Briefings in Bioinformatics Advance Access published August 24, 2016

Briefings in Bioinformatics, 2016, 1–14

doi: 10.1093/bib/bbw077
Paper

Characterization of MinION nanopore data for resequencing analyses

Alberto Magi, Betti Giusti and Lorenzo Tattini

Corresponding author: Alberto Magi, Department of Experimental and Clinical Medicine, University of Florence, Largo Brambilla, 3-50134 Florence, Italy. Tel.: +39 055 7948909; FAX: +39 055 7948909; E-mail: albertomagi@gmail.com

Abstract

The Oxford Nanopore Technologies MinION is a new device, based on nanopore sequencing that is able to generate reads of tens of kilobases in length with faster sequencing time with respect to other platforms. To evaluate the capability of nanopore data to be exploited for resequencing analyses we used the largest MinION data set to date and we compared with Illumina and Pacific Biosciences technologies. By using five different mapping approaches we estimated that the global sequencing error rate of MinION reads, mainly caused by inserted and deleted bases, is around 11%. The study of error distribution showed that substituted, inserted and deleted bases are not randomly distributed along the reads, but mainly occur in specific nucleotide patterns, generating a significant number of genomic loci that can be misclassified as false-positive variants. With 40× sequencing coverage, MinION data can produce at best around one false substitution and insertion every 10–50 kb, and one false deletion every 1000 bp, making use of this technology still challenging for small-sized variant discovery. We also analyzed depth of coverage distribution and we demonstrated that nanopore sequencing is a uniform process that generates sequences randomly and independently without classical sources of bias such as GC-content and mappability. Owing to these properties, the MinION data can be readily used to detect genomic regions involved in copy number variants with high accuracy, outperforming other state-of-the-art sequencing methods in terms of both sensitivity and specificity.

Key words: nanopore; resequencing; third generation sequencing

Introduction

In the past decade, genomic science has been revolutionized by the advent of second-generation sequencing (SGS) technologies [1]. The first commercial SGS platforms emerged in 2005 to overcome the low throughput and the high cost of first-generation Sanger sequencing.

At present, the predominant SGS approach (Illumina) consists of sequencing a huge amount of DNA molecules in parallel by anchoring millions of small clusters of the same DNA fragment in a solid surface and read them in a process that consists of sequential washing and scanning operations. The wash-and-scan cyclic process [2] consists in incorporating fluorescence-labeled nucleotides in the DNA fragments, stopping the incorporation reaction, washing the excess reagent and scanning the solid surface to detect the incorporated bases by means of fluorescence emission. Thanks to these technologies, today a human genome can be sequenced quickly at affordable prices [3]. The emergence of these platforms, together with the development of powerful computational tools, have transformed biological and biomedical research over the past several years allowing the achievement of large-scale population sequencing projects, such as the 1000 Genomes Project [4] and The Cancer Genome Atlas (www.cancergenome.nih.gov), and opened a new era for personal genomics [5–7].

Although these platforms have totally changed our ability to study the genome of any organism, they have technological

Alberto Magi, PhD, is an Assistant Professor at the University of Florence, Italy. His research interests focus on the development of computational methods for the identification of genomic variants.
Betti Giusti, PhD, is an Associate Professor of Clinical Pathology at the University of Florence. Her research activity is focused on genetics of cardiovascular diseases and extracellular matrix disorders.
Lorenzo Tattini, PhD, is a Postdoctoral Researcher at the Department of Computer Science, University of Pisa. His research is focused on computational methods for the analysis of DNA and protein data.

Submitted: 28 July 2016; **Received (in revised form):** 4 April 2016

limits. During the wash-and-scan process, the incorporation reaction of each DNA strand population becomes more and more asynchronous as each base is added. This phenomenon (dephasing) generates noise and sequencing errors, limiting the sequence size to 100–400 bp. Moreover, these platforms are based on polymerase chain reaction (PCR) to grow clusters of a given DNA template and this process introduces errors in the template sequence.

The past few years have seen the emergence of a third-generation of sequencing (TGS) technologies that include single-molecule real-time (SMRT) [8] sequencing and nanopore sequencing [9]. All these new sequencing approaches interrogate single molecule of DNA and do not need synchronization and amplification, overcoming the classical sequencing biases introduced by PCR and dephasing.

The SMRT sequencing method developed by Pacific Biosciences is based on directly observing a single molecule of DNA polymerase as it synthesizes a strand of DNA. It requires minimal amounts of reagent and sample preparation and because there are no scanning and washing steps, the time to result is faster than SGS methods [3]. Moreover, exploiting the processivity of the DNA polymerase, SMRT sequencing generates longer read lengths than any other first- and second-generation sequencing methods, producing average read lengths in the order of 10 kb.

The nanopore-sequencing approach consists in the transit of a DNA molecule through a pore-forming protein embedded in a membrane and measuring its effect on an electric current. From the 1990s, several researchers suggested the use of nanopores as biosensors [10] and demonstrated that ionic current passing through a nanopore depends on the identity of nucleic acid bases interacting with and transiting the nanopore [11–13].

Oxford Nanopore Technologies (ONT, https://www.nanoporetech.com) is a company founded in 2005 to develop a single molecule sensing system based on this proof-of-concept study. In 2012, it announced the smallest high-throughput sequencing platform, the MinION, at the Advances in Genome Biology and Technology meeting [14].

The MinION is a pocket-sized (90 g and 10 cm in length) device that is able to sequence long single-stranded DNA molecule. The two strands of a DNA molecule are linked by a hairpin and sequenced consecutively. When the two strands of the molecule are read successfully, a consensus is built to obtain a more accurate read (called the 2D read). Otherwise only the forward-strand sequence is provided (called the 1D read). When the DNA strand passes through the nanopore, a sensor measures ionic current changes with a sampling frequency of 3000 Hz and the raw current data are then subjected to base calling by means of a Hidden Markov Model (HMM): base-calling is first performed for template and complement strands separately (1D) and then are used to constrain the 2D base-calls of the DNA fragment.

In April 2014, ONT launched the MinION Access Programme (MAP), an independent beta-testing program for a developers' community made of more than 1000 laboratories (https://www.nanoporetech.com/community/the-minion-access-programme). Each participant received the MinION starter pack that included the MinION device, a USB cable, a Configuration Test Cell, two flow cells, a nanopore sequencing kit and a wash kit.

Although the MAP allowed several research groups to test the power of this novel TGS instrument, it became clear to all participants that it was impossible to evaluate reproducibility and quality of the MinION data from few sequencing runs. For these reasons, a group of MAP participants decided to form the MinION Analysis and Reference Consortium (MARC) with the aim of 'evaluating and providing standard protocols and reference data to the scientific community' [15]. During MARC phase 1, five laboratories sequenced the same *Escherichia coli* strain using the same protocol and generating a total of 20 data sets. The results of these experiments were recently published in F1000Research [15] reporting comprehensive analyses regarding sequencing protocols, base throughput, read quality and the performance of the MinION device itself.

In this article, we present the results of the first and most comprehensive analysis for understanding the capability of ONT data to be used in resequencing experiments. By using the data generated during the MARC phase 1, we studied the main characteristics of ONT reads and we evaluated the performance of different alignment approaches to map ONT sequences against a reference genome. Aligned data allowed us to quantify sequencing error rate for substituted, inserted and deleted bases and to study the stochastic properties of error distribution. Moreover, by using a complex strategy to simulate synthetic reference genomes, we evaluated the sensitivity and specificity of ONT data to detect substitutions, small insertions and deletions (InDels) and genomic regions involved in copy number variants (CNVs).

Results

ONT MARC data

The MARC phase 1 experiments were performed by five laboratories that sequenced the same *E. coli* strain, in duplicate, by using the R7.3 flow cells with two different sequencing kits: the SQKMAP005 (Phase 1a) and the SQKMAP005.1 (Phase 2b). Each sequence produced by the MinION was classified as pass and fail on the base of base quality and converted to fastQ files by using poreTools [16] (see Methods section).

Because the main goal of this article is to evaluate the capability of ONT data to be used in resequencing analyses, in this section, we briefly report the principal characteristics of MARC experiments in terms of sequencing throughput, read length and quality distribution. A deep and comprehensive analysis of the characteristics of data generated by the MARC can be found in [15].

The total throughput of each experiment is variable between and within different laboratories, ranging from a minimum of 28 Mb to a maximum of 385 Mb and with a total number of sequenced reads that goes from around 6000 to 45 000 (Table 1).

The average sequence length ranges between 5 and 7 kb for the great majority of the MARC experiments, and single reads range from hundreds bp to tens kb (Table 1 and Figure 1) for all the 20 experiments. On an average, pass sequences are longer (4–8 kb) than fail sequences (4–6 kb) and the amount of pass reads represent more the 60% of the total sequencing throughput in almost all experiments.

The read GC content distribution is close to *E. coli* reference genome for both pass and fail reads (Figure 1), and although the average quality of pass reads is clearly larger than fail reads, base quality does not depend on read position (except for around the first 100 bases, Figure 1E), demonstrating that the DNA strand translocation through the nanopore is not affected by position biases. This result is of fundamental importance because it suggests that nanopore-sequencing approach can generate high-quality sequences with no theoretical limits on length, except those introduced during sample preparation.

