# Accurate detection of single nucleotide polymorphisms using nanopore sequencing

Espada Rocio[1], Zarevski Nikola[1,2], Drame-Maigne Adele[1], and Rondelez Yannick[1]

[1]Gulliver, ESPCI Paris, PSL University, CNRS, 75005 Paris, France
[2]CRI Paris, Paris, France

## Abstract (350 words)

**Background** Nanopore sequencing is a powerful single molecule DNA sequencing technology which provides a high throughput and long sequence reads. Nevertheless, its relatively high native error rate limits the direct detection of point mutations in individual reads of amplicon libraries, as these mutations are difficult to distinguish from the sequencing noise.

**Results** We propose a computational method to reduce noise in nanopore detection of point variations. Our approach uses the fact that all reads are expected to be very similar to a wild type sequence, for which we experimentally characterize the position-specific systematic sequencing error pattern. We then use this information to reweight, in individual reads from the variant library, the confidence given to nucleotides read that do not match the wild type. We tested this method on two sets of known variants of Klen Taq, where the true mutation rate was 3.3 mutations per kb, well below the sequencing noise. We observed that the actual mutations became more distinguishable from sequencing noise after correction. This approach can be used, for example to help the clustering of variants, or to decrease the number of reads necessary to call a consensus.

**Conclusions** The computational method is simple to implement and requires only a few thousands reads of the wild type sequence of interest, which can be easily obtained by multiplexing in a single minION run. The approach does not require any modification in the experimental protocol for sequencing and can be simply implemented downstream standard base calling.

## Keywords

minION, nanopore sequencing, next generation sequencing, amplicons, SNP detection, logistic regression.

## 1 Background

Nanopore is a powerful technology for high throughput DNA sequencing, currently commercialized by Oxford Nanopore Technologies [1]. It provides sequences base calls reconstructed from conductivity records during the translocation of a single DNA molecule through a protein pore. This approach can be implemented in small devices, such as minION, which offers portability, real time sequencing, simple protocols, and a relatively low cost. A minION device can provide sequences for DNA strands of various lengths, from PCR products up to megabases genomic fragments, and provides a minimum of $5{\cdot}10^9$ bases in one run. These characteristics make it an attractive device for, among other applications, sequencing amplicon libraries that are too long for other next generation sequencing technologies. Here we will discuss the case where the amplicon library has high diversity but low variability, i.e. it contains many different sequences but differing from each other by just a few point mutations. This is the case for example, when the genetic sample originate from a single ancestral sequence (the wild type) that has been submitted to error-prone replication. In these cases, although minION can provide full length read irrespective of the size of the DNA fragments, its relatively high error rate (≈5-10%) prevents the accurate detection of point genetic variation directly from individual reads [2, 3].

Several experimental approaches have been developed to mitigate these errors and provide more accurate sequences. One possibility is to read several times each amplicon, and then construct a consensus sequence. This has been done by creating sequence concatenates, for example using rolling circular amplification, [4, 5, 6] or via gene barcoding prior to amplification [7]. These methods usually reduce the number of different variants that can be studied, because a part of minION throughput is invested in reading duplicated reads, and is sensitive to bias occuring during the amplification steps [8]. Other approaches, used for genome assembly, combine long minION reads with shorter and more accurate reads obtained via other technologies, such as Illumina sequencing-by-synthesis approach [9].

In this paper, we propose a computational protocol to improve variant detection in individual reads from libraries for which a reference gene is known, using standard 1D protocol minION sequencing. We base our method on two observations made during the sequencing of many (identical) copies of the parent sequence. First, the confidence or quality scores ($Q_{score}$) assigned by the base calling process to each nucleotide are usually low when a wrong nucleotide is assigned (Suppl. figure S2), as expected. Second, the errors are not homogeneously distributed, and they are more frequent in some positions of the DNA (Suppl. figure S1). These observations suggest that it should be possible to reduce the non-random part of the sequencing errors, using the information contained in the ($Q_{score}$). The method we propose has two steps: the first one uses the reference reads to build a statistical model of the error pattern. Here we used a position and nucleotide-specific logistic regression. In the second step, this information is used to re-analyze minION base calls for the variant library and to update the confidence value of each nucleotide read in this dataset.

