TECHNICAL ADVANCE

Open Access

ECNano: A cost-effective workflow for target enrichment sequencing and accurate variant calling on 4800 clinically significant genes using a single MinION flowcell

Amy Wing-Sze Leung[1†], Henry Chi-Ming Leung[1†], Chak-Lim Wong[1], Zhen-Xian Zheng[1], Wui-Wang Lui[1], Ho-Ming Luk[2], Ivan Fai-Man Lo[2], Ruibang Luo[1*] and Tak-Wah Lam[1*]

Abstract

Background: The application of long-read sequencing using the Oxford Nanopore Technologies (ONT) MinION sequencer is getting more diverse in the medical field. Having a high sequencing error of ONT and limited throughput from a single MinION flowcell, however, limits its applicability for accurate variant detection. Medical exome sequencing (MES) targets clinically significant exon regions, allowing rapid and comprehensive screening of pathogenic variants. By applying MES with MinION sequencing, the technology can achieve a more uniform capture of the target regions, shorter turnaround time, and lower sequencing cost per sample.

Method: We introduced a cost-effective optimized workflow, ECNano, comprising a wet-lab protocol and bioinformatics analysis, for accurate variant detection at 4800 clinically important genes and regions using a single MinION flowcell. The ECNano wet-lab protocol was optimized to perform long-read target enrichment and ONT library preparation to stably generate high-quality MES data with adequate coverage. The subsequent variant-calling workflow, Clair-ensemble, adopted a fast RNN-based variant caller, Clair, and was optimized for target enrichment data. To evaluate its performance and practicality, ECNano was tested on both reference DNA samples and patient samples.

Results: ECNano achieved deep on-target depth of coverage (DoC) at average > 100× and > 98% uniformity using one MinION flowcell. For accurate ONT variant calling, the generated reads sufficiently covered 98.9% of pathogenic positions listed in ClinVar, with 98.96% having at least 30× DoC. ECNano obtained an average read length of 1000 bp. The long reads of ECNano also covered the adjacent splice sites well, with 98.5% of positions having ≥ 30× DoC. Clair-ensemble achieved > 99% recall and accuracy for SNV calling. The whole workflow from wet-lab protocol to variant detection was completed within three days.

Conclusion: We presented ECNano, an out-of-the-box workflow comprising (1) a wet-lab protocol for ONT target enrichment sequencing and (2) a downstream variant detection workflow, Clair-ensemble. The workflow is cost-effective, with a short turnaround time for high accuracy variant calling in 4800 clinically significant genes and regions

*Correspondence: rbluo@cs.hku.hk; twlam@cs.hku.hk

[†]Amy Wing-Sze Leung and Henry Chi-Ming Leung contributed equally to this work

[1] Department of Computer Science, The University of Hong Kong, Hong Kong, China
Full list of author information is available at the end of the article

using a single MinION flowcell. The long-read exon captured data has potential for further development, promoting the application of long-read sequencing in personalized disease treatment and risk prediction.

Keywords: MinION sequencing, Medical exome sequencing, Third-generation sequencing, Target enrichment, Ensemble variant calling

Background

Screening of single genes with traditional techniques such as Sanger sequencing [1] is tedious and time-consuming, especially in heterogeneous diseases, where variants in different genes can result in similar phenotypes [2]. High-throughput sequencing allows efficient and comprehensive screening of all clinically significant genomic positions for patients with potential genetic defects. However, obtaining sufficient depth of coverage (DoC) for accurate variant calling with whole genome sequencing (WGS) is still costly for routine clinical tests [3]. Thus, target enrichment sequencing of a subset of the genome, using medical exome sequencing (MES), for example, has become a common alternative solution. MES uses customized or commercially available capture panels that target a subset of the entire exome, covering genes and positions with clinical significance. The effectiveness of exome sequencing in variant detection of rare autosomal recessive monogenic disorder [3, 4] and diseases of high genetic heterogeneity [5–7] have been well examined. MES also requires less input DNA than WGS does [8].

Oxford Nanopore Technology (ONT) Sequencing [9–11] provides a cost-effective solution for long-read sequencing with minimal laboratory setup. A regular ONT MinION sequencing run generates approximately 12–20 Gbp data, which covers, on average, 4X to 8X of the human genome in a WGS run. ONT sequencing involves real-time generation of molecular signatures while the nucleotide polymer passes through a biological CsgG protein pore [11], which can be used for both DNA and RNA sequencing, as well as nucleotide modification detection. The simultaneous basecalling while sequencing significantly shortens the data-processing time. The technology, therefore, has diverse potential applications in genomic medicine [12].

Owing to the high sequencing error rate of ONT, it remains challenging to use a single MinION WGS run to achieve high-quality variant detection. Target enrichment resolved the issue by improving the DoC in clinically important regions. A recent development of ONT sequencing incorporated the use of CRISPR/Cas9, which can effectively capture and enrich large genomic fragments for single-nucleotide variants (SNV) and structural variant detection. However, CRISPR/Cas9 targeted ONT sequencing can be performed on only a small scale, tested on 10 loci [13]. The sequencing cost, therefore, is still high in routine clinical applications targeting over 1000 genes. In addition to the CRISPR/Cas9 enrichment protocol, ONT has developed an amplicon sequence capture protocol that can be applied to exome sequencing. The protocol can be performed with an average DoC of about 30× on whole-exome sequencing [14], which is insufficient for high-quality variant calling, especially for positions with $<30\times$ DoC. Further optimization is needed to increase the average DoC.

A couple of existing bioinformatics tools are available for preprocessing and variant detection using ONT data. The recommended data processing pipeline on the ONT proprietary analysis platform EPI2ME (https://epi2me.nanoporetech.com/) provides quality control of data based on the alignment result. The workflow is incomplete without further downstream analysis. The existing third-generation sequencing (TGS) variant calling tools, such as Medaka (https://nanoporetech.github.io/medaka/index.html), LongShot [15], and Clair [16], are designed and trained mainly for WGS data, assuming a relatively even DoC in the sampled regions. However, owing to the variation of capturing and PCR efficiency in different targeted regions, the average DoC among the captured blocks fluctuates across the genome. Some of these variant callers generate consensus sequences and perform haplotype phasing [17, 18] for error correction in their workflow, which significantly increases the runtime and lowers the sensitivity in high DoC regions. While the availability of high-depth models could improve the performance of callers at high DoC positions, this is not as sensitive as some low-depth models at positions with 100X or below.

In this study, we developed a workflow, ECNano, for accurate variant calling of MES of 4,800 clinically significant genes using a single ONT MinION flowcell. The workflow comprises (1) a wet-lab protocol for the target enrichment ONT sequencing and (2) a bioinformatics pipeline, Clair-ensemble, for subsequent variant calling. The ECNano wet-lab protocol was designed to work with the solution-based target enrichment Agilent SureSelect Focused Exome panel, which strikes a balance between panel size and obtaining a sufficient DoC for high-quality variant calling in one sequencing run. Since the average exon size is about 164 bp [19, 20], our workflow targets an average fragment size of 1000 bp, which uniformly covers the targeted exome region, as well as the close-by splice sites. Alignment quality, and therefore variant-calling accuracy, were significantly improved with long reads. Another advantage of longer reads is obtaining more even coverage in positions with nearby homopolymers and small repeats, which is often excluded in the design of the capture probes [21]. The bioinformatics pipeline Clair-ensemble adopts the fast ONT variant caller Clair [16] for variant calling. Clair can achieve more accurate and refined variant classification, including the multiple bi-allelic variants, as well as the long INDEL variants. To improve the performance of variant calling with amplicon data, Clair-ensemble applies a by-positional subsampling strategy for high DoC positions and provides the ensemble of the results of Clair using multiple models. To ensure stable performance and reproducibility, the ECNano workflow was tested with both standard reference DNA and patient samples. Good-quality, high-coverage long-read data were generated for high-accuracy single-nucleotide polymorphism (SNP) calling. The precision of INDEL calling with ONT data was also significantly improved and was benchmarked against other variant callers. The whole workflow was completed within three days. The application is therefore suitable for urgent genetic testing. Further development of the workflow is promising, extending towards effective variant phasing and trio analysis. This work is also significant in promoting the application of long-read sequencing in personalized disease treatment and risk prediction.