Table 1. MARC experiments statistics

Exp name	Phase	Pass reads	Fail reads	Pass base	Fail base	ARS pass	ARS fail	Base prop pass:fail	Read prop pass:fail
Lab1_run1	1a	32 548	14 806	228.3	86.8	6825.5	5626	0.72:0.28	0.69:0.31
Lab1_run2	1a	17 805	11 303	120.9	65.7	6658	5729	0.65:0.35	0.61:0.39
Lab2_run1	1a	8289	4790	59.3	30.3	7060	6138	0.66:0.34	0.63:0.37
Lab2_run2	1a	2901	1708	21.1	10.1	7200	5681.5	0.68:0.32	0.63:0.37
Lab3_run1	1a	18 765	7951	121.5	40.7	6367	4744	0.75:0.25	0.7:0.3
Lab3_run2	1a	19 169	7538	156.8	48.4	8007	6152	0.76:0.24	0.72:0.28
Lab4_run1	1a	13 836	10 858	69.4	44.2	3931	3479	0.61:0.39	0.56:0.44
Lab4_run2	1a	19 024	12 341	98.3	52.8	4352	3563	0.65:0.35	0.61:0.39
Lab5_run1	1a	23 566	6069	153.7	33.2	6242	4780	0.82:0.18	0.8:0.2
Lab5_run2	1a	17 673	26 351	48.2	39.4	1528	439	0.55:0.45	0.4:0.6
Lab1_run1	1b	12 258	17 511	69.4	78.1	5664	4464	0.47:0.53	0.41:0.59
Lab1_run2	1b	14 235	10 162	72.4	24.7	4738	667	0.75:0.25	0.58:0.42
Lab2_run1	1b	5165	5960	28.9	28.6	5438	4517.5	0.5:0.5	0.46:0.54
Lab2_run2	1b	28 054	30 044	206.5	178.2	7261.5	5944	0.54:0.46	0.48:0.52
Lab3_run1	1b	30 364	11 757	225.1	70.4	7235	5819	0.76:0.24	0.72:0.28
Lab3_run2	1b	14 800	6569	94.1	34.1	6285	5164	0.73:0.27	0.69:0.31
Lab4_run1	1b	1493	4673	8.4	20.0	5612	4042	0.3:0.7	0.24:0.76
Lab4_run2	1b	11 484	5856	65.5	27.3	5381	4371	0.71:0.29	0.66:0.34
Lab5_run1	1b	12 844	7876	83.8	43.0	6454	5257.5	0.66:0.34	0.62:0.38
Lab5_run2	1b	11 126	5894	72.8	31.7	6382.5	5113	0.7:0.3	0.65:0.35

Columns report the main characteristics of each experiment generated by the five laboratories of the MARC. For each experiment we reported the phase (Phase), the number of reads (Pass reads and Fail reads), the throughput in Mb (Pass base and Fail base), the average read length (ARS Pass and ARS Fail) and the proportion of reads and throughput between pass and fail reads (Base Prop and Read Prop). All the statistics were calculated from MARC fastQ files.

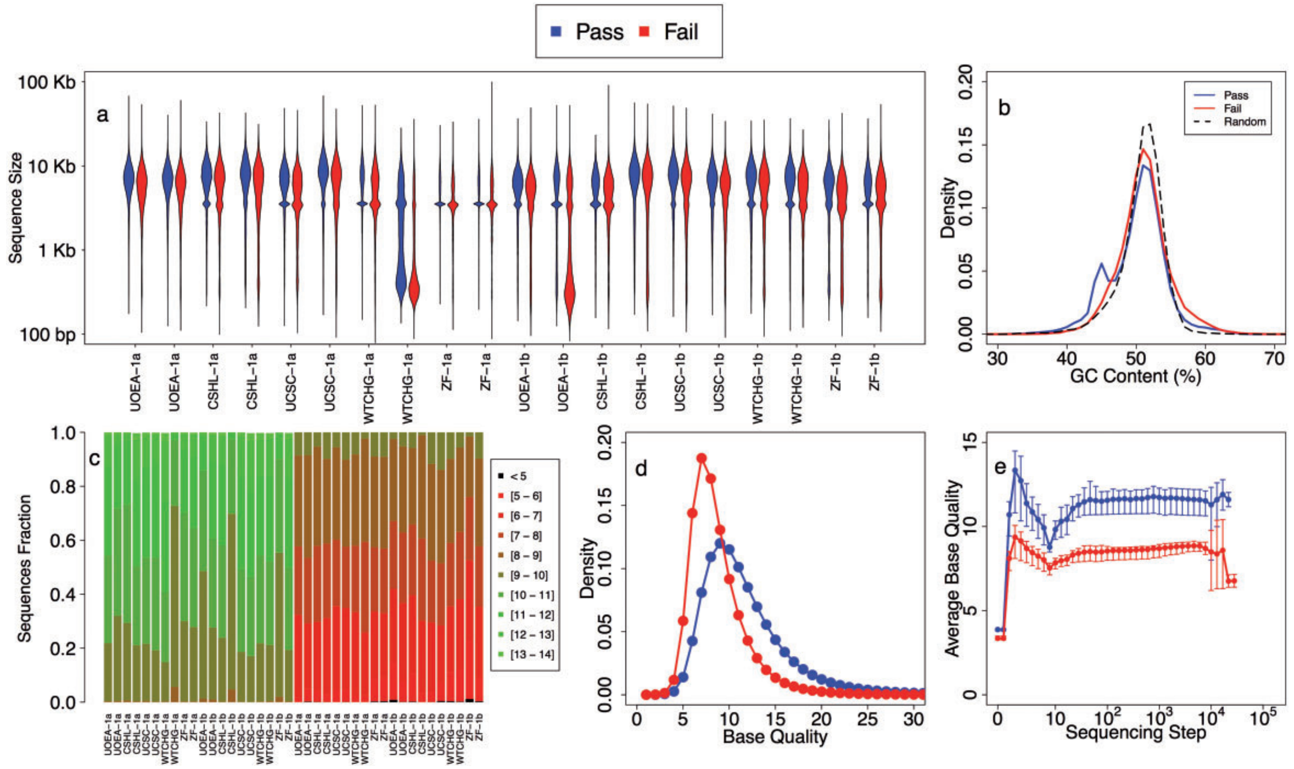

Figure 1. Size and quality distribution of pass and fail ONT sequences. Panel (A) shows read length distribution for pass and fail sequences of the 20 experiments performed by the MARC. Panel (B) reports the GC content percentage for pass and fail reads compared with randomly selected regions of the *E. coli* genome. Panels c, d and e show average read quality (C, the first 20 barplots are related to pass reads while the second 20 to fail reads), base quality distribution (D) and base quality as a function of sequence position (E) for pass and fail reads, respectively.

The results summarized in Table 1 and Figure 1 show that there are no significant differences between the experiments generated in Phases 1a and 1b in terms of sequencing throughput and median read lengths, in accordance with the data reported in [15].

Aligners and error rate estimation

The alignment of TGS sequences can be particularly challenging for the large number of long reads that they generate (from kb to tens of kb) and for the high error rates that are primarily InDels rather than substitutions [17]. The principal computational problem is how to align long (many kilobase) reads with moderate divergence from the genome (up to 20% divergence, concentrated in InDels) with the same speed and sensitivity of SGS alignment methods.

At present, few methods have been tested or developed to properly map long reads generated by TGS platforms. Chaisson et al. [18] proposed a novel method (Basic Local Alignment with Successive Refinement, BLASR) that combines the data structures used in short read mapping with alignment methods used in whole genome alignment (see Methods section for more details). Heng Li extended the BWA-MEM algorithm [19] by combining relaxed scoring of Smith–Waterman with heuristics filtering to support PacBio and ONT reads. Several papers proposed to map nanopore and PacBio data by using the approach adopted by LAST [20]. LAST follows a three steps approach in which first finds initial matches between reads and genome, then extend them with a gapless X-drop algorithm and finally extend them using a gapped X-drop algorithm [21]. Recently, Jain et al. [17] proposed a novel approach, marginAlign [17], properly devised for ONT data that realigns reads against a reference genome by combining a HMM with the alignments generated by LAST and BWA (burrows wheeler aligner). Henceforth, we will refer to marginAlign with LAST as HMML and to marginAlign with BWA as HMMB.

To understand the capability of different alignment approaches to properly map ONT sequences against a reference genome, we applied the five aforementioned long reads alignment methods (BWA, BLASR, LAST, HMML and HMMB, see Methods section for more details) to the pass and fail sequences of the 20 MARC experiments.