We tested our method using the gene of Klen Taq DNA polymerase, a truncated variant of the well-known Taq polymerase. We trained our model using approximately 6000 reads of the wild type and applied it on two toy libraries containing 10 known variants with 2 to 9 point mutations. Our computational correction reduced the sequencing noise, allowing a better identification of true point mutations. The corrected confidence values were used to clusterize sequencing reads of the same variant with an accuracy of 96% leaving aside barcodes or other physical links. Furthermore, in the context of point mutant libraries, the corrected confidence values reduce the number of reads needed to obtain an exact consensus sequence by at least 5 times, with good performance obtained using only typically 10 reads.

## 2 Method

Like other sequencing approaches, nanopore data analysis pipeline provide reads where each base is associated with a confidence values ($Q_{score}$). This number reflects the probability that the assigned nucleotide at that position is the correct one $p_{right}$, via $p_{right} = 1 - 10^{-\frac{Q_{score}}{10}}$.

We first characterized the errors and $Q_{score}$ distributions on more than 6000 reads of the wild type Klen Taq gene (length 1665 nucleotides), for which we have a ground truth sequence (see Suppl. table S1). In this data set, we can confidently attribute mismatches between the read nucleotide and the wild type as sequencing errors. At each position, and for each of the three non wild type nucleotides and deletions, we quantified the proportion of errors made by minION and guppy base-caller or, equivalently, the proportion of correct reads (which we denominate $n_{correct}$), and grouped them according to the $Q_{score}$ provided by the basecaller. $n_{correct}$ is generally smaller for low $Q_{score}$ (figure 1A and S4), reflecting that most errors appear where guppy base-calling confidence is low (figure 1C). However, the pattern strongly differs base-to-base and position-to-position (see examples in Suppl. fig. S4). We fitted this relation position and base-wise by a logistic regression, which provided a classifier able to convert the reported $Q_{score}$ to the probability that this read is indeed correct ($p_{right}$). This was made separately for the sequences read for sense and antisense strands. To include deletions in this analysis (which do not have a $Q_{score}$ assigned by the basecaller), we fixed their confidence value as the average of the $Q_{score}$ of their nearest neighbors in the nucleotide sequence. This decision was inspired by the observation that $Q_{score}$ are correlated between consecutive nucleotides (Fig. S3). Insertions were ignored. All together, we obtained 13320=1665·4·2 regressions, one for each position and non-wild type nucleotide/deletion, in the forward and reverse sense of sequencing.

With this information in hand, we then looked at mismatches in reads originated from mutated variant sequences. For each of them, we can now use the corresponding logistic regression to compute $p_{right}$, the probability of actually being a mutation according to the associated $Q_{score}$. The higher the value of $p_{right}$, the more likely it is a true mutation.

However, the training set uses wild type sequences, which do not contain true mutations and hence our naive classifier is heavily biases against classifying mutations, as represented in figures 1A and 1C . This is not representative of the proportion of errors/correct reads present in the mismatches of a set of mutants, even if the mutation rate is low. To adapt the classifier, we need an *a priori* expectation of mutations ($p_{prior-right}$), that should come from independent information. In the present case, the variant sequences originates from an error prone PCR (ePCR) process, for which we possess an estimate of the mutation frequency. We also obtain the *a priori* expectation of an observed mismatch to be a sequencing error ($p_{prior-error}$) as the sequencing error rate at that position in the wild type set. We therefore compute $p_{prior-right}$ and $p_{prior-error}$, which we use to reweight the wild type set before fitting. This shifts the logistic regression towards the higher $Q_{score}$, allowing the classifier to accept a number of observed mismatches consistent with the prior expectation (figure 1B). We then proceed to re-score the mismatches observed in the mutant library, taking into account position, nucleotide and sense of the strand.

## 3 Results

To test our method, we built two toy libraries of Klen Taq mutants and sequenced them using the minION device (see experimental methods). The first one consists