Method

Overview

The ECNano workflow, comprising a wet-lab protocol and bioinformatics analysis, can be completed within three days, including target capture, target enrichment, ONT library preparation, MinION sequencing, data pre-processing, and variant calling. Our protocol is optimized for the use of SureSelect^{XT} Focused Exome (Agilent, Santa Clara, CA, USA), which targets approximately 17 Mbp positions. Other custom panels of similar target size are also expected to be applicable. The summary workflow within the three-day timeframe is illustrated in Fig. 1 and a step-by-step protocol is available in the Additional file 2.

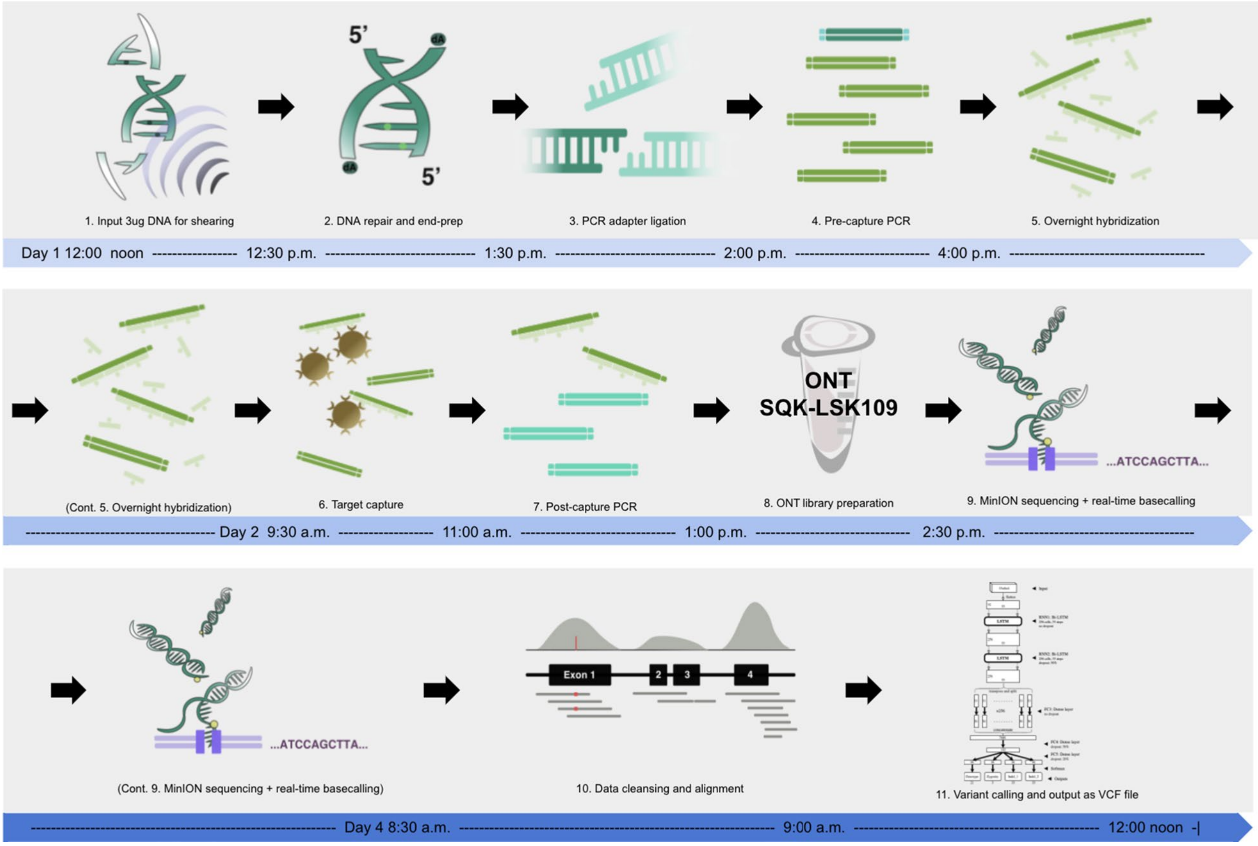
5'
dA
dA
5'
1. Input 3ug DNA for shearing
2. DNA repair and end-prep
3. PCR adapter ligation
4. Pre-capture PCR
5. Overnight hybridization
Day 1 12:00 noon ---------------- 12:30 p.m. ---------------------------- 1:30 p.m. ------------------------------ 2:00 p.m. ------------------------- 4:00 p.m. -------------------------------------
ONT
SQK-LSK109
...ATCCAGCTTA...
(Cont. 5. Overnight hybridization)
6. Target capture
7. Post-capture PCR
8. ONT library preparation
9. MinION sequencing + real-time basecalling
----------------------------------- Day 2 9:30 a.m. ----------------- 11:00 a.m. ---------------------------- 1:00 p.m. ---------------------------- 2:30 p.m. -------------------------------------
...ATCCAGCTTA...
Exon 1
2
3
4
(Cont. 9. MinION sequencing + real-time basecalling)
10. Data cleansing and alignment
11. Variant calling and output as VCF file
-- Day 4 8:30 a.m. -- 9:00 a.m. -- 12:00 noon -|

Fig. 1 Overview of the ECNano workflow diagram. The workflow comprises both wet-lab and bioinformatics components and was completed within three days (Day 1 noon to Day 4 noon)

Wet-lab protocol of ECNano

The patient DNA samples used in this study were extracted from EDTA blood samples. The automated DNA extraction was performed using the Promega Maxwell® RSC Instrument (Promega, Madison, WA, USA) with the Promega Maxwell® RSC Blood DNA Kit. The pure standard DNA sample HG001 and HG002 were purchased from Coriell Cell repositories. In our protocol, several experimental procedures and parameters were optimized based on the ONT sequence capture protocol using the SQK-LSK109 ligation kit for the SureSelect^{XT} Focused Exome capture panel as follows:

(1) *DNA preparation (Additional file 2: Sect. 1–3)*: The protocol requires 3–3.5 µg of HMW DNA as input. DNA is sheared and size selected to retain approximately 1000 bp fragments for DNA repair and end-prep.

To obtain a stable protocol that gives sufficient throughput using a single MinION flowcell for variant calling, we tested the workflow on a target fragment length of 300 bp, 700 bp, 1000 bp, and 2000 bp. The sequencing throughput is optimal and was therefore fixed at the fragment length of 1000 bp. Based on the target fragment size, a DNA purification procedure after each reaction was performed using 0.6X Agencourt AMPure XP beads and washed with 80% ethanol to retain the most DNA fragments at about 1000 bp.

We tested the feasibility of using Covaris microTUBE AFA Fiber Snap-Cap with the M220 Focused-ultrasonicator as a better alternative for DNA shearing compared to DNA fragmentase. The input DNA amount was tested with approximately 3.5 µg, which provided spare DNA for the construction of extra libraries if needed. It is also possible to use less DNA as input (2–3 µg) for downstream processing if materials are limited. The DNA volume was adjusted to 130 µl Tris-HCl (pH 8.0) and was sheared into 1,000 bp fragments using Covaris microTUBE (Covaris, Inc., Weburn, MA, USA) with the following settings: 20 °C, 50 W peak incident Power, 2% Duty Factor, 200 cycles per burst, and 20-second treatment time. Optional validation steps with QIAxcel Advanced System (Qiagen, Hilden, Germany) or a run on 1% TBE-agarose gel electrophoresis can be done to check the quality of shearing and AMPure XP beads clean-up.

