Journal of Human Genetics (2019) 64:1107–1116
https://doi.org/10.1038/s10038-019-0654-9

ARTICLE

Comparison of mitochondrial DNA variants detection using short- and long-read sequencing

Ahmed N. Alkanaq[1] · Kohei Hamanaka[1] · Futoshi Sekiguchi[1] · Masataka Taguri[2] · Atsushi Takata[1] · Noriko Miyake[1] · Satoko Miyatake[1] · Takeshi Mizuguchi[1] · Naomichi Matsumoto[1]

Received: 19 May 2019 / Revised: 29 July 2019 / Accepted: 4 August 2019 / Published online: 13 August 2019

Abstract

The recent advent of long-read sequencing technologies is expected to provide reasonable answers to genetic challenges unresolvable by short-read sequencing, primarily the inability to accurately study structural variations, copy number variations, and homologous repeats in complex parts of the genome. However, long-read sequencing comes along with higher rates of random short deletions and insertions, and single nucleotide errors. The relatively higher sequencing accuracy of short-read sequencing has kept it as the first choice of screening for single nucleotide variants and short deletions and insertions. Albeit, short-read sequencing still suffers from systematic errors that tend to occur at specific positions where a high depth of reads is not always capable to correct for these errors. In this study, we compared the genotyping of mitochondrial DNA variants in three samples using PacBio's Sequel (Pacific Biosciences Inc., Menlo Park, CA, USA) long-read sequencing and illumina's HiSeqX10 (illumine Inc., San Diego, CA, USA) short-read sequencing data. We concluded that, despite the differences in the type and frequency of errors in the long-reads sequencing, its accuracy is still comparable to that of short-reads for genotyping short nuclear variants; due to the randomness of errors in long reads, a lower coverage, around 37 reads, can be sufficient to correct for these random errors.

Introduction

The last decade represents an unprecedented era of genetic research; tremendous amounts of genomic data are being generated, together with a parallel development in computational power and advancements of bioinformatics algorithms to decipher genomic patterns in these data. The time- and cost-effectiveness of the next-generation sequence-by-synthesis technology, made it the most widely used sequencing technology in research and clinical investigations. Most of today's bioinformatics algorithms are, therefore, designed to analyze short-read sequencing data.

Following successful genetic discoveries using short-read sequencing, the research community started to face new obstacles due to the inability of short-read sequencing to effectively resolve specific characteristics of the human genome. These obstacles can be summed into: (a) inability of short-reads to accurately map onto complex parts of the genome [1, 2], (b) the need for very complex algorithms, which in turn require expensive computational power, to accurately identify structural variants (SVs), (c) despite all the advancements in bioinformatics, some quantitative analyses like copy number variations (CNVs), are still hard to be accurately identified and assessed using short reads. The fact that parts of the human genome are still yet to be fully constructed in the reference genome, is another representation of the need for longer sequences to understand the complexity of genomic sequences.

The emergence of Next-Next-Generation sequencing technologies by PacBio (Pacific Biosciences Inc., Menlo Park, CA, USA) single molecule real-time (SMRT) technology [3], and Oxford Nanopore (Oxford Nanopore

Supplementary information The online version of this article (https://doi.org/10.1038/s10038-019-0654-9) contains supplementary material, which is available to authorized users.

✉ Naomichi Matsumoto
naomat@yokohama-cu.ac.jp

[1] Department of Human Genetics, Yokohama City University Graduate School of Medicine, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan

[2] Department of Data Science, Yokohama City University School of Data Science, 22-2 Seto, Kanazawa-ku, Yokohama 236-0027, Japan

Technologies Ltd., Oxford Science Park, Oxford, UK) long sequencing technologies [4], brought new opportunities for genetic researchers to overcome the shortcomings of short-read sequencing. However, these long-read sequencing technologies still have their own limitations, represented mainly by inaccuracies at the base-by-base level. These errors are mainly due to low signal to the noise ratio [5]; in addition, studies showed that the single nucleotide errors of SMRT long-read sequencing can be partially attributed to base substitution errors of polymerase enzyme [6, 7] with random distribution across long reads. Short insertion and deletion errors (indels), represent the majority of SMRT errors, with a tendency to occur around homopolymer regions and can also be the result of polymerization slow-down around non B-form DNA conformations, like G-quadruplexes [8]. Nonetheless, these errors are still random in nature and as the number of polymerization passes increases, the resulting consensus sequence accuracy would increase; this is being exploited by the circular consensus sequencing (CCS) [9] and the very recently developed high fidelity (HiFi) sequences [10].