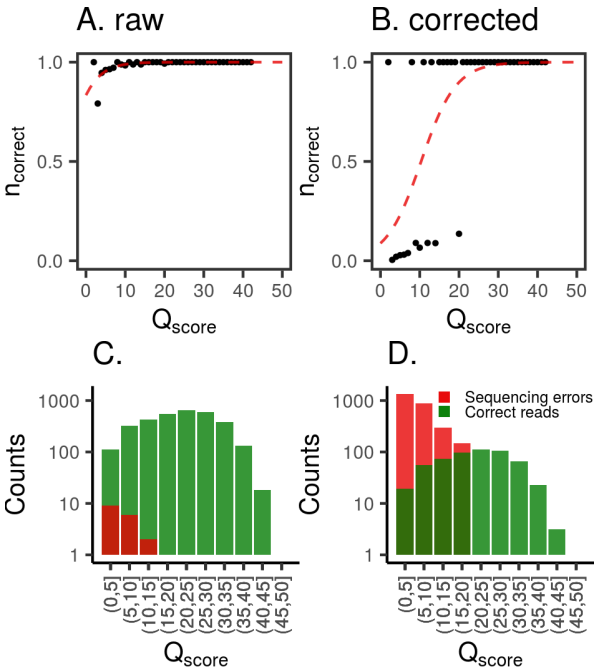


**Figure 1:** Example of logistic regression over reads of a known wild type sequence to re-scale $Q_{score}$ at one particular position and nucleotide. A: black dots are the proportion of correct reads against the $Q_{score}$ reported by guppy base-caller. Red line is the logistic regression performed over this data. C: Distribution of correct reads (green) and sequencing errors (red). B and D: Same plots as A and C, when each read is weighted according to the prior probabilities.

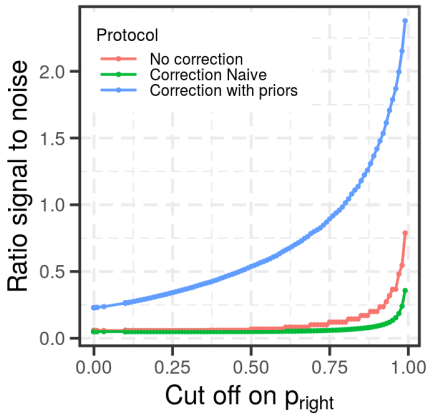


**Figure 2:** Signal to noise ratio as a function of the cut-off on the quality values $p_{right}$, for the three methods analyzed: using $p_{right}$ provided by minION (red), using $p_{right}$ after naive correction (green) and $p_{right}$ after correction with priors (blue). The last method improves the signal to noise ratio for any $p_{right}$ cut off.

of 7 known variants uniquely barcoded before sequencing, thus we know exactly what mutations to expect in each read. The second toy library is a mixture of three known variants. In this case we know which mutations to expect in the library but we do not have a mark on each sequencing read. The complete list of mutations is available in Suppl. Table S2.

We used the logistic regressions performed on the wild type data to convert $Q_{score}$ into a probability of being a true mutation the $p_{right}$ of each nucleotide, as described in section Method.

## 3.1 Signal to noise ratio

MinION's high error rate makes it difficult to distinguish actual mutations from sequencing errors in single reads, especially when the mutation rate is low. A straightforward procedure to filter errors is to only trust read position which have a high probability of being correct. In this section, we compare how it performs to carry this process using the raw $p_{right}$ returned by guppy base-caller on minION reads, the $p_{right}$ after fitting with a logistic regression (or correction naive), or $p_{right}$ after fitting with a logistic regression using prior probabilities.

We used the library containing seven known mutants to quantify the signal to noise ratio. In this example each sequence is barcoded, so we know if each mismatch is a sequencing error or an actual mutation. We defined *signal* as the number of mismatches known to be mutations with a $p_{right}$ higher than the threshold (true positives), and *noise* as the number of those mismatches known to be non mutated (false positives). The counts are weighted by $p_{right}$ for each nucleotide. Results are shown in figure 2. For all thresholds, the proposed correction has a higher signal to noise ratio, thus facilitating the identification of actual mutations. This remains true when no cut off is applied (cut off = 0). Results are similar if we count number of mismatches over the threshold without weighting them by $p_{right}$ (Suppl. fig S5). We noticed that the signal to noise ratio improvement is mainly driven by the reduction of the false positive rate, i.e. reduction of sequencing noise. Nevertheless, the true positive rate also decreases faster (Suppl. fig S6). This should be taken into account if the goal is to detect mutations which are poorly represented in the sequenced set.