BWA, HMML and HMMB produced soft-clip alignments that represent 1.5–5% of mapped pass reads (1.5% for BWA and HMMB and 5% for HMML) and 10% of mapped fail reads (see Supplementary Table S1 for more details). Moreover, BWA was the only aligner to produce split mapping: 1% of pass reads and 8% of fail reads were splitted (Supplementary Table S1). Around 99% of pass reads and 80% of fail reads (Figure 2A) were aligned against the *E. coli* reference genome and mapping performance strongly depends on sequence size (Figure 2B and C and Supplementary Figure S1) as the longer the reads and the higher the fraction of sequences mapped by each method. At present, the best way to evaluate the likelihood that an alignment is correct is mapping quality (MQ). This score is generally estimated by considering various factors, such as the number of base mismatches and the sizes of inserted or deleted regions in the alignment [22]. We analyzed the MQ values generated by BLASR and BWA (mappers being evaluated that generate MQ) and we found that around 99% of mapped pass reads and 90–95% of mapped fail reads (see Supplementary Table S1) have $\text{MQ} \geq 20$. For this reason, all the subsequent analyses for BWA and BLASR will be performed using reads with $\text{MQ} \geq 20$. Although all the five alignment strategies gave similar results, the LAST algorithm obtained the worst global performance and resulted as the most influenced by sequence length for both pass and fail reads. All the mapping strategies tested in this work were not able to align 10% of pass reads shorter than 1 kb and 40% of fail reads shorter than 3 kb, suggesting that short reads with lower base qualities contain more sequencing errors than short reads with higher base qualities.

As a further step, aligned data were used to obtain a raw estimation of ONT error rate for the three main sources of local errors: mismatch, deletions and insertions. To this end, for each mapping algorithm, we calculated the number of bases that are substituted, inserted and deleted with respect to the reference genome as a function of sequence position and read quality.

Although the results of these analyses give a combination of sequencing and alignment errors, the use of five different mapping strategies allowed us to mitigate the alignment effect obtaining a good estimation of sequencing errors. To better evaluate the error rates estimated for ONT data, we compared these results with those obtained by the Illumina MiSeq and Pacific Bioscience platforms (see Methods).

Taken as a whole, panels d–o of Figure 2 show that the sequencing error rate slightly depends on read position, while it is highly influenced by read quality. The fraction of substituted, inserted and deleted bases (Figure 2D–I) increases with sequence position until reaching a constant value at around 50–100 bp for all the five aligners, with the exception of LAST on single-base substitutions. On the contrary, the error rate for the three variant categories decreases as the average read base quality increase (Figure 2J–O), suggesting that read quality and read errors are highly correlated.

To evaluate this correlation, we used the alignments data generated by BLASR, BWA and LAST to estimate the Phred-scaled mismatch rate as $Q = -10\log_{10}P$ (where P is the fraction of mismatches for each aligned read) and we compared it with the predicted quality scores. The results of these analyses are reported in Supplementary Figure S2 and show that predicted quality score accurately reflects measured mismatch rate for both pass and fail reads (although for fail reads, the predicted higher quality scores are underestimated).

All these analyses also demonstrate that the five mapping methods returned different results for the three error categories. The two marginAlign methods obtained the smallest error rate for single base substitutions, while LAST and BWA showed the best performance for small InDels (see Table 2 for more details).

The total error rates (sum of the three error rates) for BWA and the two marginAlign approaches (the three best methods in terms of performance) is around 11% (see Table 2) and in accordance with the total percent error estimated in the first paper released by the MARC [15]. As expected, the average error rate for pass reads (11%) is much smaller than that obtained for fail sequences (around 21%, see Table 2). Interestingly, although PacBio sequences present low error rate for substitutions (around 1%) they generate a total error rate comparable with ONT data as a consequence of the high insertions errors, and this is in accordance with previously published paper [23, 24]. As expected, the total error rate estimated for the SGS MiSeq reads is almost two order of magnitude (0.24%, Table 2) smaller than the other TGS technologies (around 11% for both PacBio and ONT, Table 2).

Error rate distribution

In resequencing studies, once the reads have been properly mapped, genomic variants are discovered by searching for

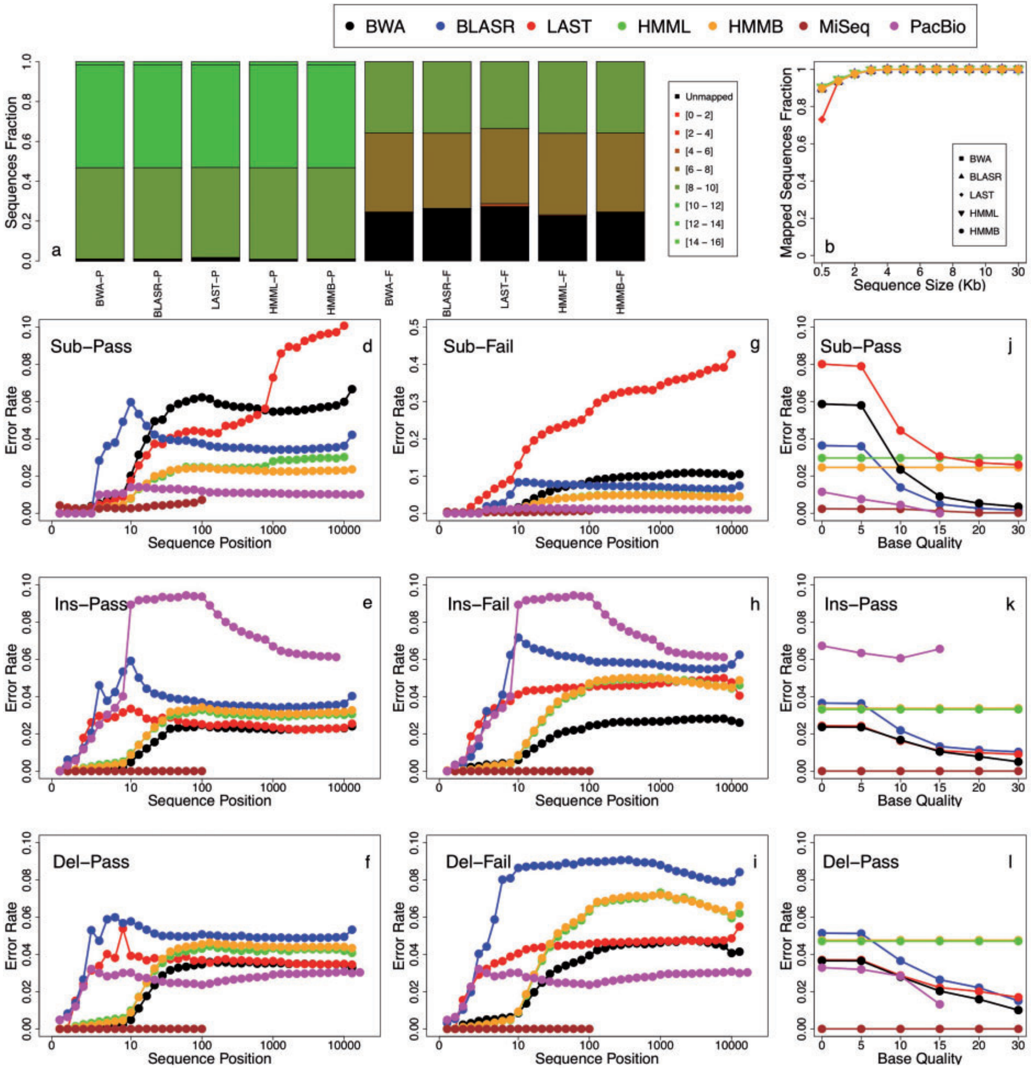

Figure 2. Mapping algorithms comparison and sequencing error rate estimation. Panels (A) show the proportion on mapped and unmapped reads for all the five aligners (colors represent average read qualities). Panels (B) and (C) report aligners performance as a function of read length for pass (B) and fail (C) sequences. Sequencing error rate was estimated as a function of sequence position (D–I) and base quality (J–O). Error rate was estimated for substituted (D, G, J, M), inserted (E, H, K, N) and deleted bases (F, I, L, O) for pass (D–F, J–L) and fail (G–I, M–O) sequences. To simplify subplots grouping, panels (D–O) contain additional labels that describe variants and read type: Sub- (substitutions), Ins- (insertions), Del- (deletions), -Pass (pass reads) and -Fail (fail reads).

Table 2. Error rate statistics

Aligner	Sub pass	Ins pass	Del pass	Total pass	Proportion pass	Sub fail	Ins fail	Del fail	Total fail	Proportion fail
BWA	5.87	2.37	3.66	11.9	49:20:31	13.82	3.6	6.16	23.58	59:15:26
BLASR	3.64	3.65	5.14	12.43	29:29:42	7.44	6.12	9.25	22.81	33:27:40
LAST	8.01	2.43	3.71	14.15	57:17:26	39.11	5.19	5.11	49.41	79:11:10
HMMB	2.97	3.32	4.71	11	27:30:43	6.35	6.23	8.96	21.54	29:29:42
HMML	2.46	3.37	4.77	10.6	23:32:45	6.06	6.15	8.83	21.04	29:29:42
MiSeq	0.24	0	0	0.24	100:0:0	0.24	0	0	0.24	100:0:0
PacBio	1.15	6.73	3.29	11.17	10:60:30	1.15	6.73	3.29	11.17	10:60:30

Columns report the most relevant information about error rate for substitutions (Sub), insertions (Ins) and deletions (Del) for pass and fail reads. 'Total' columns report the sum of substitution, insertion and deletion error rates. 'Proportion' columns report the relative percentage of each error class (Sub:Ins:Del).

differences between the reference genome and the aligned reads. For each genomic position, substitutions and small InDels (hereafter 'events') are inferred by comparing the number of reads that do not contain the reference allele and the total number of reads aligned with that position: a variant is called when the number of reads containing the same alternative allele is significantly large with respect to the total number of reads (for haploid genomes, a variant can be roughly called when half reads contain the same alternative allele). In this framework, although the error rate estimated in the previous section is a good approximation of sequencing accuracy (the capability of a sequencing technology to correctly sequence a DNA fragment), it is not able to predict the number of false-positive events generated by a resequencing analysis because it depends on recurrent errors aligned at the same genomic position.