For the end-repair reaction, there were significantly more DNA fragments and thus more free-ends to be repaired in a solution of 1000 bp DNA compared with those of ultra-high molecular weight DNA (i.e., > 10 kbp). For the DNA repair and end-prep step, the reaction was incubated at 20 °C for 10 mins and subsequently at 65 °C for 10 mins on a thermocycler, instead of 5 mins each in the ONT sequence capture protocol. The increase in incubation time improved the yield of end-repaired DNA, with larger amounts of fragments.

(2) *Target capture and enrichment (Additional file 2: Sect. 4–8):* The end-repaired DNA is ligated with the ONT PCR adapters to perform genome-wide amplification before overnight target capture using the Agilent Sureselect XT Focused Exome RNA probes. The captured products are amplified to obtain enough DNA for ONT library preparation.

The ECNano protocol uses the Agilent SureSelect^{XT} Focused Exome (Cat no. 5190-7787) capture panel with SureSelect TE Reagent kit, PTN (Cat no. G9605A) for target capture and enrichment. The target enrichment kit was originally designed for Ion Proton sequencing, which is more suitable for our target fragment length than those designed for Illumina sequencing. Most of the steps were conducted following the manufacturer's protocol for probe library size over 3 Mb.

For the pre- and post-capture PCR reaction, we tested the performance of Tks Gflex™ DNA Polymerase (Takara Bio Inc., Japan) and LongAmp™ Taq 2× Master Mix (New England Biolabs, Ipswich, MA, USA) for 1000 bp fragment amplification. The use of LongAmp™ Taq with PRM primers included in the ONT extension kit EXP-PCA001 (ONT) performed better with more PCR products generated under the same number of amplification cycles. The quantification of the PCR product generated was evaluated using Qubit 4 Fluorometer, and visually inspected on 1% TBE-agarose gel. The PCR condition was optimized to generate enough PCR product for downstream library preparation, thus avoiding the need for further secondary PCR amplification. The settings of the PCR cycles were as follows: (1) initial denaturation 95 °C for 3 mins; (2) 14 cycles for pre-capture PCR and 17 cycles for post-capture PCR of denaturation at 98 °C for 20 s, followed by annealing at 62 °C for 15 s and extension at 65 °C for 3 mins; and (3) the final extension at 65 °C for another 3 mins.

Note that during hybridization and capture, a lower rotation speed of 1400 rpm was used during shaking incubation at room temperature on a 96-well plate mixer for long biotinylated RNA–DNA complexes to bind better on the Streptavidin magnetic beads.

(3) *ONT library preparation (Additional file 2: Sect. 9–11):* After the hybridization and enrichment steps, the product yielded approximately 1 μg purified amplicons for a standard ONT SQK-LSK109 library preparation. Sequencing proceeded for 24–48 h with a MinION flowcell. Signals were recorded in fast5 files by MinKNOW (ONT), ready for base-calling and downstream analyses.

After we completed our experiments, ONT released an upgraded ligation kit, SQK-LSK110. The new kit is advertised as an in-place replacement for SQK-LSK109. Our protocol can use the new kit without any changes, and we expect to see a similar or better read length distribution and on-target rate using the new kit.

Bioinformatics workflow of ECNano

The bioinformatics workflow and test data are available at GitHub (https://github.com/HKU-BAL/ECNano). We benchmarked the performance using existing ONT variant callers, including LongShot, Medaka, and Clair. We tested, but excluded, PEPPER (https://github.com/kishwarshafin/pepper) from our benchmark because of its low running speed and lack of model support for exome sequencing. While Clair performs the best in terms of sensitivity and precision, the calling time is significantly shorter than the others. We therefore further optimized a variant detection workflow, Clair-ensemble, for the ECNano exome capture data. The bioinformatics workflow could also be applied to other ONT amplicon data with uneven depth distribution among the targeted regions. The major adaptations are as follows:

(1) *Data pre-processing*: The fast5 signals generated were real-time base-called using ONT basecaller Guppy v.3.4.4 using HAC config during sequencing. In addition, homopolymer correction and a minimum q-score of 3 were set for base-calling. The adapters in reads were trimmed using Porechop v0.2.1 (available at https://github.com/rrwick/Porechop). Reads with middle adapter identified were removed or split, based on a user-defined setting to avoid possible chimeric reads that might cause false positives during alignment. Reads were aligned to human reference genome hg38 (GRCh38) using minimap2 with the ONT genomic read alignment setting [21].

(2) *By-position resampling*: To ensure only high confidence alignment was retained for variant calling, only primary alignments with mapping quality of 60 or above were retained. Instead of retraining the model using capture data with different sequencing depths, we proposed a more flexible method of resampling the sequencing data into multiple datasets for each high-depth region. The resampling function is a customized partitioning method embedded in Clair. During the initial data pre-processing steps, the user (1) can specify the maximum depth per position within the target BED region, (2) can specify the maximum number of partitions to be subset for positions above the specified maximum depth, and (3) can apply optional base quality filtering at a subsampled position before variant calling when resampling is applied. The partitioning is mostly non-overlapping. If the last partition has a smaller number of alignments than the specified depth, extra alignments can be resampled from the total reads until the required depth is reached. To preserve the sensitivity of variant calling in low-depth positions (e.g., below 10x), resampling and the base-calling filter are not applied if the depth is below the maximum depth cut-off value, which can therefore minimize the information lost, especially in low-depth regions before variant calling. In our benchmarks, we set the maximum depth at 100x, with a maximum of 5 partitions in positions that required downsampling and set the base quality cut-off at q-score 5.

(3) *Ensemble variant calling*: Variant calling on each partition of candidate positions was performed using Clair [16], with four models trained on different combinations of ONT WGS reference data. The details of each model used are listed in Table 1. Clair performed individual classification tasks on genotype, zygosity, INDEL length per haplotype, and output as a probability. Variant calling was performed in each of the candidate variant positions within the MES bed region using these models to obtain the probability of each classification task. To ensemble the classification results of the four models, the probability of each task per called position was averaged across partitions and models used to provide the most robust probability calculation. The averaged probabilities were used for post-processing of Clair to decide on the variant calling output.

We tested the alignment downsampling performance using VariantBAM [22] with maximum coverage option (i.e., m option) and SAMtools random subsampling (random seed set using -s option). For variant calling, we evaluated the performance on ECNano target enriched ONT data using LongShot, Medaka, and Clair. The benchmarking was done following the standard described in detail in Clair. The output was generated

Table 1 Information on the four variant calling models of Clair used in Clair-ensemble (CE) for variant detection of ECNano data

Model	Training set	Specifications*
CE1-3-4	hg001 + hg003 + hg004 ONT data	Normal depth WGS model
CE1-2-2HD-3-4	hg001 + hg002 + hg002 very high depth (up to 500x) + hg003 + hg004 ONT data	Very high depth and normal depth WGS mixed model
CE1-2	hg001 + hg002 ONT data	Normal depth WGS model
CE1-2-3-4	hg001 + hg002 + hg003 + hg004 ONT data	Normal depth WGS model

The models were trained with combinations of different ONT WGS datasets (Additional files in Ref. [16]). * Normal depth models were trained with datasets of maximum 168 × DoC, and the model with mixed high depth data included a dataset of maximum 578 × DoC

in standard VCF format, which can be used as input for other downstream processing, such as variant phasing.