The higher error rate of long-read sequencing, in comparison to the short-read one [5], has led scientists to resort to long-read only when trying to fathom genetic research muddles that involve a complex part of the genome, or structurally challenging for short-read to handle efficiently. At the same time, the short-read sequencing is still the technology of choice for identifying single nucleotide variants (SNVs) and short insertions and deletions (indels). Therefore, bioinformaticians started to find ways of hybridizing the results of both short- and long-read data, to get reliable genomic sequences by exploiting the lower error rate of short-reads in combination with the length of long-reads, which are long enough to accurately map to complex parts of the reference genome, to identify SVs, CNVs and repetitive regions.

Several studies tried to measure error rates of short-read sequencing. A study by Nakamura et al. [11] was among the first to describe specific systematic errors produced by illumina sequencers. Despite the following development in illumina's technologies, short-read sequences, nonetheless, are still suffering from systematic errors unequivocally associated with specific base-sequences. Pfeiffer et al. [12] performed a systematic evaluation of error rates, and they determined the error rate to be 0.24 ± 0.06% per base and 6.4 ± 1.24% of the reads are mutated for illumina's short-read sequencing technology.

Nanopore sequencing errors were shown to have some systematic patterns and less random than PacBio's sequencing errors [5], however, despite the higher frequency of errors in long reads, the extended length of the PacBio's and Nanopore's reads still provide more randomness of errors-per-read in comparison to short reads.

In this study, we compare SNV detection in three cases that have had whole-genome sequencing using both, illumina's short-read sequencing, and PacBio's SMRT sequencing technologies. The comparison was done using genotyping of the mitochondrial DNA (mtDNA) rather than the nuclear one for the following reasons: (a) compared to nuclear DNA, the number of mtDNA copies inside a cell is tremendously high, exceeding the nuclear one by thousands of folds in some cells, therefore, it naturally provides higher depth of coverage for any sample's whole-genome sequencing which is necessary for accurately comparing variant allele frequencies (VAF) in reads; (b) haploid-phasing of variants in nuclear DNA is necessary for obtaining a higher recall rate, as has been described in multiple studies [13–16]; therefore, being a haploid DNA, mtDNA makes variants identification more comparable between short and long sequences; (c) mtDNA is a very short sequence of DNA compared to any nuclear chromosome, therefore, its reassembly against the reference is more accurate compared to nuclear DNA, for identifying baseline heteroplasmy fractions.

Synthetic long reads, generated by technologies like 10X Genomics' bar coding (Pleasenton, CA, USA) [17] can provide chromosomal reads that are exponentially longer than mtDNA with high confidence of being from the same DNA fragment. However, this study aims to make a direct comparison between the standard output of long-read sequencing and short-read sequencing technologies, without resorting to costly and sophisticated technologies in avoiding the haplotype mix-up.

Materials and methods

Sample selection

Three samples of unrelated individuals were selected in our laboratory, where both short- and long-read whole-genome sequencing analyses were done for each sample. Sample-1 is of an 8-year-old female who was diagnosed with Krabbe disease (OMIM# 245200), she has beta-galactocerebrosidase deficiency and a heterozygous mutation. No mitochondrial variants were found to be responsible for her clinical diagnosis. Sample-2 is of a 40-year-old female who was diagnosed with benign adult familial myoclonus epilepsy (BAFME) (OMIM# 601068), she is referred to as individual [**III 2**] in a BAFME family that was studied by Mizuguchi et al. [18]. Sample-3 is of a 31-year-old male patient with definite hereditary hemorrhagic telangiectasia (HHT) (OMIM# 187300), based on the Curaçao's diagnostic criteria [19]. His three-generation family history is suggestive of autosomal inheritance of HHT and no mitochondrial variants were found to be possibly linked to his HHT diagnosis.

Long-read library preparation

Genomic DNA of the three samples was extracted from peripheral blood leukocytes using QuickGene (Kurabo) for samples 1 and 3, and standard phenol-chloroform for sample 2. DNA size and integrity were assessed using pulse-field agarose gel electrophoresis, followed by DNA concentration measurement using Qubit fluorometer (Life Technologies). Fragmentation, using g-TUBE (covaris) and 1500 × *g* centrifugation, was done before purifying the fragmented DNA by AMpure PB magnetic beads (Beckman Coulter).

Five micrograms of each sample's fragmented DNA was utilized for SMRTbell library reparation using SMRTbell Template Prep Kit 1.0 SPv3, Sequel Binding Kit 2.0, SMRTbell Clean-Up Column v2 Kit, and MagBead Kit v2 (Pacific Biosciences). Briefly, the resultant SMRTbell template was enriched for DNA fragments of >10 kb via BluePippin (Sage Science) size-selection. Purification of the size-selected was done using AMpure PB before performing DNA repair reaction. SMRTbell template DNA was annealed with Sequel Polymerase 2.0. The Clean-up Column kit was used to purify the SMRTbell template DNA/polymerase complex, before the diluting the purified complex to a concentration of 20 pM. Finally, the purified complex was mixed with MegaBead to produce MegaBeadbout SMRTbell complex which was loaded onto Sequel SMRT Cell 1M v2. A total of four cells were used for samples 2 and 3 and six cells were used for sample 1, with a data collection time of 6 h for each SMRT cell.

