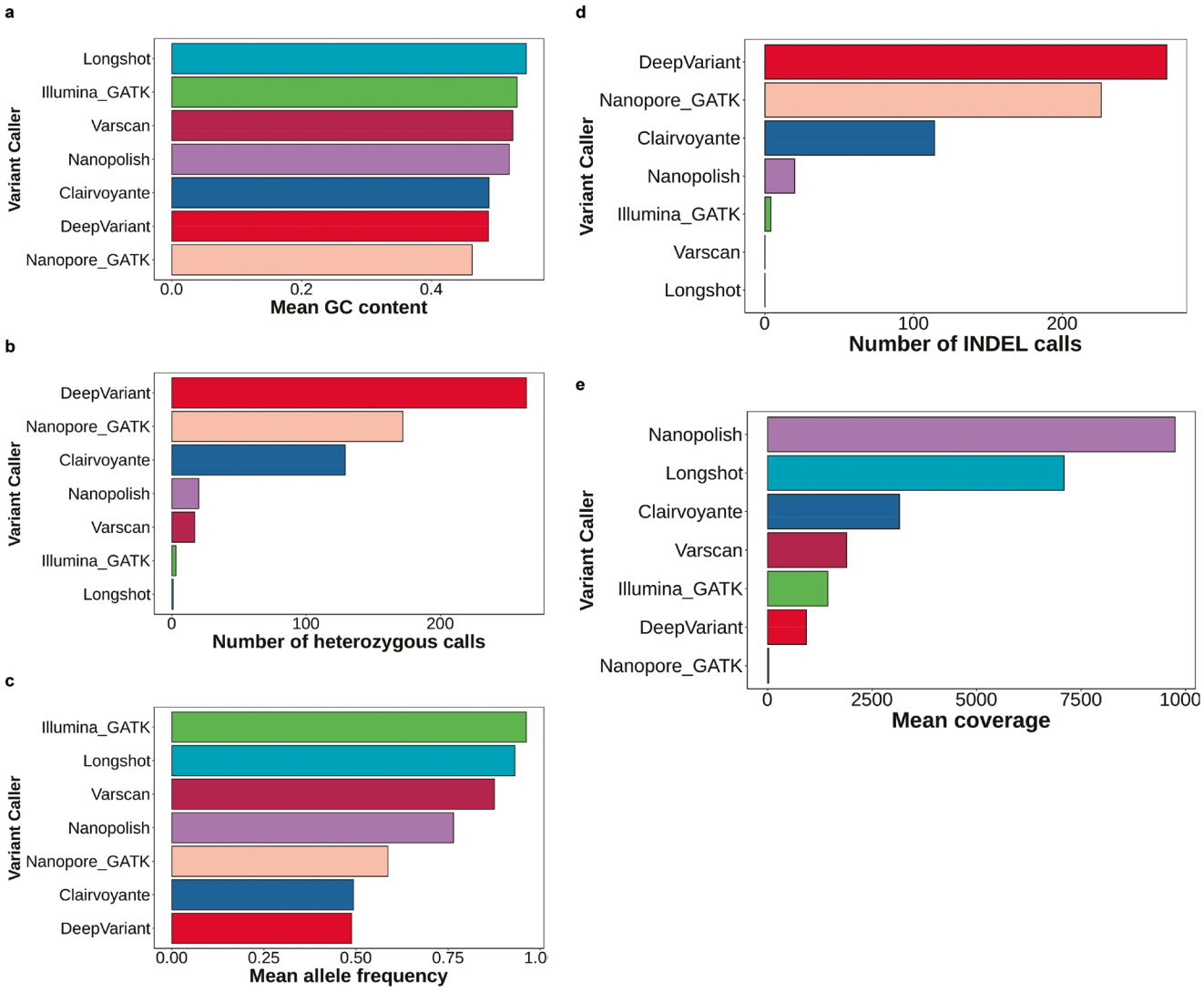
Supplementary 3

The heatmap presented summarizes a pairwise comparison of variant callers. The intensity of the color is proportional to the Jaccard coefficient indicating the similarity of the outputs.