Results

Reference DNA and clinical samples for performance evaluation

The wet-lab protocol of ECNano was tested using the standard HG001 and HG002 DNA samples, which are commercially available, ensuring that the performance assessment of the wet-lab protocol was unrelated to the quality of input DNA. In addition, these two datasets are among the best-annotated human references, which allows precise variant calling evaluation of Clair-ensemble using ECNano data. To assess the reproducibility of ECNano workflow in actual clinical practice, ECNano was tested on three in-house clinical samples that require detection of different variant types and genotypes. The patients were diagnosed, and disease-causing pathogenic variants of the samples were previously identified using the next-generation sequencing approach. In all these six sequencing runs using the same ECNano protocol, we observed similar performance in terms of read length distribution, coverage profile at target positions, total throughput, and the precision and recall of small variant calls, which are the primary measurements of the protocol. Thus, we confirmed that our workflow, which is sensitive to pathogenic variant detection using clinical specimens, is robust and reproducible.

Stable, high-quality performance in long-read target capturing for MinION sequencing

Among the various target enrichment methods, using RNA or DNA biotinylated probes for hybridization-based target capture is stably applied with short-read sequencing for genetic screening in clinical genetics laboratories. The method is standardized and robust, and therefore highly reliable. ECNano therefore integrates hybridization-based medical exome capture with MinION sequencing, which can be easily incorporated into existing clinical practice as a stable enhancement of the genetic screening workflow. To ensure the workflow was very steady, we tested the balance between the capture efficiency of Agilent SureSelect Focused Exome probes and fragment length. The workflow was optimized to capture input DNA fragments of approximately 1000 bp. After MinION sequencing, the N50 of the captured reads was 1378 bp, and the average read length was 968 bp after minimum quality filtering and adapter trimming (Fig. 2).

Per-base accuracy is often constrained by sequencing technology. Low-quality reads with a mean Phred score below 3 were removed during base-calling. The average read Phred score in the tested sequencing library was Q11, and the best was Q24. Over 70.2% of the reads achieved Q10. Although the highly accurate base-calling model of base-caller Guppy was used, the error rate of ONT reads was still high, so conducting accurate variant calling is still challenging if the DoC is not high enough. The DoC per position deviated from the sequencing throughput. Depending on the quality and number of active pores available in the flowcell used, each sequencing run is expected to yield minimum 10Gbp throughput. Even with a sequencing run of only 10 Gbp throughput, ECNano achieved DoC at target positions of 133x, on average. About 98% of the target positions listed in the panel were covered by at least 30 reads, about 87% of the positions had 60× DoC, and about 55% of the positions had 100× DoC (Fig. 3). With long ONT reads, over 96% of reads could be uniquely mapped to the reference genome for variant calling. Out of the 72,417 target regions within the MES bed, on average, only 17 regions did not have any read covered. Most of these regions had over 80% GC content, so they might be difficult to capture.

The margins of the target exon were uniformly covered by long reads of ECNano. The uniformity of the sequencing data was over 98% in the target regions. For adjacent exons with a short intron between them, our data showed a broader covered region, and the intron positions were also covered (Fig. 4). This allows ECNano data to have high discovery potential for

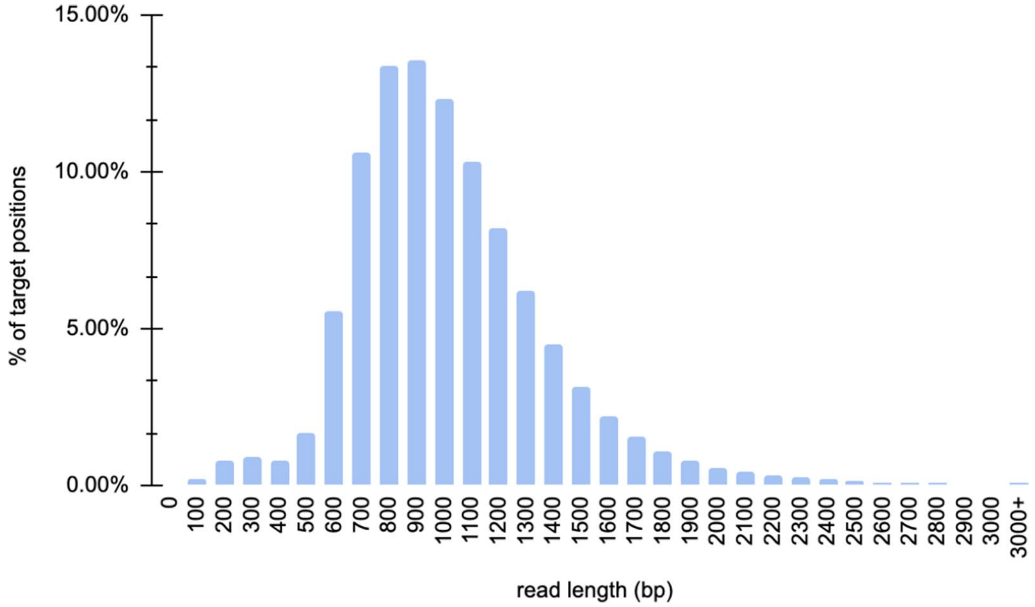

Fig. 2 Read length distribution of an ECNano target captured HG001 library, sequenced using an ONT MinION sequencer with a single flowcell with 10 Gbp throughput

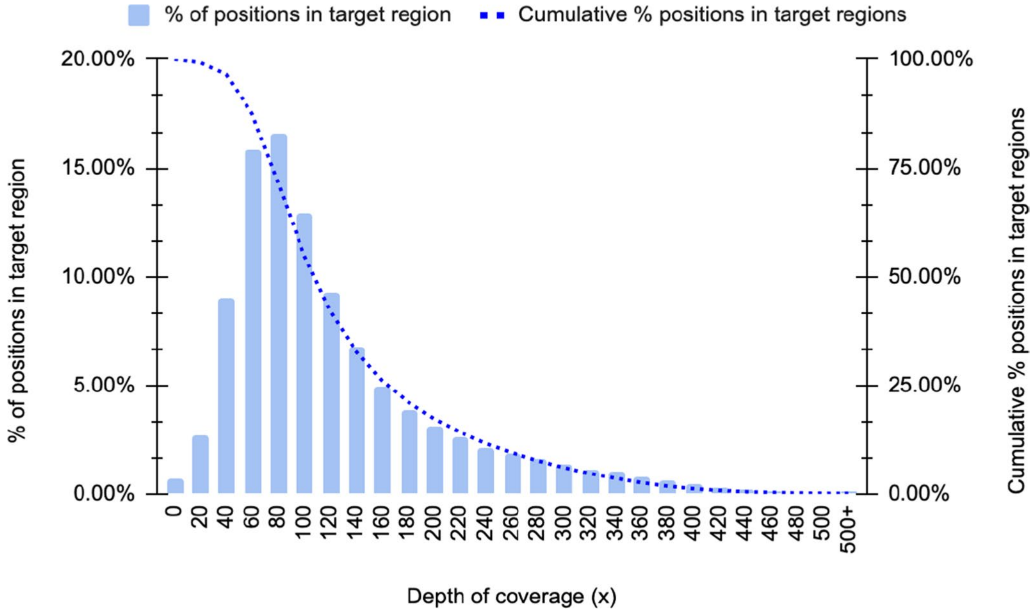

Fig. 3 Depth distribution of the ECNano target positions in a sequencing run using standard reference HG001 samples. Positions with 30X coverage or above are considered to be highly confident for variant calling

novel pathogenic variant detection in a broader range around these medically relevant positions. Although there were still considerable deviations in the DoC in the captured regions owing to variation in capture efficiency and PCR bias, in-depth normalization was well performed by Clair-ensemble to achieve precise variant calling.