## 3.2 Sequences clustering

To show that the reduction of noise is relevant to characterize an ensemble of reads into sequence clusters, we used the corrected probabilities $p_{right}$ to group reads by weighted sequence similarity. We defined the dissimilarity between two reads $s_1$ and $s_2$ as a modified hamming distance:

$$d(s_1, s_2) = \sum_{n=\text{position}} f(s_1^n, s_2^n) \qquad (1)$$

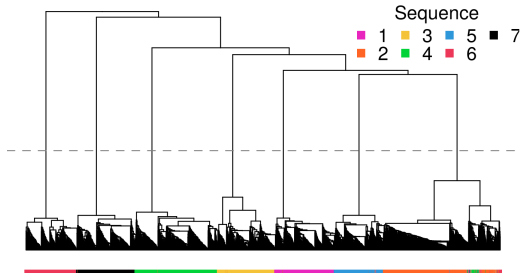


**Figure 3:** Hierarchical clustering of reads of a toy library constituted by 7 mutants of the Klen Taq gene, each of them experimentally identified by a barcode. The colors below each leaf indicate to which variant the read corresponds to. The clustering groups reads of the same variant.

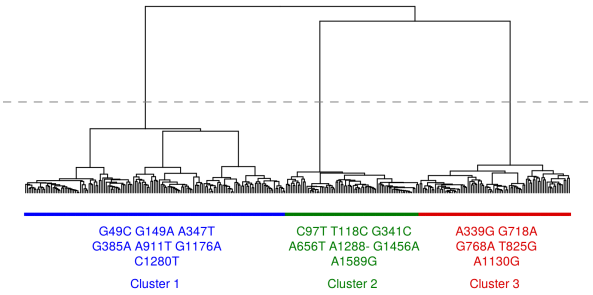


**Figure 4:** Hierarchical clustering of reads of a toy library constituted by 3 variants. Clusters are marked in colors. After clustering, we calculated the consensus sequence within each of the four groups detected. Detected mutations with respect to the wild type gene are annotated below. Each cluster correspond to a known variant.

where

$$f(s_1^n, s_2^n) = \begin{cases} 0 & s_1^n = \text{wt and } s_2^n = \text{wt} \\ +p_{\text{right}}(s_1^n)p_{\text{right}}(s_2^n) & s_1^n \neq s_2^n \\ -p_{\text{right}}(s_1^n)p_{\text{right}}(s_2^n) & s_1^n = s_2^n \neq \text{wt} \end{cases} \quad (2)$$

where $p_{\text{right}}(s_i^n)$ is $\text{p}_{\text{right}}$ assigned to sequence $i$ at position $n$. We summed zero if both sequences are equal to wild type (wt), we summed the product of corrected probabilities of both sequences if they are different (one mutation present in only one of them), and we subtracted the product of corrected probabilities of both sequences if they are the same and differ from wild type (both reads display the same mutation). Then we used R implementation of Ward's minimum variance method to clusterize the reads.

We evaluated this procedure on the toy library of 7 variants. The resulting dendrogram in displayed in figure 3, where each leaf is a different read, and the color of the point below represents which Klen Taq variant it is. When cutting the dendrogram at a fixed height to obtain seven clusters, we can find a prevailing variant on each of them, which allows us to use this as a classification method. Out of 3538 reads analyzed, 3381 were correctly clustered (96%). This was not true when using guppy base-caller's $\text{p}_{\text{right}}$ values, in which only 63% were correctly classified (see suppl. fig S7).

We also evaluated the toy library of three mixed variants (named variants 8, 9 and 10, see sequences at suppl. table S2). We measured dissimilarities between reads and obtained the dendrogram in figure 4. We cut the dendrogram in a fixed height to obtain three clusters denoted in colors in the figure. We calculated the consensus sequence for the reads belonging to the same cluster. In the figure, we annotated the mutations for each consensus. We found that all three match one of the known variant: cluster 1 is variant 8, cluster 3 is variant 9 and cluster 2 is variant 10 (in this last case, the consensus sequence adds a deletion at position 1288).

## 3.3 Consensus calculation

One strategy for lowering noise in nanopore sequencing consists in computing a variant consensus sequence (VCS), i.e. reading several times a variant and use these reads to calculate the most likely sequence [4, 10]. Usually, average on more reads leads to a more accurate VCS. We hypothesized that the corrected weights should help to converge faster on the actual sequence, as they increase the confidence on mutations over the sequencing errors. We tested this hypothesis over the set of 7 known Klen Taq variants sequenced using minION.