For this reason, for each position of the reference genome, we counted the number of reads that contain the same substituted, inserted or deleted bases. In this way, for each alignment, we estimated the 'recurrent' error distribution that gives the probability to find N reads containing the same error aligned at the same position of the reference genome. The study of these distributions allowed us to estimate the probability of detecting false-positive events and to understand the randomness of recurrent errors (the probability to find *N* errors in the same position by chance).

To study the stochastic nature of recurrent errors, by using the sequencing error rates estimated in previous section, we simulated synthetic reads with randomly distributed substituted, inserted or deleted bases and we calculated their recurrent error distribution. By using this recipe, we obtained the probability distribution of finding *N* recurrent errors by chance and we compared with that generated by each alignment by means of the Kolmogorov–Smirnov statistics.

On one hand, the Kolmogorov–Smirnov statistics *D* quantifies the distance between two empirical distribution function and the smaller is *D* the closer are the two distributions. In our analyses, a small *D* value indicates that the recurrent error distribution of real and randomly generated reads are close and consequently real errors are randomly distributed along each read independently by the genomic position in which have been mapped. On the other hands, large *D* values indicate that errors in real reads are not randomly distributed but fall in recurrent positions of the genome.

On one hand, all the *D* statistics estimated for MiSeq and PacBio alignments (Figure 3) are close to zero, indicating that substitution, insertion and deletion errors are randomly generated during the sequencing process of these technologies. On the other hands, the *D* statistics obtained from ONT alignments suggest that the error distribution along the reads generated by nanopore sequencing process is not completely random (Figure 3 and Supplementary Figures S3–S8). LAST-based alignments (LAST and HMML) obtained *D* statistics close to one for all the three error classes, while BWA, BLASR and HMMB gave *D* values larger than those obtained by MiSeq and PacBio, in particular for deleted bases.

To evaluate the effect of recurrent errors on producing false-positive events, we counted the total number of genomic positions in which more than half of mapped reads contain the same substituted, deleted or inserted bases. As expected, the frequency of false-positive events depends on sequencing coverage and read base quality. Figure 3 shows that increasing the coverage mitigates the effect of recurrent error biases and reduces the total number of false-positive events. In the same way, the removal of reads with low base quality increases false-positive frequency by reducing sequencing coverage (Supplementary Figures S9 and S10). Surprisingly, although PacBio data show high sequencing error rate (on the same order of magnitude of ONT reads), they obtained the lowest false-positive rate for all the three variant classes, detecting less than one false substitution every 100 kb and around one InDel every 1 Mb. The reason of these results can be mainly ascribed to the nearly random nature of errors distribution along PacBio sequences (small D values). The performance of MiSeq sequencer are similar to those of PacBio and this is a direct consequence of the low sequencing error rate of SGS reads reported in previous section.

Concerning ONT data, although BLASR resulted the best aligner (in terms of false-positive frequency) for substitutions and insertions and LAST for deletions, the global performance obtained by this sequencing approach is poor for all the three variant classes. In the best experimental/computational setting (best aligner and coverages larger than 30×) ONT experiments produce around one false substitution and insertion every 10–50 kb and one false deletion every 1 kb, making a hard challenge the use of this data for small variants discovery. Moreover, the combination of pass and fail reads has little effect on reducing false-positive frequency (Figure 3).

As a further step, to understand the experimental and computational nature of recurrent errors, we studied the nucleotides content and the size distribution (for inserted and deleted bases) of all the false-positive events generated by each alignment. Although the five aligners produced slightly different results, the bar plots of Figure 3 and Supplementary Figure S11 show that recurrent errors follow specific nucleotide patterns that can be ascribed to intrinsic biases of the nanopore sequencing process. On one hand, recurrent substitution errors mainly affect C and G and, independently of the nucleotide they affect, substituted bases are enriched of C and G. On the other hand, recurrent-deleted bases principally involve A and T and mainly occur after A and T nucleotides of the genome. Supplementary Figure S11 also show that realignment strategy of marginAlign, irrespective of the mapper chosen for the primary alignment, introduces a bias which results in the missing of one or more nucleotides in poly-X homopolymers. Remarkably, inserted bases do not suffer of any apparent bias being equally distributed among the four nucleotides.

Moreover, we found that although the great majority of InDel calls are 1-base events for all the alignments, the two TGS data contain a significant fraction of inserted (PacBio) and deleted (PacBio and ONT) bases larger than 1 bp (Supplementary Figure S12).

Taken as a whole, these results suggest that the translocation of C (G) through the nanopore is preferentially miscalled with G (C), while the translocation A and T may result in the loss of one (or more) subsequent nucleotides by the sequencing/base-calling process.

Although it is difficult to completely explain the reasons of these errors, we speculate that both deletions and C–G miscalling can be mainly ascribed to algorithmic limits of the HMM at the base of the Metrichor base caller. Taken as a whole, the results reported in this section can be of fundamental importance for improving the performance of base-calling methods and for the development of novel algorithms for the identification of small variants by using ONT data.

Depth of coverage

At present, the most powerful method for the identification of CNVs in resequencing analyses is the depth of coverage (DOC) approach [25, 26].

The DOC approach is based on the simple idea that during the sequencing process, the reads are randomly and independently sequenced from any location of the genome. Under this assumption, the number of reads mapping into a window of the reference genome should be proportional to the number of times the region appears in the DNA sample and follow a Poisson distribution. Following this assumption, the copy number of any genomic region can be estimated by calculating the