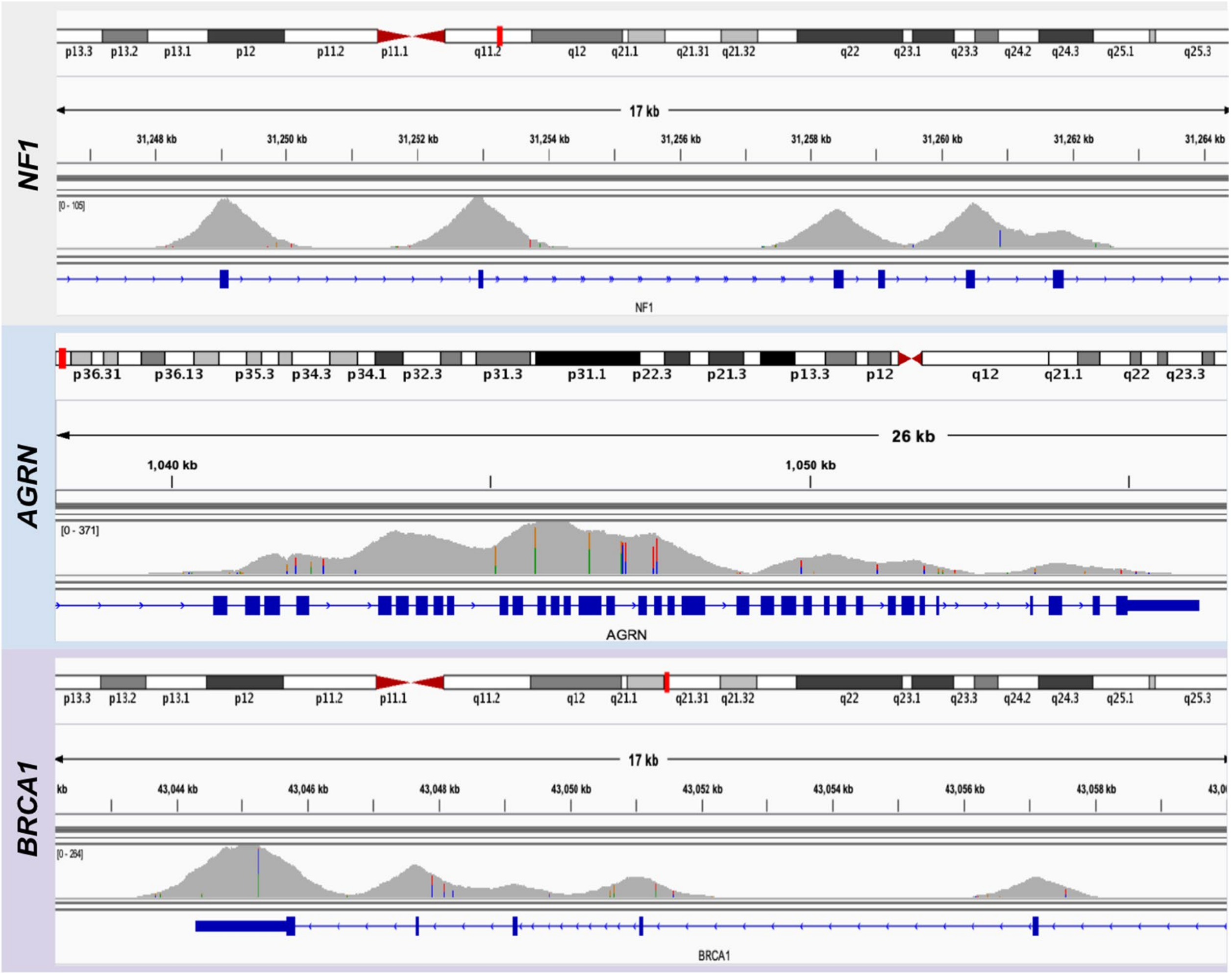

Fig. 4 Sample depth distribution comparison of target regions in three genes: (upper: *NF1*; middle: *AGRN*; bottom: *BRCA1*) in HG001 with ECNano target enrichment protocol

Sufficient DoC is guaranteed at known pathogenic variant positions and nearby splice sites for accurate variant calling

Known pathogenic variant positions and nearby splice sites are most relevant to medical diagnosis. Since most of the recorded pathogenic missense variants were found in protein-coding sequences [23, 24], screening variants in exome regions is sufficient for predicting a large proportion of genetic diseases. It is essential to maintain sufficient DoC at these positions to guarantee that the variant detection is highly sensitive. To evaluate the performance of target enrichment at the known pathogenic variant positions, we examined 23,866 positions in the targeted region of ECNano, which are listed in the ClinVar database as pathogenic using one of the HG001 sequencing runs. Over 98.9% of these positions achieved above 30× DoC, which allows highly confident variant calling at these critical positions.

Since the average size of an exon is 164 bp [19, 20], a library with a read length of over 1000 bp can cover the entire exome region, as well as potential variants close to the splice sites. In total, 2,114,539 positions of potential splicing donors and receptors adjacent to the target region were extracted from the intropolis database [25]. The referenced study annotated potential donor and acceptor splice sites on the reference genome using 21,000 NGS transcriptome data. Over 90% of these positions (± 10 bp included) had at least 60 × DoC, and 98.5% had at least 30 × DoC. This coverage allows accurate identification of canonical pathogenic splice sites and their nearby variants, which are known to cause some monogenic disease and developmental disorders [26, 27].

Accurate variant calling and performance enhancement with Clair-ensemble in target enrichment datasets

To ensure consistent and fair evaluation of variant-calling performance, Clair-ensemble was evaluated using the two sets of standard HG001 and one set of HG002 ECNano target-enrichment data against the known variants in the GIAB truth set (Table in Additional file 1). Following the best practice in Luo et al. [16], positions in high confidence regions defined by GIAB were used for benchmarking. In total, 14,976,390 positions were included in the benchmarking, which was approximately 84% of the Agilent BED positions. We benchmarked the performance of Clair-ensemble against the original Clair, LongShot, and Medaka. They are listed by precision and efficiency in Table 1. In addition to different variant callers, another downsampling method using VariantBam before variant calling was tested.

All the evaluated tools, including the original Clair, LongShot and Medaka, are capable of SNP detection. With our high DoC ECNano data, the SNP calling accuracy of these tools does not significantly deviate from each other. However, the sensitivity is inferior with Medaka, in addition to its lengthy calling time, possibly because of the use of consensus calling. The performance of Indel prediction with Medaka was also dissatisfactory. The performance of LongShot and the original Clair with low-DoC WGS models was similar in SNP calling. However, LongShot does not provide a function for Indel detection. Although we still included LongShot in our benchmarking, it is not recommended in the ECNano workflow for clinical context since it is functionally incomplete. Among all the benchmarked tools, Clair-ensemble achieved the best precision and sensitivity, achieving an F1-score in the overall calling results of over 98%. Both the built-in downsample option and VariantBam were tested in Clair-ensemble. The calling results were similar with both resampling methods, with the built-in resampling method having slightly higher precision and overall performance. However, the use of the built-in method is preferred since no extra time or disk storage is required. It is expected that with runs of even higher throughput, the built-in resampling method will be much more effective than using VariantBam because of the increase in DoC variation with regions. The performance of the ECNano bioinformatics workflow was significantly better than the original Clair, especially for INDEL calling (Fig. 5; Additional file 1: Table S1). Accurate INDEL prediction with ONT data is difficult owing to random base shifts during base-calling. Although the sensitivity remained at a similar level, the ensemble method improved the precision in INDEL calling, which is preferred in clinical applications.