We computed the VCS over a set of aligned reads of the same Klen Taq variant (using the fact that these reads were barcoded). For each position, we calculated the weighted frequency of each nucleotide and deletions. The weights used are the $\text{p}_{\text{right}}$ returned by guppy base-caller, or the ones obtained in the logistic regression. The nucleotide (or deletion) with the highest weighted count is kept as the consensus in that position.

We computed the consensus on a set of 2 to 50 sequences drawn randomly, and repeated 100 times for each set size. In figures 5 and S9 (red curves) we show how many times the VCS matches exactly the actual sequence. In all cases where there are no deletions (variants 1-4 and 7), or where there is a deletion in a non-homopolymer position (variant 6), the convergence on the actual sequence is faster when using the corrected weights: perfect consensuses are more that 95% of attempts starting from typically 10 sequences. When using raw $\text{Q}_{\text{score}}$ provided by guppy base-caller this performance is not obtained in sets of up to 50 sequences. Results are different for variant 5 which has a deletion within a homopolymer ('GG' at positions 489 and 490), making it a challenging target. The basic VCS matches performed worse with corrected scores and including more sequences even reduces the correct percentage (5B, red solid and dashed curves). We noticed that this is produced by the misrepresentation of the deletion. The VCS is com-

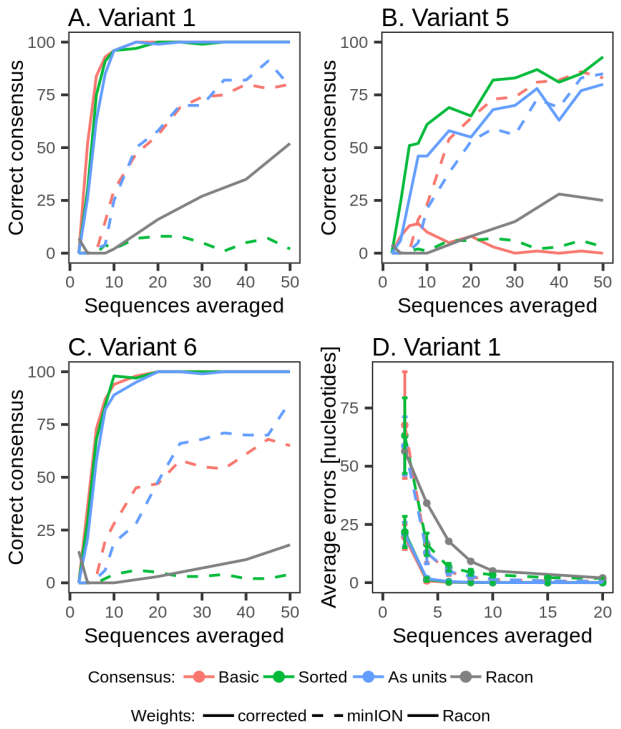


**Figure 5:** A-C Number of times the VCS matches the actual sequence according to the number of reads used to compute it. We used different strategies for computing the consensus: In red, no pre-processing of sequences. In green, sorting homopolymers to move deletions towards the end. In blue, treating homopolymers as a unit. Panel A show the results for variant 1 (no deletions), panel B for variant 5 (which has a true deletion within a homopolymer), and panel C for variant 6 (which has one deletion in a non-homopolymeric position). Panel D shows the average error per consensus sequence according to the number of sequences used. In all cases we compare the use of guppy base-caller weights (dashed line) and weights corrected by the logistic regression method (solid line). In grey, results using racon software.

puted position by position, and the aligner used (LAST) does not always represent the deletion in the same position (sometimes is 'G-' and sometimes is '-G'). Thus the presence of a deletion is averaged out. We tested two strategies to address this issue: on the one hand we tried to sort all homopolymers so that the deletions are always at the end before computing the VCS. Results are shown in green curves in fig. 5). In this case, the results for variant 5 improved, reaching a performance slightly better than using $Q_{score}$ provided by guppy base caller, while similar performance for all other variants compared to the basic consensus calculation. Notably, this strategy is conter-productive for the VCS calculation using guppy raw scores in all variants (green dashed lines). On the other hand, we tried evaluating the homopolymers as a unit instead of position by position (blue curves in fig. 5). Here again more, the results for variant 5 improved, reaching a performance as good as using $Q_{score}$ from guppy base-caller, without this affecting other variants. For a deeper discussion on this point, refer to the Supplementary material.