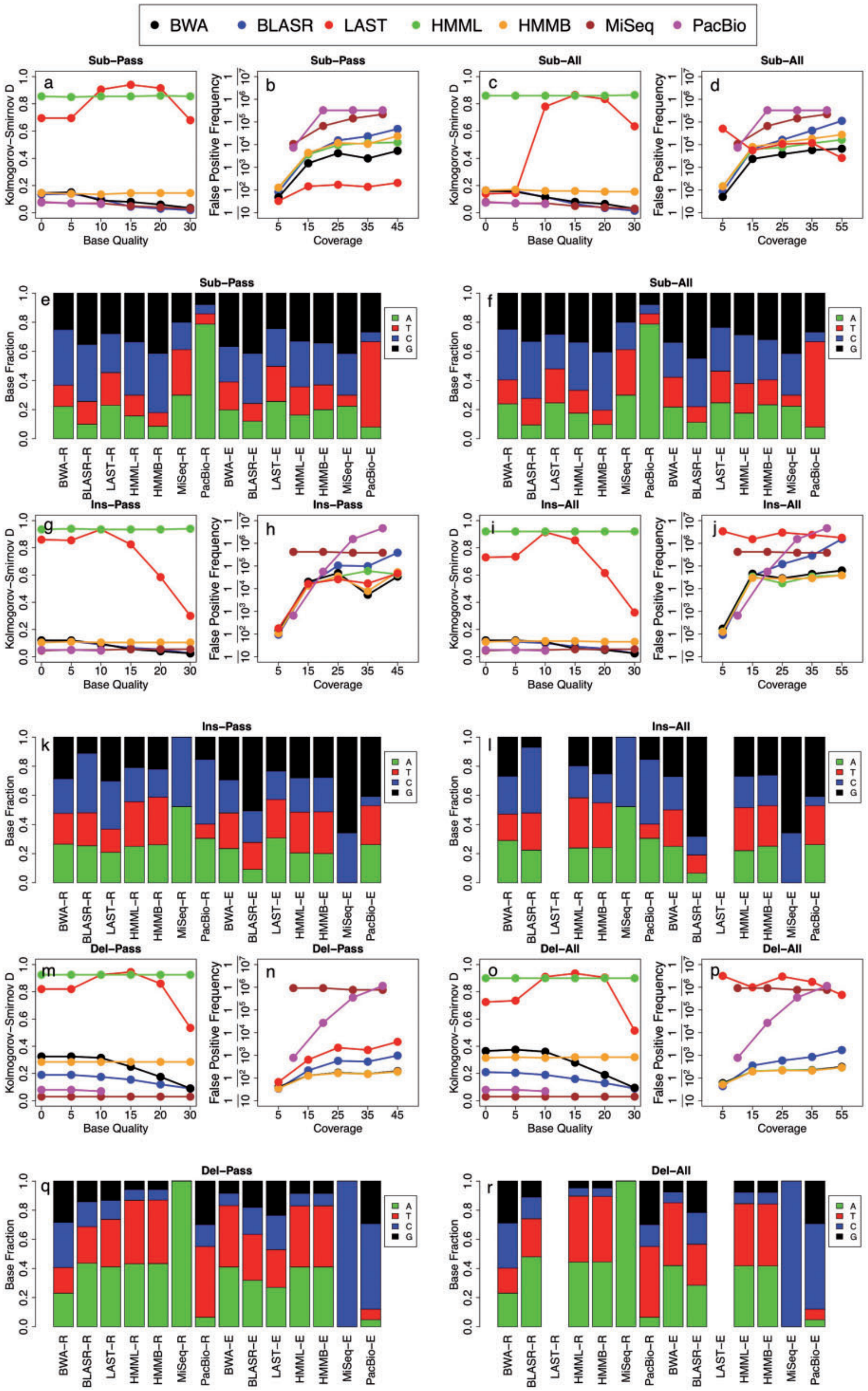

Figure 3. Recurrent errors distribution analysis. Summary of the recurrent error distribution analyses for substituted (A–F), inserted (G–L) and deleted (M–R) bases. Panels (A, C, G, I, M and O) report the Kolmogorov–Smirnov statistic as a function of read quality, while panels (B, D, H, J, N and P) show the false-positive frequency as a function of average sequencing coverage. False-positive frequency is defined as the ratio between the total number of false positive events and the size of *E. coli* genome in bp. False-positive events are defined as genomic loci in which more than half of the aligned reads contain the same error. The barplots of panels (E, F, K, L, Q and R) report the base content of false positive events for substituted (E, F), inserted (K, L) and deleted (Q, R) bases. Each bar with suffix -R reports the distribution of nucleotides in which the false event occurs (for InDels the nucleotide before the event). Each bar with suffix -E contains the base content of the substituted/deleted/inserted bases. Panels (A, B, E, G, H, K, M, N and Q) report results for pass reads, while panels (C, D, F, I, J, L, O, P and R) for pass+fail reads. To simplify subplots grouping, all panels have title labels that describe variants and read type: Sub- (substitutions), Ins- (insertions), Del- (deletions), -Pass (pass reads) and -All (pass+fail reads).

DOC of reads aligned to consecutive and non-overlapping windows of the genome. To understand the capability of ONT data to identify genomic regions involved in CNVs, we studied the statistical properties and biases of DOC distribution and we compared it with the other two sequencing technologies.

As a first step, we studied the relationship between DOC and classical genomic biases: local GC content and mappability (defined as the inverse of the number of times that a sequence originating from any position in the reference genome maps to the genome itself) calculated as in [27].

On one hand, the correlation between DOC and GC content has been previously reported in several papers for SGS data and is mainly owing to the amplification step of the sequencing process. On the other hand, mappability bias is owing to the fact that the genome contains many repetitive elements and aligning reads to these positions leads to ambiguous mapping. In Magi et al. [25], by analyzing Illumina, 454 and SoLID reads, we observed that DOC is maximum for values of GC content between 35% and 60%, while it decreases at both extremes. In the same paper, we also found that DOC distribution for highly mappable regions is closer to Poissonian than genomic regions with low mappability.

On one hand, the results summarized in Figure 4 clearly show that ONT reads are slightly affected by the two classical sequencing biases with the exception of LAST alignment that is highly influenced by mappability. On the other hand, PacBio and MiSeq coverages strongly depend on local GC content and this can be mainly ascribed to the PCR chemistry at the base of these two sequencing approaches.

As a further step, to understand the stochastic properties of coverage distributions, we calculated the index of dispersion (ID) for different window sizes (10 bp, 20 bp, 50 bp, 100 bp, 200 bp, 500 bp, 1 kb, 2 kb, 5 kb, 10 kb and 20 kb). The ID, defined as the ratio between variance and mean, is used to quantify whether a set of observations are clustered or dispersed. In particular, ID larger than one indicate overdispersed data that follow a negative binomial distribution, ID smaller than one refer to underdispersed data that follow a binomial distribution, while $\mathrm{ID} = 1$ indicates data with Poisson distribution. In [25], we demonstrated that DOC distribution from SGS sequences exhibit an ID largely greater than one and that this over dispersion can be accounted to local GC content and mappability.

All the ONT DOC distributions, with the exception of LAST alignments, have an ID close to one (Figure 4) that demonstrate the Poissonian nature of the nanopore-sequencing process as a direct consequence of the low influence of GC content and mappability on these data. The large ID obtained by LAST can be mainly ascribed to the mappability bias of this alignment method, while the overdispersion of PacBio distributions is principally owing to the GC content bias described above. Although MiSeq data are strongly affected by GC content, they show small ID values that are the consequence of the small variance of these data.

As a final step, to evaluate the false-positive rate of CNV events, we calculated the fraction of genomic windows in which the 1-copy normalized DOC is larger than 1.5 (for duplication) and smaller than 0.5 (for deletions). ONT data obtained the best results for both duplicated and deleted regions, while PacBio reads gave the highest error rate demonstrating a poor suitability for CNVs analysis. Concerning ONT alignments, the BWA mapping data obtains the smallest error rate outperforming the other four methods. Moreover, the results reported in panels v1-w3 of Figure 4 show that ID and error rate decrease at the increasing of the window size and this trend is highly correlated with the read size: MiSeq data start to decrease from window size larger than 100 bp, while TGS data from window sizes larger than 2 kb.

Taken as a whole, these analyses demonstrate that the nanopore-sequencing process is a uniform process in which reads are randomly and independently sequenced. Notably, the error rate produced by ‘all’ reads (combining pass and fail) is much smaller than the error rate obtained with pass reads: although fail reads contain a large fraction of substituted, inserted and deleted bases they produce an increase in coverage that decrease the variance of DOC distribution and consequently the number of false-positive windows.

Variants detection accuracy

To evaluate the detection rate of ONT data for substitutions, small InDels and CNVs, we aligned the MARC data (pass and combined pass and fail) and the other sequencing experiments against synthetic *E. coli* reference genomes.

Synthetic reference genomes were generated by substituting, inserting and removing bases from the *E. coli* reference genome (see Methods section for more details). By using this approach, we were able to simulate substitutions, small InDels from 1 to 50 bp and deletions from 200 bp to 5000 kb in size. Moreover, by using a sophisticated strategy based on removing segmental duplicated regions from the *E. coli* reference genome, we were able to simulate multiple copy duplications (see Methods). After read mapping against the synthetic reference genomes, the detection rate for substitutions and small InDels was roughly estimated by calculating the proportion of modified loci in which more than half of the aligned reads contain the original reference allele. Detection rate was studied as a function of the local DOC of modified loci and as a function of variant size for small InDels.

The results of these analyses are summarized in Figure 5 and show that, as expected, MiSeq outperforms TGS methods for both substitutions and small InDels detection accuracy. PacBio obtained good results for substitutions discovery but completely failed the detection of small Indels. ONT data reached a detection rate of 0.9 for the discovery of substitutions with the two marginAligner mappers, and although the accuracy for small insertions was poor (<0.1), for small deletions BWA and marginAlign obtained detection rates in the order of $\sim$0.3, much larger than that obtained by PacBio. As expected, the larger the InDel size, the smaller the capability of all the alignment data to detect them. Remarkably (with the exception of MiSeq data), local DOC has little effect on detection rate, while the use of combined pass and fail reads reduces the sensitivity for both substitutions and small InDels identification with respect to using only pass sequences.

At present, few methods have been developed for calling variants with ONT data and these methods, that include Nanopolish [28] and marginCaller [17], are capable to search for substitutions only. The Nanopolish variant caller first selects candidate variants on the base of mismatches between aligned reads and the reference genome and then groups them into sets of close variants. Each cluster of variants is used to generate a set of candidate haplotypes from the possible combinations of SNVs and the haplotype that maximizes the probability of the event-level data is called as the sequence for the region. The marginCaller (marginAlign tool) computes posterior alignment match probabilities between the bases in the reads and the reference by using a realignment strategy based on HMM.