Short-read library preparation

Genomic DNA was extracted from peripheral blood lymphocytes. Using TruSeq DNA PCR-free library preparation kit, genomic DNA library was constructed before sequencing with illumina's HiSeqX10, using single index. Generated sequence data had an average of 32.8 million of 150 nucleotide-long paired-end reads for each sample.

Long-read mitochondrial DNA data analysis

For the purpose of comparison, we kept the consistency of analysis with that of short-read by performing the long-read analysis on whole genome single-pass subreads obtained from PacBio Sequel sequencer. PacBio's single-pass subreads were generated by obtaining long sequences from the SMRTbell templates, after the removal of adapter sequences. Each sample's subreads BAM file contains all the subreads generated from all cells used for a specific sample. In addition to the nucleotide-sequence information, a full set of quality and kinetic parameters are attached to each subread; therefore, for a full utilization of these technology-specific data, the mapping and analysis were done using the standard software included in PacBio's SMRT tools v.6.0.0 (Pacific Biosciences).

Subreads produced by cells of each sample were aligned to mtDNA rCRS reference (NC_012920.1), using BLASR (v5.1) [20] with default mapping options.

Following alignment to the rCRS reference genome, the average N50 length of polymerase reads for the whole-genome data was 14,761 bp, and the average number of subreads per sample was 378, with an average length of 3906 bp (Supplementary Table 1). The average concordance of samples' data with the reference is 0.8261.

Short-read data analysis

The Short-read data analysis of mtDNA was done following best-practice guidelines of Genome Analysis Toolkit (GATK v.4.1) [21] (Broad Institute), since GATK is still regarded as the gold standard and one of the most widely used software toolkit for genotyping short-read data. In version 4.1 of GATK, the *Mutect2* tool, which was primarily designed to call somatic short nuclear variants using local assembly of haplotypes, has been revised to include the "mitochondria mode," where the LOD score is set to 0 for the capability to annotate possible nuclear mitochondrial sequences using Poisson distribution of the median autosomal coverage. Therefore, utilizing *Mutect2* can provide a robust detection of very low fractions of mitochondrial variants after a statistical exclusion of "nuclear mitochondrial DNA segments" (NuMT) which represent transposed mitochondrial sequences in the nuclear DNA. In addition, *Mutect2* utilizes the original DREAM challenge-winning engine [22], together with the HaplotypeCaller machinery of local de novo reassembly. Therefore, *Mutect2* can provide high sensitivity in combination with specificity in calling variants of mtDNA.

Following the GATK best-practice guidelines, short-reads were mapped to GRCh38 genome reference that includes the revised Cambridge Reference mitochondrial Sequence (NC_012920.1) [23], using Burrows-Wheeler Alignment Tool (bwa v0.7.17-r1188) [24]. Since bwa aligner is not designed to evenly align circular reads of the mtDNA, the alignment process included two branches where the second branch was aligned to the mtDNA reference that was shifted by 8000 nucleotides. Following alignment, the resultant two BAM files for each sample were passed through a pipeline of Genomic Analysis Toolkit (GATK v.4) [21] tools that included duplicate reads marking, local indel realignment, and quality scores recalibration, before being genotyped using *Mutect2* in mitochondrial mode.

Long-read data analysis

The genotyping of long-read mapped data were done using variantCaller tool (v2.2.2) of PacBio's SMRTtools (v6.0).

variantCaller is provided by PacBio in the GenomicConsensus package; when it runs in default settings, as we have done for our samples, it utilizes the Arrow consensus model for variant calling against the reference. Arrow algorithm is an improved model of Quiver [25] which is based on hidden Markov principle that utilizes the consensus data of long reads to filter out random errors.

Haplogroup assignment

The haplogroup assignment for each sample was done using mitolib v.0.1.2 software (https://github.com/haansi/mitolib) integrated into the contamination analysis step of GATK 4.1 tools [21]. The mitolib's haplockecker checks for mtDNA contamination using Phylotree 17 and assigns the most probable haplotype for each mtDNA short-read BAM file.

Sanger sequencing

A total of seven discrepant variants between both, short- and long-read sequencing analyses, were chosen for Sanger sequencing. All variants that have VAF below 0.1, except for one variant in sample-1 that has a borderline VAF of 0.096, were excluded from Sanger sequencing confirmation. Sequences of primers used are available upon request.