We also evaluated how different the VCS are from true sequences. In figures 5D and S10, we plotted the mean number of wrong nucleotides in the VCS. When using the corrected weights, it is enough to use 6 reads to obtain a mean error of less than one nucleotide in the VCS, (in all cases except variant 5). This number increases up to 10 sequences when using raw $Q_{score}$ as weights. Also, there are no significant changes among the different consensus strategies used.

As a reference, we computed consensus sequences using Racon, a state of the art tool [10] (grey curves in fig. 5, S9 and S10). In all cases, we observed that the convergence for Racon requires more reads than the methods presented in this manuscript. However, we note that Racon does not use a reference sequence, which makes direct comparison difficult.

# 4 Conclusions

Single molecule MinION sequencing comes with a relatively high error rate, which limits some applications such as the analysis of libraries containing many different, but genetically similar, sequences. This is for example the case for amplicon libraries used in directed evolution. The approach that we propose here leverage the fact that the errors are partly systematic, as previously noted [11]. We therefore accumulate many reads from the reference gene to build a sequence specific error model that can locally correct for the sequencing biases. Applying this procedure on two toy libraries of the Klen Taq gene with an average mutation rate of 3.3 bases/kb, we showed that correcting the confidence values provides large increase in the signal to noise ratio. This can be used to cluster reads and compute accurate consensus with a minimized burden on the sequencing throughput.

An important ingredient of our approach is a correct prior for the number of mutations in the train and test set. Here, our reference sequence was assumed to be perfect, and we could precisely evaluate the average mutation rate in the test set, because it originated from a controlled experimental mutagenesis protocol. In other situations it would be possible to use short read sequencing, for example Illumina, to evaluate this number. If the full sequence is submitted to short read high quality sequencing, it would even be possible to obtain more precise priors, for example specific to each position and nucleotide. Our approach would then be used to associate these mutations on single long reads.

An underlying assumption of our method is that the distribution of $Q_{score}$ observed at a particular position for the wild type base appropriately reflects the distribution

of $Q_{score}$ that would be observed for a variant base at that position. This approximation is necessary since the error model is build from a single sequence and hence has a single "true" base per position. Fortunately, the difference in $Q_{score}$ distributions for "true" versus "error" seems large enough for our method to perform well within that approximation. Moreover, when working with large variant libraries, it would in principle to use an iterative process to complete the error model using observation made on the variants.

The protocol proposed needs to characterize the sequencing errors done on appropriate reference sequence. As such it is limited to analyze variants which are close neighbors of the reference, and where mutations can be considered to be independent. We did not try to adapt the method to detect alterations beyond point replacements or deletions, which may require a more complex analysis pipeline. Encouragingly , variant 6 contained two contiguous mutations, and was properly analysed by our consensus approach. We also note that, in the case of highly diverse sequences, and if the library can be sequenced at sufficient depth, it become easier to cluster sequences and compute direct consensuses. Finally, our approach provides large improvement of signal to noise at very little experimental effort or throughput reduction. This is because only the reference DNA needs to be sequenced many times, and this can be done simultaneously with the libraries, using standard barcoding protocols. There is no other modifications to the experimental protocol, and the computational error correction process can be simply added to any analysis pipeline after base calling.

# 5 Experimental methods

## Samples preparation

Klen Taq wild type gene was amplified using a high fidelity PCR (Q5 polymerase from NEB) from a stored plasmid. See DNA sequence in Suppl. table S1.

Mutants were obtained via error prone PCR (ePCR) using Agilent's kit GeneMorph II. We started from 1.1 nM of dam-methylated DNA. We used primers GGGATTATTCTTTGGCGCTCAGCCAAT and ACCATGCGTCTGCTGCATGAAT. Thermocycling was performed as follows: 95$^o$C for 2 min, followed by 25 cycles of [95$^o$C for 30sec + 65$^o$C for 30sec + 72$^o$C for 2min] and a final extension at 72$^o$C for 10 min. We digested the product with DpnI (NEB) and purified it using columns (Macheray-Nagel). We put the mutagenized genes in a pIVEX vector via Gibson assembly (NEB Hi-Fi DNA assembly) using 125ng of gene DNA, 100 ng of vector in a 2:1 insert:vector molar ratio and incubated for 15min at 50$^o$C. We purified and concentrated DNA with a Zymo Research kit. We transformed the product into chemocompetent KRX bacteria. We spread them on a Petri dish with LB and Ampicillin. We incubated overnight and picked some clones randomly. We verified the presence of the plasmid via colony PCR (using DreamTaq polymerase from Thermofisher). 10 positive clones were grown overnight in liquid LB with antibiotic and mini-prepped to obtain the plasmid DNA. A fraction was used for high quality sequencing (Sanger sequencing), and another fraction used for minION sequencing.