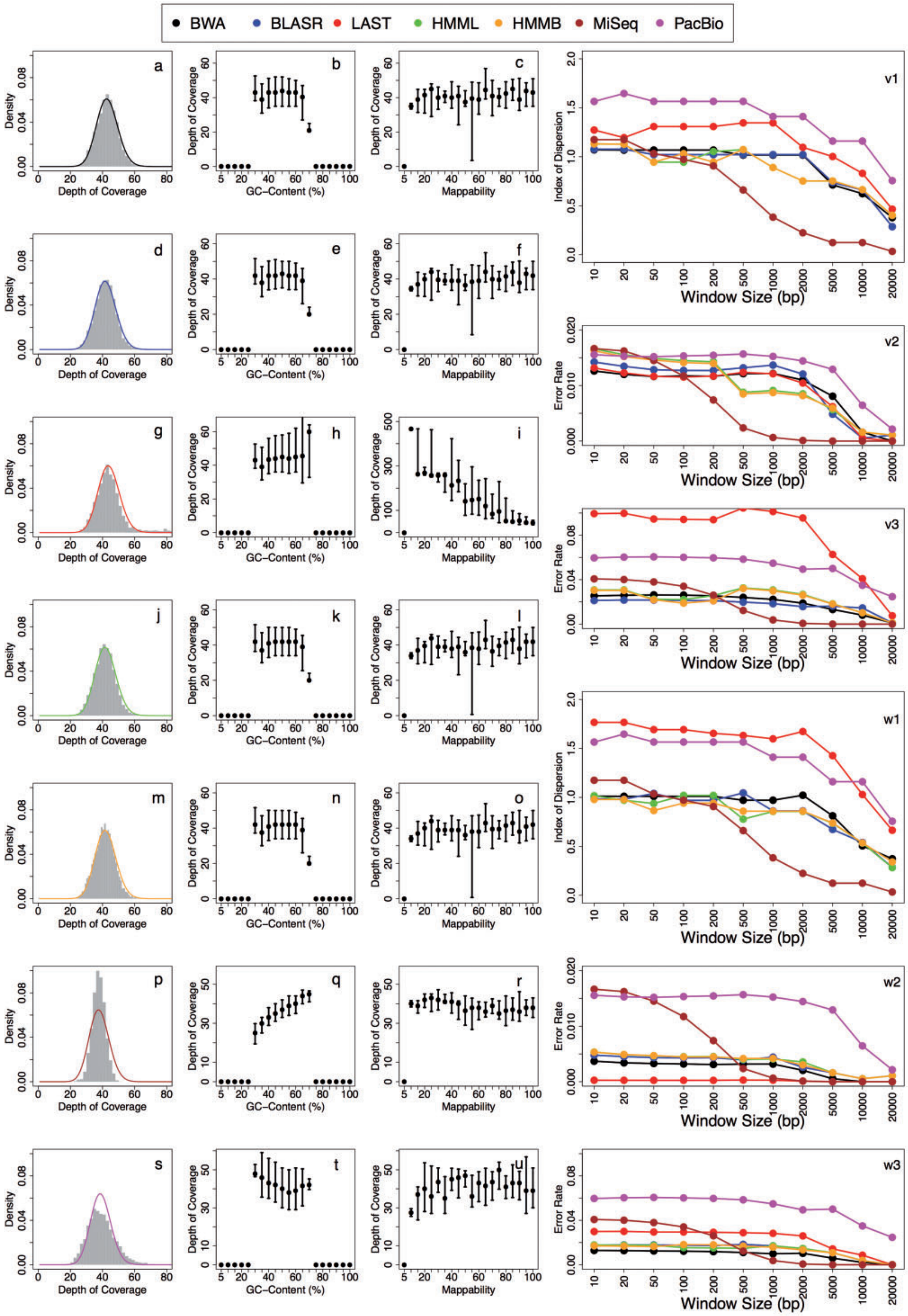

Figure 4. DOC distributions and biases. The first column of panels (a–u) reports the histograms of DOC and the superimposed Poisson distribution (solid lines) for all the alignments. Second column shows the correlation between DOC and GC content, while third column the correlation between DOC and mappability. The rows of panels (a–u) show the results for different aligners/platforms: BWA (A–C), BLASR (D–F), LAST (G–I), HMML (J–L), HMMB (M–O), MiSeq (P–R) and PacBio (S–U). Panels (v1) and (w1) report the ID as a function of window size, (v2) and (w2) the error rate for duplications and (v3) and (w3) for deletions. Panels (v1), (v2) and (v3) refer to pass reads, while panels (w1), (w2) and (w3) refer to all (pass+fail) reads.

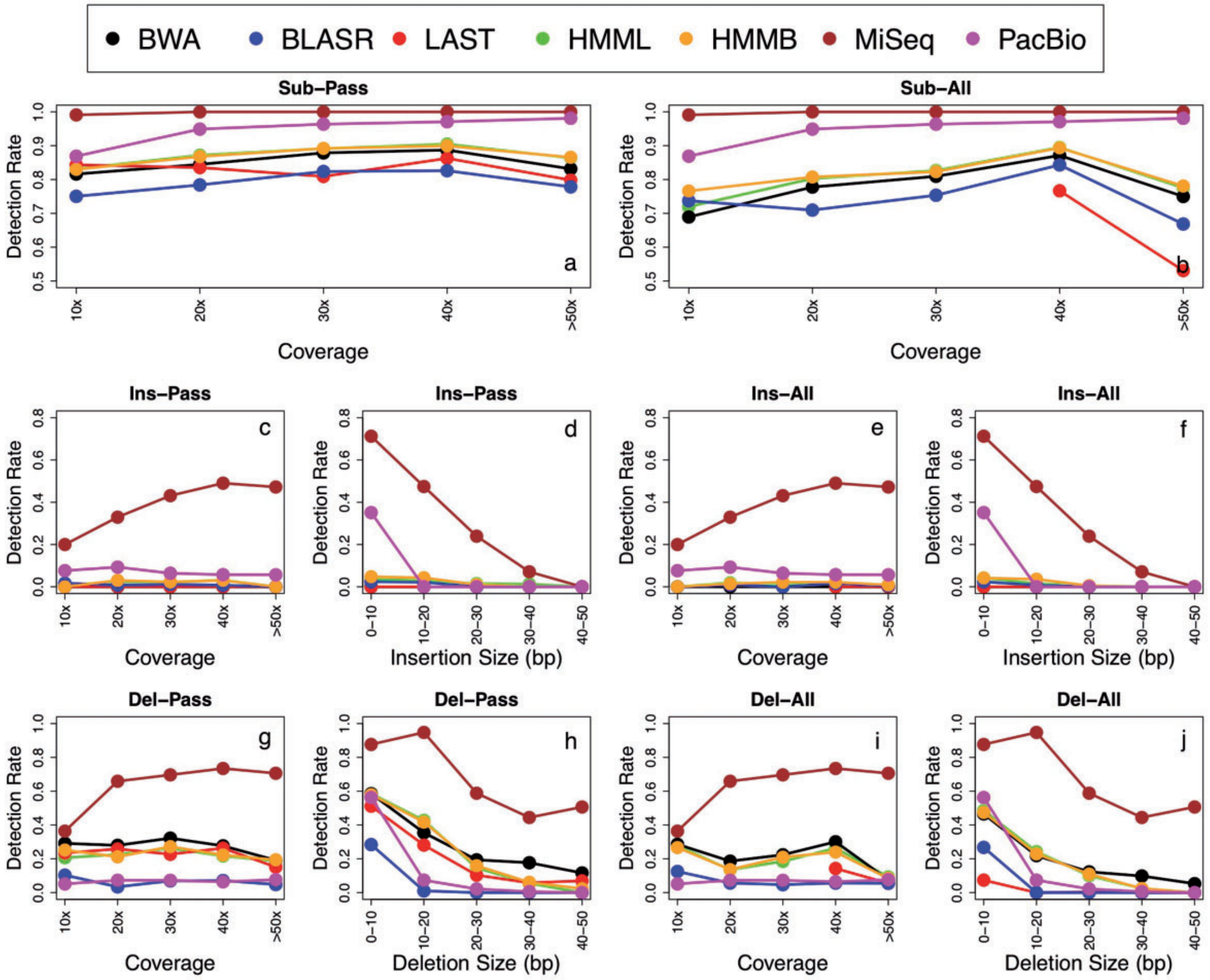

Figure 5. Detection rate for substitutions and small InDels. Summary of the detection rate estimated with synthetic reference genomes. Panels (A) and (B) report the detection rate for substitutions as a function of base coverage. Panels (C–F) and (G–J) report the detection rate as a function of coverage and size for InDels respectively. The analyses were performed for pass reads (A, C, D, G, H) and for combined pass and fail reads (B, E, F, I, J). To simplify subplots grouping, all panels have title labels that describe variants and read type: Sub- (substitutions), Ins- (insertions), Del- (deletions), -Pass (pass reads) and -All (pass+fail reads).

Unfortunately, nanopolish performance could not be tested owing to the lack of raw Fast5 files, but the analyses performed with marginCaller on synthetic variants data set (Supplementary Figure S13) demonstrate that this tool is capable to reach a detection rate around 0.99. However, these analyses also show that the high-detection rate of marginCaller is obtained at the expenses of a significant number of false-positive substitutions in C and G demonstrating that the HMM algorithm at the base of this tool is not capable to mitigate the effect of recurrent error bias of ONT reads.

To evaluate the accuracy of different sequencing technologies to identify genomic regions involved in CNVs, we calculated the 1-copy normalized DOC for different window sizes. The absolute number of DNA copies of each simulated variant was estimated by calculating the median DOC of the windows within the region and a deletion is called if this value is smaller than 0.5, while a duplication is called if it is larger than 1.5. The results reported in Figure 6 and Supplementary Figure S16 clearly show that although all the sequencing technologies are capable to correctly identify deleted regions (0-copies), only MiSeq and ONT reads aligned with BWA are able to identify duplications with high accuracy and to estimate the exact number of their DNA copies even for highly duplicated regions. Moreover, ONT-BWA data obtained the best correlation between simulated and predicted copy number, outperforming the MiSeq data. Notably, Supplementary Figure S14 demonstrates that sequencing coverage has little effect on CNV detection rate.

These results, combined with those reported in previous section, demonstrate that ONT data can be readily used to identify CNVs with high accuracy.