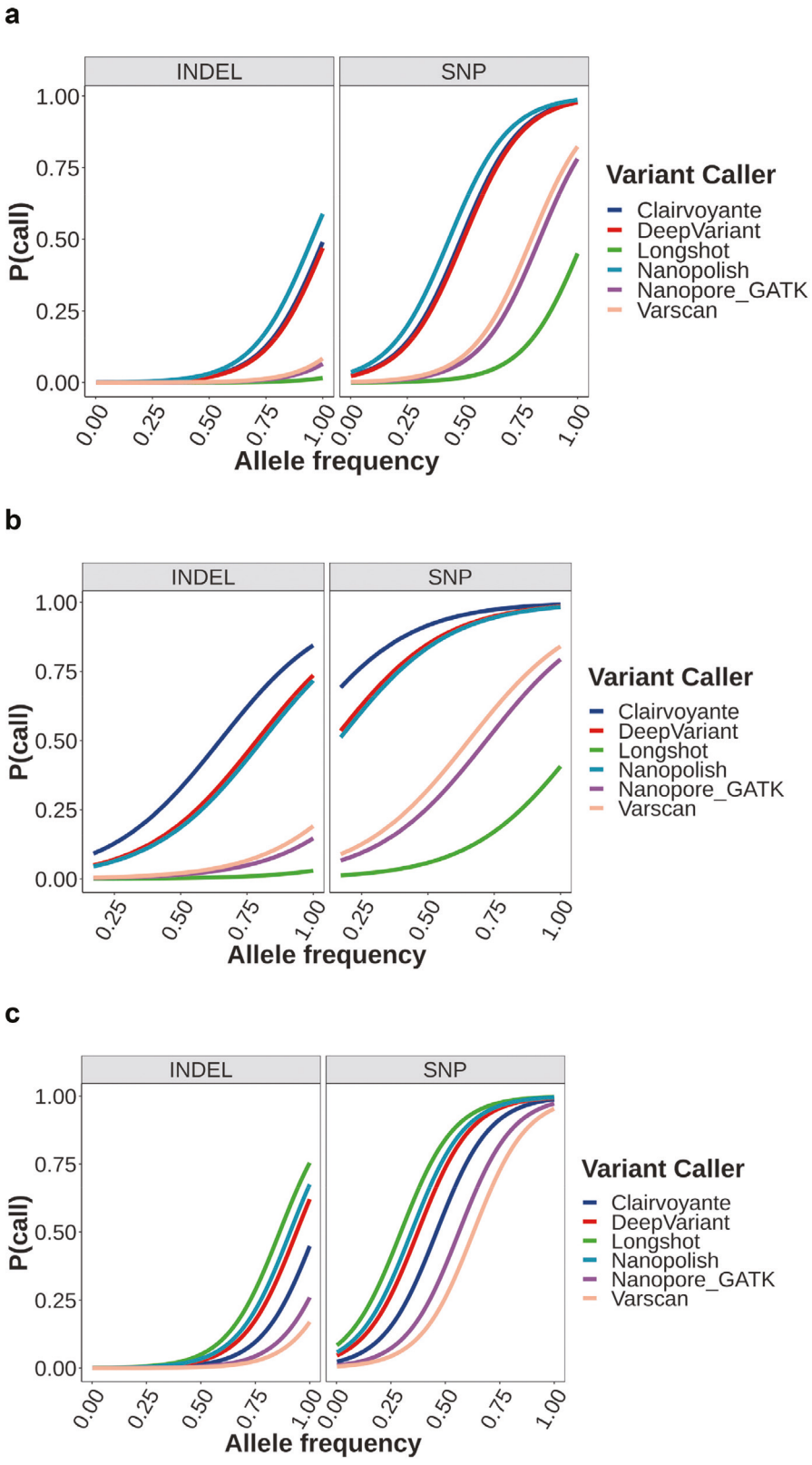
Supplementary 4

Instrument-wise statistics of the variants identified. **(a)** Mean GC-content for sites proportional to their allele frequency. **(b)** The cumulative sum of heterozygous polymorphic sites. **(c)** Mean allele frequencies of alternative alleles. **(d)** Total number of INDELs for each pipeline. **(d)** Mean per-site coverage.



Supplementary 5

Logit regression depicting the consistency with the Illumina data for variants callers applied at long reads for all sites (**a**), false-negative calls (**b**), and false-positive calls (**c**).



Supplementary 6

Logit regression depicting the consistency with the Illumina data for variants callers applied at long reads, including the Illumina statistics in the analysis, for all sites (**a**), false-positive calls (**b**), and false-negative calls (**c**).