Following standard PCR amplification, and capillary electrophoresis using ABI 3130xl, we had to modify the PCR protocol to accommodate for a GC-rich region.

Tagging variants

Several studies concluded the high likelihood for illumina's short-read variants with VAF below 1% to be erroneous [26, 27]; in fact, most of these reads in our study had multi heteroplasmic variants which were annotated by *Mutect2* as "chimeric original alignment" or as "strand artifacts". It is also important to mention that a number of studies describe specific mtDNA variants associated with illumina's short-read sequencing as sequencing artifacts [26, 28]. Generally speaking, there are two main sources of errors: (a) technology-specific systematic errors, as with variants flanked by low-complexity regions [29]; (b) bioinformatics errors, as in the miscalling of variants around the "N" placeholder at 3107 position of the rCRS reference. Therefore, for an unbiased comparison between the long- and short-read sequencing technologies, we tagged all variants that belong to any of the aforementioned categories as likely erroneous variants. These variants lie in the following positions of mtDNA, where it is characterized by low complexity sequences or the place holder: (301, 302, 310, 316, 3107, and 16182–16192).

According to a comprehensive study by Spencer et al. [30]. In order to accurately detect variants with VAFs less than 0.01, specialized library preparation methods are required; otherwise, variants called using common methods with VAFs less than 1% are very likely to be erroneous. Therefore, following the short-read genotyping, variants with VAF less than 0.01 were removed before proceeding with further analyses.

Statistical analysis for interrater reliability

Due to the small number of samples being analyzed, and the necessity to mask tagged variants, the utilization of standard statistical analyses, like *t*-test and *chi*-square analyses, becomes inapplicable. We can assume, however, that long- and short-read genotyping analyses are two raters for each sample where: (a) each sample's mtDNA result is independent from other sample's, since they are not related individuals; (b) the probability for each position in a sample's mtDNA to be mutated is independent from other positions', and it can be either identical to the reference or not, with no preference for the rater to report each position to be identical to the reference or not, in other words, each position in a sample's DNA has a mutually exclusive probability of being mutated or not, and being identical to the reference or not is independent among different positions; (c) the raters, short- and long-read technologies, are operating independently from each other.

According to the proposed conditions by Jacob Cohen in 1960 [31] weighted Cohen's kappa statistics represents the best statistical model for comparing the agreement between long- and short-read data analyses. Since both technologies cannot provide accurate genotyping at the masked positions, it became necessary to adopt weighted Kappa's statistical analysis, since in these masked regions, no technology was proven to be superior to another, and giving these masked regions' position a third status of 'unknown' can safely evaluate the agreement between the two raters by providing low weight to these masked regions and reduce the effect of their obscurity on the analysis. If we were to choose another more familiar statistical measurements like *t*-test, we would need larger number of cases in order to have a statistically significant analysis of the agreement between the two tests.

The formula for Cohen's kappa [31] calculation is:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

Where *P(a)* is the accuracy, or the actual agreement between the two raters, and *P(e)* is the estimated or hypothetical probability of agreement between the two raters.

In our study, we implemented the quadratic weighted Cohen's kappa [32] calculation, which takes the three possibilities at each position: reference, mutated, unknown, as independent variables. Therefore, the disagreement between the two raters are treated equally, with the different levels of agreements contribute to the value of kappa.

The formula for our quadratic weights will therefore be:

$$w_i = 1 - \frac{(LongRead\ value - ShortRead\ value)^2}{(Total\ number\ of\ categories - 1)^2}$$

Since we have a total of three categories:

$$w_i = 1 - \frac{(LongRead\ value - ShortRead\ value)^2}{4}$$

Where w_i is the weighted agreement score at position i. Values are 1 for reference, 2 for mutation, and 3 for unknown.

Additionally, since we are analyzing the data for three samples only, Cohen's weighted kappa scoring reduces bias through: (a) the consideration of each position of the mtDNA as a separate experiment, for both technologies to analyze, providing a stronger statistical power for the calculation of kappa coefficient; (b) the random errors of SMRT long-read sequencing, can be accounted for in the coefficient calculation for each position of the mtDNA; since we are doing the calculation in respect to the total number of possible values, which is three in our case; (c) the ability to include the masked regions' variants in the calculation, using a different weight of disagreement.

It's important to mention that the kappa coefficient is not a directly interpretable measure of agreement [33], but rather an indication of the level of agreement. Kappa coefficient values above 0.81 represent an almost perfect level of agreement with a reliability of data between 64 and 100% (see Supplementary Table 2 for description of all levels).

In calculating kappa coefficient for each sample, we used the standard weighted kappa coefficient tool of the specialized python library scikit-learn [34] by comparing long- and short-read analyzed data at each position of the 16569 mtDNA reference.