The performance of neural networks is biased towards the properties of its training dataset. These variances among callers or models are reduced with the ensemble method, achieving a more stable performance across datasets. Including multiple outputs for an ensemble, however, is time-consuming if the caller runtime is long. In a comparison of the calling speed, a single Medaka run on ECNano data took days, while the original Clair required only hours. The short processing time of Clair allows an ensemble of multiple calling results without requiring too much time.

Another important feature of Clair-ensemble is the per-positional resampling as the pre-processing function alongside the ensemble. The resampling function is implemented in a Clair pre-processing step, which ensures the calling DoC does not exceed the maximal allowed depth in the ultra-high DoC region, while no resampling is performed in regions with optimal or low DoC. Compared with the use of other global downsampling methods, such as VariantBam and Samtools, this is particularly effective in avoiding over-downsampling in low DoC regions. A huge variation in DoC is commonly observed in other target enrichment data and is also highly applicable for processing with Clair-ensemble. The resampling process in Clair-ensemble also preferentially retains higher quality bases for variant calling, and therefore further improves the precision of variant calling.

Practical application of ECNano on real patient samples

Since the whole wet-lab protocol optimization and variant-calling performance evaluation was completed with standard DNA samples, we also applied the complete workflow using three patient DNA samples to ensure that ECNano is practical in actual clinical settings. All samples were first sequenced using Illumina NGS whole-exome sequencing, and the pathogenic variant was known. The three patient samples involved different types of variants (shown in Fig. 6), including (1) a 10-base insertion in BCAP31; (2) a homozygous C > T SNP in SLURP1; and (3) two heterozygous C > T and T > C SNPs in UROC1. Using the standardized ECNano workflow, we obtained over 10 Gbp of base-called throughput with a single MinION flowcell and identified the target variants unambiguously. One of the three tested samples had a 10-based

(See figure on next page.)

Fig. 5 Performance of Clair-ensemble against other existing ONT variant callers at target positions using an ECNano HG001 dataset. The performance was evaluated in terms of overall (top), SNP (middle), INDEL (bottom). Clair-ensemble was performed with both the built-in resampling method and down-sampling with VariantBam. Other tools evaluated included the original Clair with different ONT models (model details described in Table 1); and LongShot for SNP calling

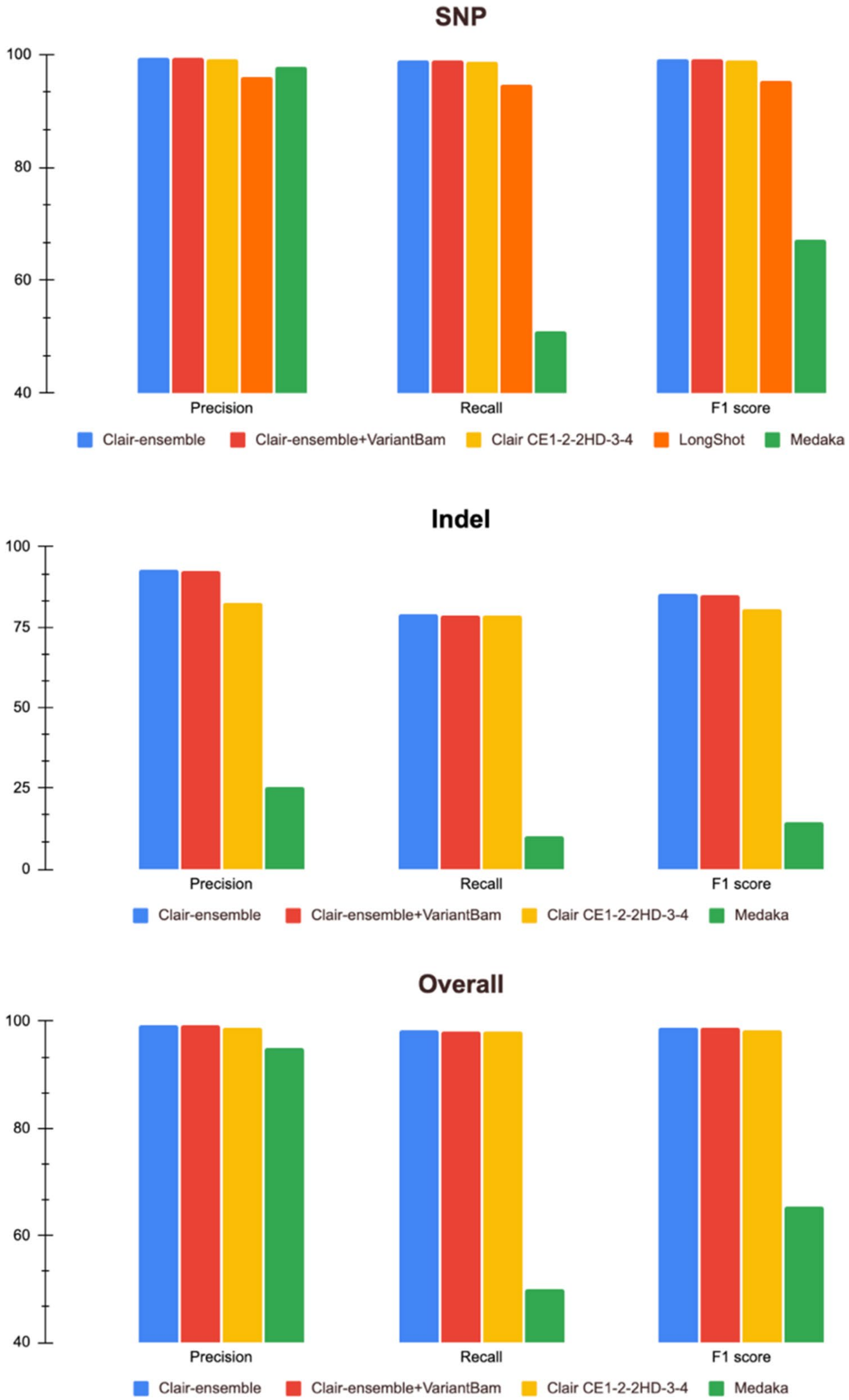

Fig. 5 (See legend on previous page.)