Discussion and conclusion

The advent of nanopore-sequencing technology is going to revolutionize our capability to study and understand the genome complexity of any organism. The advantages over current SGS and TGS technologies are the faster sequencing time and the longer read lengths that will improve *de novo* assembly and enable haplotype reconstruction and even whole chromosome phasing. At present, there is limited number of papers published in scientific journals that make use of nanopore data,

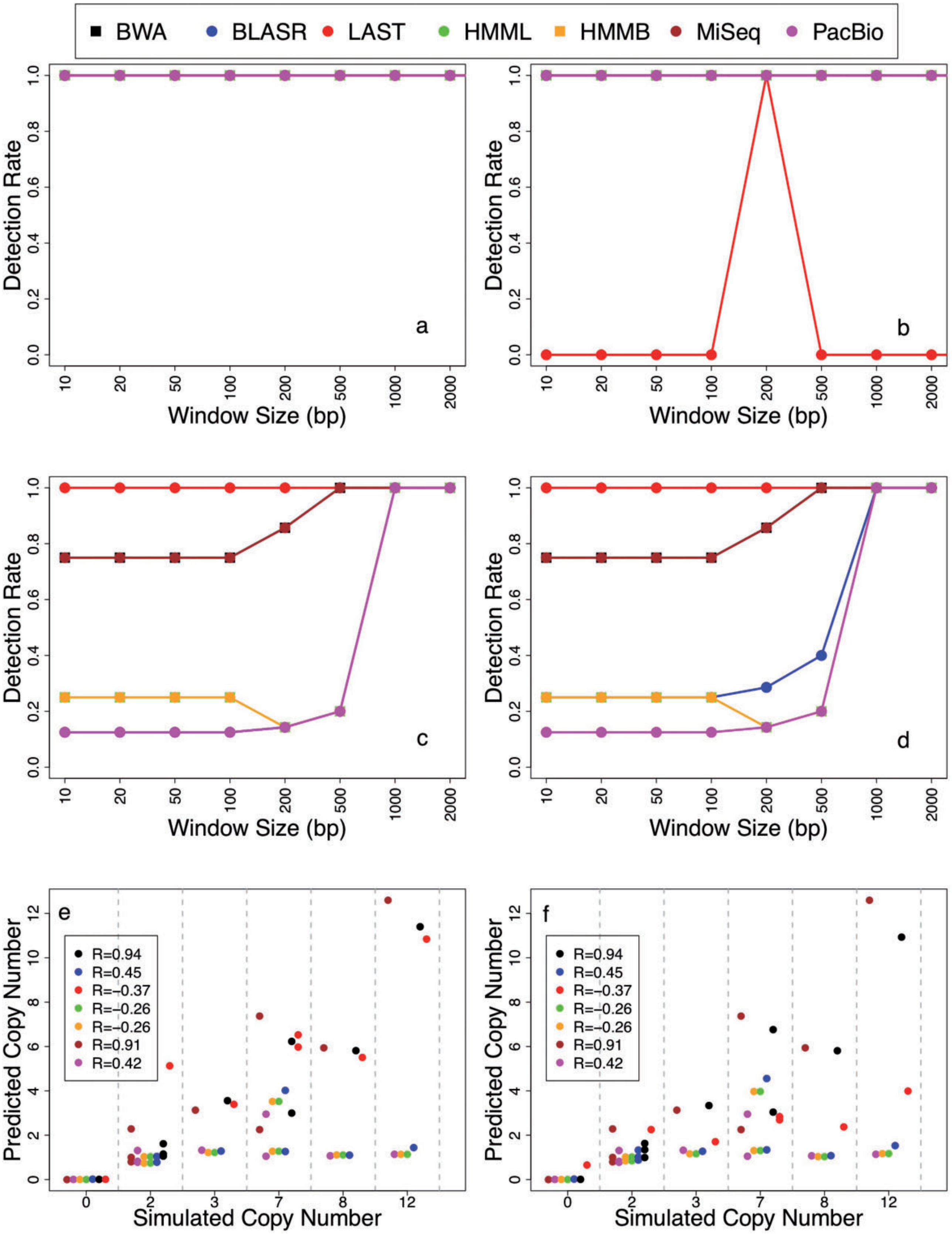

Figure 6. Detection rate for CNVs and absolute number of DNA copies prediction. Panels (A–D) show the detection rate of simulated deletions (A, B) and duplications (C, D) as a function of window size. Panels (E and F) report the correlation between the simulated and predicted absolute number of DNA copies for all the aligners/platforms. Panels (A, C and D) show the results for pass reads, while panels (B, D and F) for combined pass and fail reads.

and the great majority of these works are focused on bacterial genome assembly and bioinformatic tools for data handling and primary analysis.

Quick et al. [29] were the first to sequence an entire bacterial genome with a single run of the MinION. Loman et al. [28] demonstrated that it is possible to assemble a complete bacterial genome by using only nanopore sequencing data, while Ashton et al. [30] used the MinION to resolve the structure and chromosomal insertion site of an antibiotic resistance island in *Salmonella typhi*. Goodwin et al. [31] were the first to sequence a Eukaryotic Genome with MinION and demonstrated that nanopore data achieve improved assembly compared with Illumina sequencing alone.

Regarding computational approaches, the first generation of tools developed by the MAP community were mainly focused on the methods for evaluating and correcting raw MinION data. Poretools [16] and poRe [32] allow to convert raw data and generate quality control charts, NanoOK [33] exploits different aligners to estimate data quality and accuracy, while marginAlign [17], NanoCorr [31], Nanopolish [28], PoreSeq [34] and GraphMap [35] are properly devised to use MinION reads for genome assembly, alignment, realignment and error-correction.

Although the potential power of this sequencing approach is enormous, much work still remains to be done for improving read accuracy and developing novel computational methods tailored for this kind of data. In this scenario, the results reported in this article can be useful for guiding the development of the next generation of tools for nanopore data analyses. In this work, we evaluated the quality and accuracy of nanopore data and their capability to be exploited for resequencing analyses. To this end, we used the largest nanopore-sequencing data set to date, generated by the MARC using the first commercial nanopore-sequencing device, the MinION. The large number of experiments comprised in this data set allowed us to estimate with high confidence the main characteristics of nanopore sequences and compare them with other sequencing technologies.

The average read length produced by the MinION is around 6–8 kb, much larger than the 100–400 bp of SGS platforms and comparable with the size of the latest chemistry of PacBio reads. The global sequencing error rate is around 11%, it is mainly caused by inserted and deleted bases, it is comparable with the other TGS platform and nearly two orders of magnitude greater than Illumina sequences. The analyses performed in this work also show that nanopore-sequencing errors are not randomly distributed along the reads but mainly occur on specific nucleotides (C and G for substitutions and A and T for deletions) producing an unexpected number of genomic loci with recurrent substituted, inserted and deleted bases.

This biased error distribution generates a significant number of loci in which more than half of the aligned reads contain the same error and that can consequently be identified as false-positive variants. We estimated that in the best experimental/computational settings, ONT resequencing data can produce around one false substitution and insertion every 10–50 kb and one false deletion every 1000 bp.

For small bacterial genomes, these errors can generate 'only' hundreds/thousands of false-positive events, but when projected to higher order genomes (such as the Human genome) can lead to the identification of millions of false variants that will make difficult the interpretation of the results.

The analyses we performed on synthetic genomes showed that substitutions and deletions can be detected with accuracy comparable with SGS data, while this does not hold for insertions making still challenging the use of this technology for small-size variant discovery.

At present, publicly available tools for ONT data allow for the detection of substitutions only and, although the computational recipes implemented in these methods obtain good results in terms of sensitivity, they cannot remove the systematic biases of ONT data reported in this article. The high substitution rate mainly occurring on C and G and the high deletion rate involving A and T suggest a reassessment of ONT chemistry and/or base-calling algorithms and the development of novel variant-calling methods that include this nucleotide information a priori.

Coverage analysis demonstrated that, contrary to other PCR-based sequencing technologies, the nanopore sequencing is a uniform process that generates sequences randomly and independently without classical sources of bias such as GC-content and mappability. Thanks to these properties, the ONT data can be readily used to detect genomic regions involved in CNVs with high accuracy, outperforming PacBio and even SGS data in terms of both sensitivity and specificity. In this framework, the tuning of smoothing and segmentation models previously developed for the analysis of CGH-array and DOC signals from SGS data could be a promising starting point.

The results reported in this work are based on data generated with R7.3 flow cells and SQK-MAP005 kits and base-calling performed with the HMM algorithm implemented in Metrichor 1.12. At present, ONT is planning novel advances that include a new chemistry, termed R9, and a new base-calling algorithm based on Recurrent Neural Networks.

Methods

Experimental data

The MARC phase 1 experiments were designed to evaluate accuracy and reproducibility of the MinION data and for providing standard protocols and reference data to the scientific community. A laboratory *E. coli* strain (NCBI RefSeq NC_000913) was chosen as it has a single circular chromosome of 4.6 Mb that could be sequenced to sufficient depth in a single MinION run. A total of 20 experiments were performed by five laboratories that sequenced the same *E. coli* strain, in duplicate, by using the R7.3 flow cells with the same protocol for culture and DNA extraction and two different protocols for library preparation and sequencing: the SQKMAP005 kit was used for the Phase 1a experiments, while the SQKMAP005.1 kit for the Phase 2b. The detailed protocol for sequencing double-stranded total genomic DNA was based on the standard protocol from ONT at the time the experiment was conceived and is described in [15]. Each sequence produced by the MinION was base-called using the Metrichor 1.12 protocol and classified as pass and fail. Pass sequences are all those reads for which 2D base-calling was successful and the mean base quality is larger than 9, while fail reads include 2D reads with mean quality smaller than 9, 1D base-calling and failed base-calling. Pass and fail reads were extracted from the base-called FAST5 files and converted to FASTQ files by using poreTools version 0.5.1 [16] https://github.com/arq5x/poretools as described in [15]. The raw reads in FASTQ format for each of the 20 experiments are available from the European Nucleotide Archive project PRJEB11008 (http://www.ebi.ac.uk/ena/data/view/PRJEB11008).