Results

Short-read mitochondrial DNA variants analysis

The average total number of variants for short reads is 39.3, with averages of 35.67 SNVs and 3.67 indel variants per-sample. For the called variants in each sample, the average maximum and minimum coverage values are 4984 and 1037, respectively (Table 1).

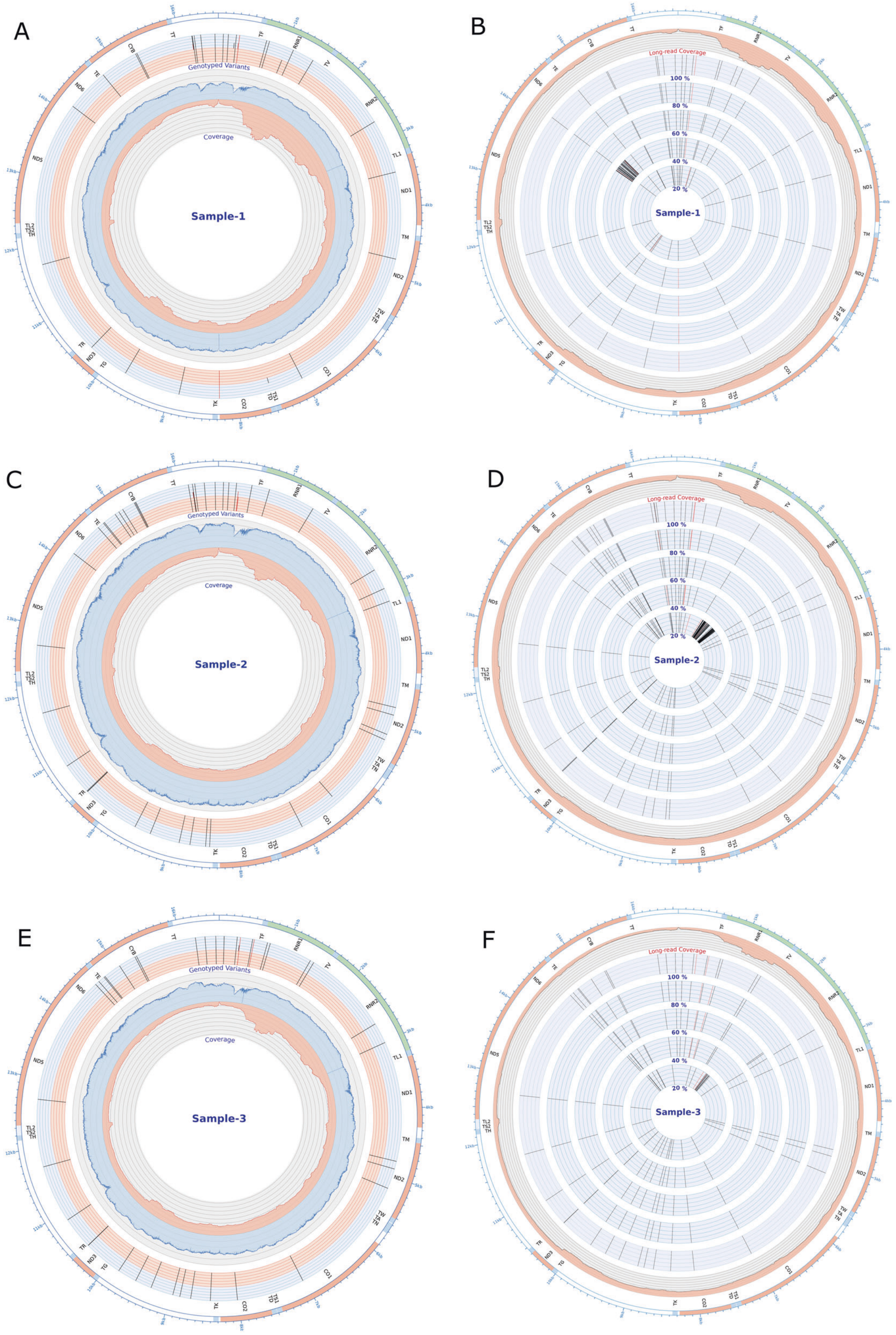
Long-read mitochondrial DNA variants analysis

The average total number of variants for long reads is 36.67, with averages of 34.3 SNVs and 2.3 indel variants per

Table 1 Summary statistics of short- and long-read sequencing genotyping for mitochondrial DNA of three samples