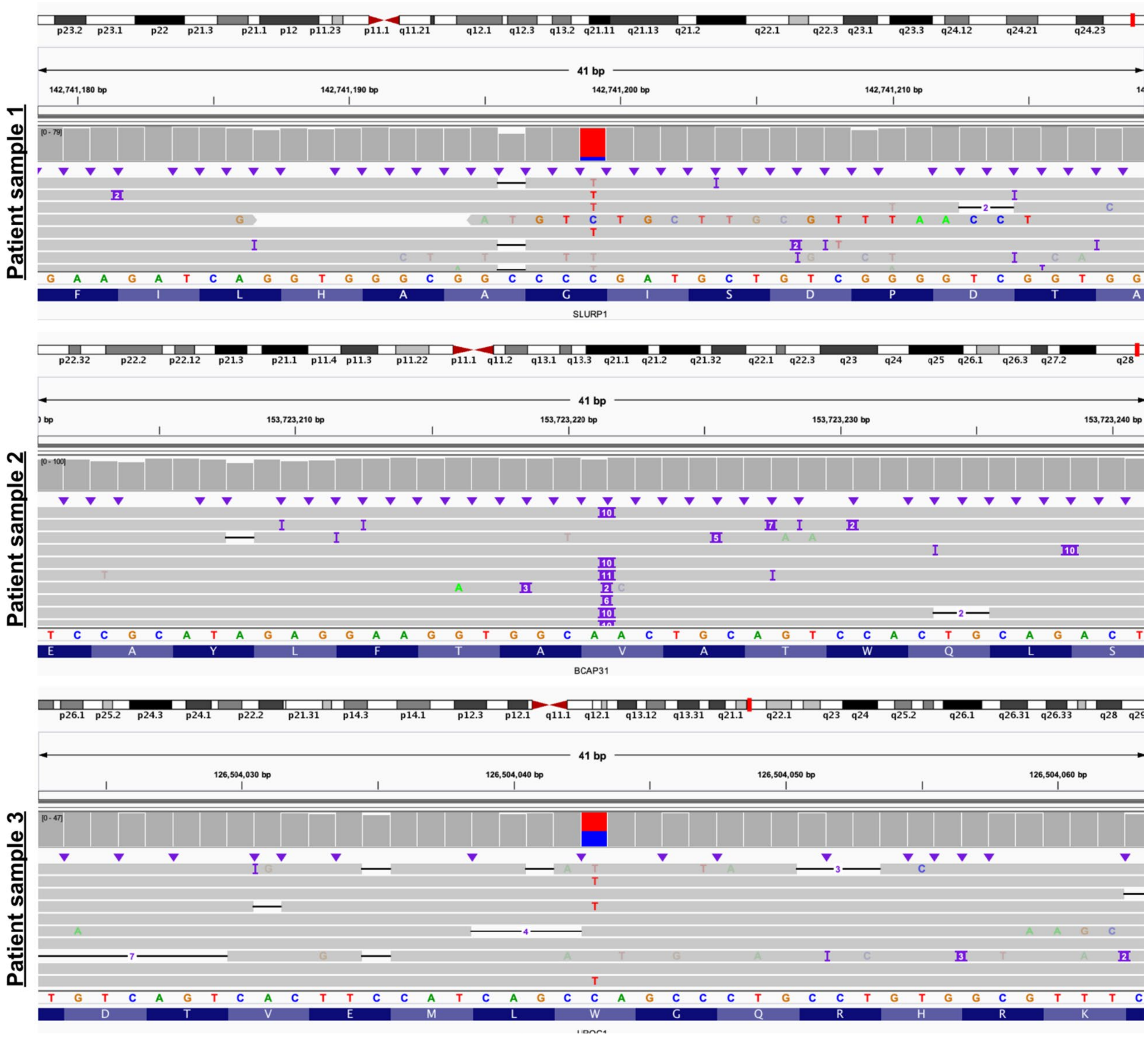

Fig. 6 Middle position showing the target variant(s) in 3 patient samples and the alignment of the adjacent positions: patient sample 1 (top) with homozygous C > T SNP in *SLURP1*; patient sample 2 (middle) with 10-base insertion in *BCAP31*; and patient sample 3 (bottom) with heterozygous C > T SNP in *UROC1*

duplication prediction. Since the precision and sensitivity of INDEL-calling with ONT data is less promising than SNP-calling, the variant could still be called with high confidence. These test set results confirmed the robust discovery power of ECNano in actual clinical use.

Comparison to medical exome sequencing using the NGS approach

To evaluate the robustness of ECNano relative to other existing sequencing, we compared the reads coverage of ECNano with a published illumina NGS dataset generated using the same target capture panel [28]. While both ECNano and NGS generated less than 1% of undetected target regions (i.e. 0.35% of the target genes or regions in NGS and 0.56% in ECNano dataset with less than 20 reads covered), the DoC distribution of ECNano is of much higher evenness compared to that in NGS (Additional file 3a). We therefore concluded ECNano is able to achieve the same if not even better depth distribution for variant calling as in NGS with the same target enrichment method.

Discussion

ONT MinION sequencing facilitates long-read sequencing at lower cost and with short turnaround time by real-time base-calling. Initial checking for the variant was often observed within the first three hours of sequencing. Users can therefore flexibly decide their sequencing runtime based on current sequencing depth with reference to real-time read distribution. Medical applications of large panel target enrichment have high discovery power, providing comprehensive screening of all medically important genes. Compared with other multiple loci variant detection methods, such as DNA Array CGH (aCGH), MES has higher resolution and can potentially identify novel pathogenic positions via a population-wide association test and trio analysis. The sequencing cost, disk storage, and computational requirement of MES are also greatly reduced compared with whole genome sequencing (WGS), as the size of most-targeted sequencing panels shrinks below 5% of the whole genome [29]. In this study, we successfully integrated these two promising technologies together in a complete workflow, ECNano, for clinical use, with high-quality performance. We carefully selected the Agilent capture panel, which is a large panel, approximately 17 Mbp in size, with medically important regions listed in three variant databases. Compared with other existing ONT target enrichment methods [30], ECNano uses hybrid-capture target enrichment for many genes, is more stable and conventional, and is easier to incorporate into routine screening practices in clinical genetics labs. ECNano also provides a complementary bioinformatics workflow, Clair-ensemble, which is designed to process amplicon data. By adopting the original Clair, which performs rapid variant calling, Clair-ensemble improves the calling accuracy without a significant trade-off in the processing time. The whole ECNano workflow can be completed within 72 h. This allows genetic diagnosis with ECNano to be rapid and precise.

Compared with the ONT long-read sequencing, regular NGS Illumina sequencing generates much shorter fragments of around 75–300 bp after target capture, depending on the platform applied for sequencing. The short reads generated suffer from misalignment, especially for those aligned to pseudogenes, genes with multiple paralogs, segmental duplication and repeat regions [31]. Since the ONT reads are longer than the NGS reads, more reads could be uniquely mapped on the reference genome for effective variant calling compared with using NGS data. Compared with the NGS data, ONT reads are covered better in the margins of the target regions. For target exons that are close to each other, the ONT reads showed more uniform coverage over a broader region and covered positions in intron regions (Additional file 3b). Due to the increased read length and coverage adjacent to target regions, this could potentially facilitate phasing of variants within a gene. Although there was still considerable deviation in depth within the target regions due to variation in capture efficiency and PCR bias, normalization in depth was done in bioinformatics analysis of ECNano to achieve more precise variant calling.

There are areas yet to be explored in the application of the ECNano workflow, in addition to accurate variant calling. With such a high sequencing depth, intermediate size structural variance (SV of > 50 bp up to 2000 bp) with a precise breakpoint within these exon regions or close by can be detected in principle [32]. The performance has been found to be poor using short reads [33]. Another potential application for ECNano sequencing is for variant phasing by region. Haplotype phasing allows more precise classification of the genetic configuration to better predict disease severity [34]. With ECNano long reads, more of these variants within individual target gene regions can be unambiguously phased.

Conclusion

We presented a complete workflow, ECNano, for MinION sequencing of 4800 clinically important genes and regions, including an optimized wet-lab protocol and a complementary bioinformatics pipeline, Clair-ensemble, for data processing and variant calling. In addition to the advantages of both hybridization-based target enrichment and long-read MinION sequencing, ECNano stably delivered high-quality results with a short turnaround time. Clair-ensemble allows accurate SNP calling by overcoming the ONT sequencing error and the uneven DoC among different captured regions. The long-read data has potential for further downstream analysis, such as variant phasing and intermediate size SV detection.

Abbreviations

aCGH: DNA array CGH; DoC: Depth of coverage; HMW DNA: High molecular weight DNA; MES: Medical exome sequencing; ONT: Oxford Nanopore Technologies; SNP: Single-nucleotide polymorphism; SNV: Single-nucleotide variants; SV: Structural variance; TGS: Third-generation sequencing; WGS: Whole genome sequencing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-022-01190-3.

Additional file 1: The performance of Clair-ensemble at target positions using two HG001 sequencing datasets and one HG002 sequencing dataset against the GRCh38 reference genome. The benchmarking results were compared against the original Clair settings with different models, LongShot and Medaka, using ECNano data. LongShot does not provide an INDEL calling function.