Other sequencing data

To better evaluate the main characteristics of nanopore sequences, we compared with publicly available data produced by two other sequencing platforms, including the Pacific Biosciences and the Illumina MiSeq platforms. PacBio data were downloaded from the DevNet project on github (https://github.com/PacificBiosciences/DevNet). The DevNet project contains data sets of many organisms generated by the PacBio RS II platform with different sequencing chemistry. For our analyses, we used the sequencing data gathered with a PacBio RS II System and the latest P6-C4 chemistry on a size selected 20 kb library of *E. coli* K12 strain. PacBio data were downloaded in h5 format and converted to FASTQ by using the DESTRACTOR tool (https://github.com/thegenemyers/DEXTRACTOR) filtering out reads with quality smaller than 0.80 (−s800). The PacBio data set contain around 76 000 sequences with length ranging between 500 bp and 42 kb and with a median size of 8 kb. PacBio sequences were aligned against the *E. coli* reference genome by using the BLASR aligner with '-bestn 1 -m 0' options, according to SMRT PacBio resequencing pipeline (https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Evaluating-Assemblies). MiSeq data were downloaded from the European Nucleotide Archive project PRJNA196622 (http://www.ebi.ac.uk/ena/data/view/SRR826444). MiSeq paired end reads were aligned against the *E. coli* reference genome with the BWA-MEM method by using default settings. Both PacBio and MiSeq aligned data were downsampled to simulate coverage from 10× to 40×.

Aligners

All the ONT data used in this article were aligned against reference genomes by using four different mapping tools: BWA, BLASR, LAST and marginAlign. For each tool, parameters were chosen on the base of the recommendations and tweakings made by other MAP and MARC participants and reported in [17]. BWA version 0.7.12 was used with the '-x ont2d' that was properly devised for the alignment of ONT 2D-reads. BLASR (http://bix.ucsd.edu/projects/blasr/) was applied to ONT reads with the parameters -sdpTupleSize 8 -bestn 1 -m 0 as reported in [17]. For LAST mapper (http://last.cbrc.jp/), we used the parameters tuned by another MAP participant ('-s 2 -T 0 -Q 1 -a 1 -b 1 -q 1 -r 1') and that are published in [29]. marginAlign uses an HMM to realign reads previously mapped against a reference genome with BWA or LAST. The HMM is first trained on a test data set and the trained model is then used for realigning reads. MarginAlign was applied for all the ONT experiments with both BWA (HMMB) and LAST (HMML) mapping algorithms. For all the five mapping methods, pass and fail reads were aligned separately, and when necessary (for combined pass and fail reads analyses), BAM files were merged using Samtools merge [19]. All the results reported in this article were obtained by parsing BAM files with Samtools and in house bash and R scripts.

Synthetic reference genomes

To evaluate the capability of mapping methods to identify different classes of genomic variants, we generated synthetic alterations by manipulating the *E. coli* reference genome (NCBI RefSeq NC_000913). To simulate substitutions, we randomly substituted single bases of the reference genome, while to simulate small and large deletions (insertions), we inserted (removed) sequences in random positions of the reference genome. Small InDels were simulated from 1 to 50 bp in size while large events were simulated with 100, 200, 500, 1000, 2000, 5000, 10 000, 20 000 and 50 000 bp. For small variants (substitutions and InDels), we simulated 100 events for each synthetic genome, while for larger variants, we simulated 10 events for each synthetic genome. To simulate genomic regions with multiple copy duplications, we first analyzed the *E. coli* reference genome by using MUMmer [36] (http://mummer.sourceforge.net/) to search for segmental duplications. Segmental duplication is defined as genomic regions larger than 50 bp and with sequence identity larger than 90%. In *E. coli* reference genome, we found eight segmental duplications with a size from 300 to 5000 bp and with a number of copies ranging between 2 and 13. For each region, the removal of all the duplicated segments, except one, allowed us to simulate duplications with different numbers of copies. By using this recipe, we were able to simulate genomic regions with an absolute number of DNA copies ranging from 0 to 12. All the sequencing data (ONT, Illumina and PacBio) were aligned against the synthetic reference genomes by using the parameter settings reported in 'Aligners' section and BAM files were parsed with in house scripts. ONT reads aligned against the synthetic reference genomes with marginAlign (HMML and HMMB) were used to test the performance of marginCaller tool by using default settings and following the instruction found at https://github.com/benedictpaten/marginAlign.

Key Points

- The Oxford Nanopore Technologies MinION is a new device, based on nanopore sequencing that is able to generate reads of tens of kilobases in length with lower cost and faster sequencing time with respect to other platforms.
- The global sequencing error rate is around 11%, it is mainly caused by inserted and deleted bases, it is comparable with the Pacific Biosciences platform and nearly two orders of magnitude greater than Illumina sequences.
- Nanopore-sequencing errors are not randomly distributed along the reads but mainly occur on specific nucleotides (C and G for substitutions and A and T for deletions) producing an unexpected number of genomic loci with recurrent substituted, inserted and deleted bases that can consequently be identified as false-positive variants: MinION resequencing data can produce around one false substitution and insertion every 10–50 kb and one false deletion every 1000 bp.
- MinION data can be readily used to detect genomic regions involved in copy number variants with high accuracy, outperforming PacBio and even Illumina data in terms of both sensitivity and specificity.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

AM was supported by Italian Ministry of Health, Young Investigators Award, Project GR-2011-02352026 "Detecting copy number variants from whole-exome sequencing data applied to acute myeloid leukemias".

References

1. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet* 2010;**11**:31–46.

2. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010;**19**(R2):R227–40.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**(6):333–51.
4. Abecasis GR Altshuler D. 1000 Genomes Project Consortium, *et al* A map of human genome variation from population-scale sequencing. *Nature* **467**(7319):1061–73.
5. Snyder M, Du J, Gerstein M. Personal genome sequencing: current approaches and challenges. *Genes Dev* 2010;**24**(5):423–31.
6. Topol EJ. From dissecting cadavers to dissecting genomes. *Sci Transl Med* 2013;**5**(202):202ed15.
7. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med* 2011;**17**(3):297–303.
8. Eid J, Fehr A, Gray J, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**(5910):133–8.
9. Clarke J, Wu HC, Jayasinghe L, *et al*. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 2009;**4**(4):265–70.
10. Kasianowicz JJ, Brandin E, Branton D, *et al*. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* 1996;**93**(24):13770–3.
11. Akeson M, Branton D, Kasianowicz JJ, *et al*. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophys J* 1999;**77**(6):3227–33.
12. Derrington IM, Butler TZ, Collins MD, *et al*. Nanopore DNA sequencing with MspA. *Proc Natl Acad Sci U S A* 2010;**107**(37):16060–5.
13. Manrao EA, Derrington IM, Pavlenok M, *et al*. Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PLoS One* **6**(10):e25723.
14. Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat Biotechnol* 2012;**30**(4):295–6.
15. Ip CLC, Loose M, Tyson JR, *et al*. MinION Analysis and Reference Consortium: *Phase 1 data release and analysis. *F1000Res* 2015;**4**:1075.
16. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* 2014;**30**(23):3399–401.
17. Jain M, Fiddes IT, Miga KH, *et al*. Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 2015;**12**(4):351–6.
18. 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.
19. Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
20. Kiełbasa SM, Wan R, Sato K, *et al*. Adaptive seeds tame genomic sequence comparison. *Genome Res* 2011;**21**(3):487–93.
21. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**(17):3389–402.
22. Ruffalo M, Koyutürk M, Ray S, *et al*. Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics* 2012;**28**(18):i349–55.
23. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinformatics* 2016;**17**:154–79.
24. Quail MA, Smith M, Coupland P, *et al*. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;**13**:341.
25. Magi A, Tattini L, Pippucci T, *et al*. Read count approach for DNA copy number variants detection. *Bioinformatics* 2012;**28**(4):470–8.
26. Magi A, Tattini L, Cifola I, *et al*. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* 2013;**14**(10):R120.
27. Derrien T, Estellé J, Marco Sola S, *et al*. Fast computation and applications of genome mappability. *PLoS One* 2012;**7**:e30377.
28. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;**12**(8):733–5.
29. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION$^{\text{TM}}$ portable single-molecule nanopore sequencer. *Gigascience* 2014;**3**:22.
30. Ashton PM, Nair S, Dallman T, *et al*. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* 2015;**33**(3):296–300.
31. Goodwin S, Gurtowski J, Ethe-Sayers S, *et al*. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* 2015;**25**(11):1750–6.
32. Watson M, Thomson M, Risse J, *et al*. poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* 2015;**31**:114–5.
33. Leggett RM, Heavens D, Caccamo M, *et al*. NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics* 2016;**32**:142–4.
34. Szalay T, Golovchenko JA. De novo sequencing and variant calling with nanopores using PoreSeq. *Nat Biotechnol* 2015;**33**(10):1087–91.
35. Sović I, Šikić M, Wilm A, *et al*. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* 2016;**7**:11307.
36. Kurtz S, Phillippy A, Delcher AL, *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004;**5**(2):R12.