	Short reads							Long reads						
Sample	Total vars	SNV	Indels	Mean cvrg	Min DP	Max DP	Haplogroup	Total vars	SNV	Indels	Mean cvrg	Min DP	Max DP	Tagged
Sample-1	31	26	5	3097	881	4520	B4c1a1	26	24	2	118	51	311	5
Sample-2	48	44	4	4568	1638	5975	D4a1a1	46	43	3	106	55	185	9
Sample-3	39	37	2	3778	1404	4457	D4b2b1/D4b2	38	36	2	91	42	189	2
AVG	39.3	35.67	3.67	3814.33	1307.7	4984		36.67	34.3	2.3	105	49.3	228.3	5.3

Total vars: total number of variants per sample, **SNV:** number of single nuclear variants, **Indels:** number of insertions and deletions, **Mean cvrg:** the mean coverage for all called variants per sample, **Max/Min DP:** the maximum/minimum depth of coverage for variant positions per sample, **AF:** variant allele frequency, **AVG:** average values for all three samples



sample. For the called variants in each sample, the average maximum and minimum coverage values are 228.3 and 49.3, respectively (Table 1).

The average total number of variants per sample is very comparable between the two technologies (Fig. 1a, c, e), despite the disproportional difference of coverage.

◀ **Fig. 1** Circular plotting of genotyped variants and downsampling genotyping. Results of sample-1 (**a**, **b**), sample-2 (**c**, **d**) and sample-3 (**e**, **f**) are presented for circular plotting of genotyped variants (**a**, **c**, **e**) and downsampling genotyping results (**b**, **d**, **f**). **a**, **c** and **e**: *SNVs*: (black lines) and *indels* (red) are plotted in relation to the mitochondrial genes map. *Heteroplasmic short-read variants* (blue background) are shown as *short lines*, while all long-read variants are homoplasmic (orange background). Corresponding coverage for short reads (in blue) is plotted on a circular scale of 6000 reads, while the long reads coverage (light red) is plotted on a circular scale of 350. **b**, **d**, **f**: The first plot inside of the genes map represents the full coverage of long-reads. Each subsequent plot represents the genotyped variants at different levels of downsampling (100, 80, 60, 40, and 20%). Black lines represent SNVs and red lines represent indels. This figure was plotted using circus package [36]

However, four untagged variants genotyped using short-read analyses with VAFs ranging from 0.032 to 0.096, are not detected with the long-read analysis.

mtDNA haplogroups

Two samples were assigned a single haplotype (B4c1a1 for sample-1 and D4a1a1 for sample-2), while one sample (sample-3) was assigned both a major and a minor haplogroup (D4b2b1 at 98.6% and D4b2 at 87.3%) (Table 1). The two haplogroups assigned to sample-3 are not phylogenetically distant, therefore, it is very unlikely to be due to contamination.

Homoplasmy versus heteroplasmy

The high depth of reads of short-read sequence data, and its analysis using *Mutect2* tool from GATK 4.1 [21] which is specialized for detecting variants with high sensitivity at different VAFs, made it possible to reliably detect heteroplasmic variants. On the other hand, the substantially lower coverage of PacBio's long reads analyzed using Arrow algorithm did not yield any heteroplasmic variants.

In order to accurately compare the performance of the two technologies in identifying variants, including the heteroplasmic ones, it is necessary to mask variants in tagged regions, since most of these variants are artifacts, and therefore likely to present in heteroplasmic form (Supplementary Tables 3, 4 and 5).

The majority of heteroplasmic variants lie in the tagged regions, using Sanger sequencing to reliably verify heteroplasmic variants at these regions was not possible due to the low VAFs of these variants, and the long homoploymers in these regions.

Cohen's kappa coefficients

The weighted kappa coefficient for samples 1, 2, and 3 at full coverage are 0.908, 0.980, and 0.997, respectively (Table 2). Based on the standard interpretation of these values (Supplementary Table 2), these weighted kappa coefficients indicate that the levels of agreement between long- and short read mtDNA genotyping are “almost perfect” at full coverage. With 82.4%, 96%, and 99.4% reliability for samples 1, 2, and 3, respectively.

Sanger sequencing

All of the discrepant variants between two technologies we tried to confirm using Sanger sequencing were in tagged regions of low complexity. Other discrepant variants that are outside the tagged regions have VAFs below 0.1, which cannot be reliably confirmed using Sanger sequencing. However, we still tried to confirm one variant in sample-1 at position 240 which is of VAF of 0.096, but unfortunately the signal generated was unreliable to validate or reject it.

A GC-rich region, the tagged low complexity region 16181–16193, was re-sequenced using specific protocol for GC-rich regions for both sample-1 and sample-2, however, the obtained results still failed to provide clear validation of the results, despite using the special protocol.

Similar results were seen in the rest of the discrepant variants we tried to confirm using Sanger sequencing.