Additional file 2: A step-by-step wet-lab protocol of ECNano library preparation including the detailed procedures in (1) input DNA preparation,

(2) DNA fragmentation and size selection, (3) DNA repair and end-prep, (4) amplification and target capture, and (5) ONT library preparation.

Additional file 3: The comparison in depth of coverage (DoC) between medical exome sequencing using ECNano ONT protocol and NGS Illumina NextSeq 500. The NGS data was obtained from a published sequencing run by Pengelly et al., 2020. For a fair comparison, random downsampling of the NGS reads was performed so that both runs have approximately 10Gbp total throughput. (B) Example depth distribution comparison of target regions in three captured genes: (i.e. AGRN, BRCA and NF1) between ECNano ONT (upper tracks) and NGS (lower tracks) sequencing using HG001 standard DNA samples.

Acknowledgements
We thank laboratory staff at the Clinical Genetic Service of the Department of Health (Hong Kong) for extracting DNA and performing NGS for the experiments.

Authors' contributions
HCML, RL, TWL, HML and IFML conceptualized the study and revised the manuscript. AWSL and HCML designed the study and drafted the manuscript. AWSL performed the experiments. HML and IFML clinically examined the patients and collected clinical data and specimens. CLW, ZXZ and WWL developed the bioinformatics workflow. AWSL, HCML and ZXZ analyzed and interpreted the data. All co-authors have read the manuscript and agreed with its content. This manuscript was revised by all authors. All authors read and approved the final manuscript.

Funding
This work was supported by the Hong Kong Innovation and Technology Fund (ITF) Grant ITS/331/17FP, Hong Kong Research Grants Council Grant GRF 17113721, and Shenzhen Municipal Government General Program JCYJ20210324134405015. The funding body had no role in the design nor the collection, analysis, and interpretation of the data, nor the writing of the manuscript.

Availability of data and materials
To ensure patient confidentiality, data containing potentially identifiable information was not released. Raw fast5 data of ONT sequencing data obtained in our study is available from the corresponding author on reasonable request. The bioinformatics workflow, FASTQ of HG001 and HG002 datasets (also available in European Nucleotide Archive under study number PRJEB50895) generated are available at GitHub (https://github.com/HKU-BAL/ECNano).

Declarations

Ethics approval and consent to participate
The study was approved by the Human Research Ethics Committee (HREC) of The University of Hong Kong (reference no. EA210163). This was a secondary data analysis project for sequencing protocol and bioinformatics algorithm development, and informed consent was not required by the Committee.

Consent for publication
Not applicable.

Competing interests
The authors declare that there are no competing interests regarding the manuscript.

Author details
[1]Department of Computer Science, The University of Hong Kong, Hong Kong, China. [2]Department of Health, Clinical Genetic Service, Hong Kong, SAR, China.

Received: 26 February 2021 Accepted: 22 February 2022
Published online: 04 March 2022

References
1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74(12):5463–7.
2. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141(2):210–7.
3. Suwinski P, Ong C, Ling MH, Poh YM, Khan AM, Ong HS. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. Front Genet. 2019;10:49.
4. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010;42(1):30–45.
5. Gubbels CS, VanNoy GE, Madden JA, Copenheaver D, Yang S, Wojcik MH, Gold NB, Genetti CA, Stoler J, Parad RB, Roumiantsev S. Prospective, phenotype-driven selection of critically ill neonates for rapid exome sequencing is associated with high diagnostic yield. Genet Med. 2020;22(4):736–44.
6. Ilyas M, Mir A, Efthymiou S, Houlden H. The genetics of intellectual disability: advancing technology and gene editing. F1000Research. 2020;9.
7. Wang J, Wang Y, Wang L, Chen WY, Sheng M. The diagnostic yield of intellectual disability: combined whole genome low-coverage sequencing and medical exome sequencing. BMC Med Genomics. 2020;13:1–5.
8. Chen M, Chen J, Wang C, Chen F, Xie Y, Li Y, Li N, Wang J, Zhang VW, Chen D. Clinical application of medical exome sequencing for prenatal diagnosis of fetal structural anomalies. Eur J Obstet Gynecol Reprod Biol. 2020;251:119–24.
9. Rusk N. Cheap third-generation sequencing. Nat Methods. 2009;6(4):244–244.
10. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17(1):1–1.
11. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY, Popitsch N. Sequencing of human genomes with nanopore technology. Nat Commun. 2019;10(1):1–9.
12. Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr. 2020;9(2):163.
13. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, Sedlazeck FJ, Timp W. Targeted nanopore sequencing with Cas9 for studies of methylation, structural variants and mutations. BioRxiv. 2019;7:564.
14. ONT poster, 2019. Incorporating sequence capture into library preparation for MinION, GridION Mk I and PromethION. (https://nanoporetech.com/resource-centre/sequencecapture)
15. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun. 2019;10(1):1.
16. Luo R, Wong CL, Wong YS, Tang CI, Liu CM, Leung CM, Lam TW. Exploring the limit of using a deep neural network on pileup data for germline variant calling. Nat Mach Intell. 2020;2(4):220–7.
17. Martin M, Patterson M, Garg S, Fischer S, Pisanti N, Klau GW, Schöenhuth A, Marschall T. WhatsHap: fast and accurate read-based phasing. BioRxiv. 2016;085050
18. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017;27(5):801–12.
19. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27(2):182–9.
20. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111–8.
21. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
22. Wala J, Zhang CZ, Meyerson M, Beroukhim R. VariantBam: filtering and profiling of next-generation sequencing data using region-specific rules. Bioinformatics. 2016;32(13):2029–31.
23. Wiel L, Baakman C, Gilissen D, Veltman JA, Vriend G, Gilissen C. MetaDome: pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. Hum Mutat. 2019;40(8):1030–8.

24. Wiel L, Venselaar H, Veltman JA, Vriend G, Gilissen C. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. Hum Mutat. 2017;38(11):1454–63.
25. Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips RA III, Karbhari N, Hansen KD, Langmead B, Leek JT. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. 2016;17(1):1–4.
26. Ars E, Serra E, García J, Kruyer H, Gaona A, Lázaro C, Estivill X. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet. 2000;9(2):237–47.
27. Lord J, Gallone G, Short PJ, McRae JF, Ironfield H, Wynn EH, Gerety SS, He L, Kerr B, Johnson DS, McCann E. Pathogenicity and selective constraint on variation near splice sites. Genome Res. 2019;29(2):159–70.
28. Pengelly RJ, Ward D, Hunt D, Mattocks C, Ennis S. Comparison of Mendeliome exome capture kits for use in clinical diagnostics. Sci Rep. 2020;10(1):1–7.
29. de Koning TJ, Jongbloed JD, Sikkema-Raddatz B, Sinke RJ. Targeted next-generation sequencing panels for monogenetic disorders in clinical diagnostics: the opportunities and challenges. Expert Rev Mol Diagn. 2015;15(1):61–70.
30. Payne A, Holmes N, Clarke T, Munro R, Debebe B, Loose MW. Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels. BioRxiv. 2020;35:584.
31. Shigemizu D, Momozawa Y, Abe T, Morizono T, Boroevich KA, Takata S, Ashikawa K, Kubo M, Tsunoda T. Performance comparison of four commercial human whole-exome capture platforms. Sci Rep. 2015;5(1):1–8.
32. Altmüller J, Budde BS, Nürnberg P. Enrichment of target sequences for next-generation sequencing applications in research and diagnostics. Biol Chem. 2014;395(2):231–7.
33. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21(10):597–614.
34. Hoehe MR, Herwig R, Mao Q, Peters BA, Drmanac R, Church GM, Huebsch T. Significant abundance of cis configurations of coding variants in diploid human genomes. Nucleic Acids Res. 2019;47(6):2981–95.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.