Long-read random downsampling and its effect on genotyped variants

To have a better understanding of the relationship between genotyping and the depth of reads of PacBio's long-read sequencing, random downsampling of the reads was performed on each of the three samples. By removing 20% of the reads successively, and compare the genotyping results of 20, 40, 60, 80, and 100% of the total coverage. Figure 1b, d and f show the variants genotyping at different coverages for each sample.

After random downsampling, weighted kappa-coefficient of agreement between variant callings at different coverage levels against short-read results of each sample was done (Supplemntry Tables 6, 7 and 8), following the same procedure that includes tagging variants in regions described in the methods section. Table 2 shows the calculated kappa-coefficients for the three samples.

Table 2 shows calculated kappa values at different levels of coverage following downsampling. The quadratic weighted kappa coefficient eliminates any residual chance of randomness when comparing the two analyses, therefore, when comparing the mean coverage at different downsampling levels for the three samples against the corresponding mean kappa value, we can see that at a mean coverage of 51 (for the 60% coverage level) the mean kappa value is 0.946; corresponding to an ‘almost perfect’ interpretation (Supplementary Table 2). Furthermore, the mean

Table 2 Calculated kappa-coefficient for the three samples at different coverage percentages

% coverage	Sample-1		Sample-2		Sample-3		Mean coverage	Mean kappa	% agreement
	coverage	kappa	coverage	kappa	coverage	kappa			
100%	95.688	0.908	95.608	0.980	89.051	0.997	93.449	0.961	92.432
80%	57.290	0.926	73.760	0.977	67.974	0.993	66.342	0.965	93.174
60%	46.452	0.926	60.420	0.974	46.659	0.939	51.177	0.946	89.554
40%	32.057	0.512	40.458	0.976	38.872	0.980	37.129	0.823	67.656
20%	14.611	0.747	30.745	0.447	15.250	0.643	20.202	0.612	37.505

The % coverage indicates the percentage of reads remaining following random downsampling, where 100% is the original coverage data. The mean coverage and mean kappa are calculated across the three samples

kappa value at mean coverage of 37 (for the 40% coverage level) is 0.823, which is interpreted as ‘strong’ level of agreement, indicating that 67.656% of the data are reliably agreeable and not due to chance.

However, these values widely fluctuate among the three samples, since each sample has different initial full-coverage value, and due to other sample-specific values related to sample preparation or experimental conditions.

Short-read downsampling and its effect on genotypes variants

To compare the effect of downsampling on short-read data with that of long-read data, downsampling was done for each sample at seven different depths of reads, 1000, 500, 100, 50, 30, 20, and 10×. Supplementary Tables 9, 10 and 11 show the allele frequency for each genotyped variant at different depths of coverage.

Results of the short-read downsampling show three main findings: (a) despite the dramatic reduction of read-coverage, the total number of variants remains highly consistent (Supplementary Fig. 1); (b) changes in the number of variants occurs mainly within the masked regions, confirming their liability for being erroneous; and (c) the proportion of total number of downsampling-associated erroneous variants is comparable to the proportion of total number of discrepant variants between long- and short-read sequencing, (see Supplementary Tables 3, 4 and 5).

Discussion

When compared to short-read technology, SNV and Indel genotyping of PacBio’s long-read data is considerably consistent. However, due to the limitation of resources only three cases were available, where both long- and short-read whole-genome sequencing were done, therefore, the total number of heteroplasmic variants was limited. Nevertheless, the downsampling process could still provide better understanding of the relationship between the accuracy of long-read genotyping and other parameters, including depth of coverage and other possible sample-specific factors like DNA quality and library preparation. The different, fluctuating kappa values at different coverages for different samples can be partially explained by the depth of coverage, as shown in Supplementary Tables 3, 4 and 5. The random downsampling confirms the following two facts: (1) Reducing the coverage does not necessarily lead to a corresponding reduction in the number of variants. On the contrary, due to the noise in long reads caused by random indel errors, the model starts to erroneously call false variants as the reduction goes below coverage of around 37 reads, (2) long-read random errors are responsible for generating the false variants at low coverage, unlike short-reads where the removal of fractions of the reads doesn’t lead to significant increase in false variants. It indicates that, as expected, the hidden Markov model of the Arrow algorithm requires a certain percentage of reads to accurately call variants, rather than predominantly rely on the majority of votes by reads at each position.

Therefore, coverage is very critical for the reliability of using single-pass long-read data for SNV genotyping.

An attempt to confirm discrepant variants at tagged regions using Sanger sequencing was unsuccessful, as described in the “Results” section, due to a combination of low complexity of genomic sequence at the discrepant sites, and technical limitations of Sanger sequencing when trying to confirm regions with variants of low VAFs.

In a study conducted on more than 1500 cases of the ClinSeq study (NHGRI, USA) by Beck, Biesecker, and others [35], they concluded that using Sanger sequencing as the gold standard for confirming NGS variants is not always a proper thing to do. In that study, over 5800 NGS variants were analyzed using Sanger sequencer, and in some cases Sanger sequencing can reject true positive variants instead of eliminating false positive ones.

Given the high agreement between the long- and short-read technologies, it indicates that using long-read sequencing for genotyping short variants, in addition to structural variants, might be a highly cost-effective choice. However, larger studies, using more samples can provide stronger evidence for or against these conclusions.

The high consistency of genotyped variants with the downsampling of short reads demonstrates the expected robustness and accuracy of short-read data, however, this does not exclude the possibility of persistent erroneous genotypes due to systematic errors. The comparable number of fluctuating erroneous variant numbers with downsampling to the total number of discrepant variants between short- and long-read sequencing, among the three samples, is possibly due to the effects of DNA quality, regardless to the sequencing technology.

Finally, although the recent development of HiFi consensus sequences by PacBio can provide more accurate sequences than the standard subreads, by performing multiple passes over the DNA segment, HiFi is still costly and the main scope of this study is to compare single-pass long- and short-read sequencing data.

Acknowledgements This work was supported by AMED under the grant numbers JP19ek0109280, JP19dm0107090, JP19ek0109301, JP19ek0109348, and JP18kk020501; JSPS KAKENHI grant numbers JP17H01539, JP16H05357, JP16H06254, JP17K10080, and JP 17K15630; the Takeda Science Foundation; and the Ichiro Kanehara Foundation for the Promotion of Medical Science and Medical Care. We also thank T. Miyama for their technical assistance.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using singlemolecule sequencing. Nature. 2015;517:608–U163.
2. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Hum Mol Genet. 2018;27 (R2):R234–41.
3. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133–8.
4. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338.
5. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res. 2017;6:100.
6. Hestand MS, Van Houdt J, Cristofoli F, Vermeesch JR. Polymerase specific error rates and profiles identified by single molecule sequencing. Mutat Res-Fundam Mol Mech Mutagen. 2016;784:39–45.
7. Potapov V, Ong JL. Examining sources of error in PCR by Single-molecule sequencing. PLoS ONE. 2017;12:e0169774.
8. Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovská I, Kejnovsky E, et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. 2018;28:1767–78.
9. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. bioRxiv, 2019:519025.
10. Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. bioRxiv, 2019:635037.
11. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39:e90.
12. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep. 2018;8:10950.
13. Guo XG, Lehner K, O'Connell K, Zhang J, Dave SS, Jinks-Robertson S. SMRT sequencing for parallel analysis of multiple targets and accurate SNP phasing. G3-genes genomes. Genetics. 2015;5:2801–8.
14. Ebler J, Haukness M, Pesout T, Marschall T, Paten B. Haplotype-aware genotyping from noisy long reads. Genome Biol. 2019;20:116.
15. Erik Garrison GM Haplotype-based variant detection from short-read sequencing. arXiv. 2012;arXiv:1207.3907.
16. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34:816–34.
17. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784.
18. Mizuguchi T, Toyota T, Adachi H, Miyake N, Matsumoto N, Miyatake S. Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. J Hum Genet. 2019;64:191–7.
19. Shovlin CL, Guttmacher AE, Buscarini E, Faughnan ME, Hyland RH, Westermann CJ, et al. Diagnostic criteria for hereditary hemorrhagic telangiectasia (Rendu-Osler-Weber syndrome). Am J Med Genet. 2000;91:66–7.
20. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. Bmc Bioinformatics. 2012;13:238.
21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
22. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.
23. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, et al. A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. Am J Hum Genet. 2012;90:675–84.
24. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
25. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10:563–9.

26. Kloss-Brandstatter A, Weissensteiner H, Erhart G, Schäfer G, Forer L, Schönherr S, et al. Validation of next-generation sequencing of entire mitochondrial genomes and the diversity of mitochondrial DNA Mutations in oral squamous cell carcinoma. PLoS ONE. 2015;10:e0135643.
27. Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, et al. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Res. 2016;44(W1):W64–9.
28. Li MK, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet. 2010;87:237–49.
29. Zhidkov I, Nagar T, Mishmar D, Rubin E. MitoBamAnnotator: a web-based tool for detecting and annotating heteroplasmy in human mitochondrial DNA sequences. Mitochondrion. 2011;11:924–8.
30. Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, et al. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J Mol Diagn. 2014;16:75–88.
31. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
32. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213–20.
33. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22:276–82.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
35. Beck TF, Mullikin JC. NISC Comparative Sequencing Program, Biesecker LG. Systematic evaluation of sanger validation of next-generation sequencing variants. Clin Chem. 2016;62:647–54.
36. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.