## minION sequencing

We used the prepared DNA of each clone and the wild type gene for amplification by PCR with Q5 polymerase (NEB) using primers which included the minION barcodes adapters: ACTTGCCTGTCGCTCTATCTTCAGTGTGCTGGAATTCGCCCTTTTA and TTTCTGTTGGTGCTGATATTGCAGACCACAACGGTTTCCCTCTAGAAATA.

Thermocycling was performed as follows: 98$^o$C for 30sec, 23 cycles of [98$^o$C for 10sec + 59$^o$C for 30sec + 72$^o$C for 1min], final extension at 72$^o$C for 2min.

We digested the product with Dpn1 (NEB), gel purified it using Macheray-Nagel kit. We proceeded following standard minION protocols. We used one barcode for wild type, one for each of the mutants 1 to 7, and one extra barcode for mutants 8-10. We then mixed the DNA and followed standard protocol for preparing and loading the DNA into a flowcell.

We used kits EXP-PCB001 for barcoding, SQK-LSK108 for ligation, EXP-LLB001 for flowcell loading, and minION flowcell version was R9.4/FLO-MIN106, thus the sequencing was 1D.

## minION reads pre-processing

minION raw data was base called and demultiplexed using ONT Guppy version 3.3.3. We obtained 140197 reads from which 45635 (33%) did not match any barcode. Each sequence was pairwise aligned to wild type Klen Taq sequence using LAST [12]. Those sequences with low alignment score were discarded, keeping 32507 at the end (23% from original reads).

## Racon consensus

We selected a subset of sequences on the fastq file after base calling using an R script, we aligned them to the wild type sequence using minimap2 2.17-r941 [13] (default parameters), and computed the VCS with racon v1.4.11 [10] (default parameters).

## Availability

R code for the analysis is available at https://github.com/rocioespci/minIONLogReg.

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 845976 and from the European Research Council (ERC, Consolidator Grant No. 647275 ProFF).

## Conflicts of interest

The authors declare no conflicts of interest.

# References


[1] https://nanoporetech.com/.

[2] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, 15(6):461–468, 2018.

[3] Liang Gong, Chee-Hong Wong, Wei-Chung Cheng, Harianto Tjong, Francesca Menghi, Chew Yee Ngan, Edison T Liu, and Chia-Lin Wei. Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nature methods*, 15(6):455–460, 2018.

[4] Chenhao Li, Kern Rei Chng, Esther Jia Hui Boey, Amanda Hui Qi Ng, Andreas Wilm, and Niranjan Nagarajan. Inc-seq: accurate single molecule reads using nanopore sequencing. *Gigascience*, 5(1):s13742–016, 2016.

[5] Roger Volden, Theron Palmer, Ashley Byrne, Charles Cole, Robert J Schmitz, Richard E Green, and Christopher Vollmers. Improving nanopore read accuracy with the r2c2 method enables the sequencing of highly multiplexed full-length single-cell cdna. *Proceedings of the National Academy of Sciences*, 115(39):9726–9731, 2018.

[6] Szymon T Calus, Umer Z Ijaz, and Ameet J Pinto. Nanoampli-seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience*, 7(12):giy140, 2018.

[7] Søren M Karst, Ryan M Ziels, Rasmus H Kirkegaard, Emil A Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with nanopore or pacbio sequencing. *bioRxiv*, page 645903, 2020.

[8] Marc A Sze and Patrick D Schloss. The impact of dna polymerase and number of rounds of amplification in pcr on 16s rrna gene sequence data. *mSphere*, 4(3):e00163–19, 2019.

[9] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11):1750–1756, 2015.

[10] Robert Vaser, Ivan Sovic, Niranjan Nagarajan, and Mile Sikic. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, 27(5):737–746, 2017.

[11] Raga Krishnakumar, Anupama Sinha, Sara W Bird, Harikrishnan Jayamohan, Harrison S Edwards, Joseph S Schoeniger, Kamlesh D Patel, Steven S Branda, and Michael S Bartsch. Systematic and stochastic influences on the performance of the minion nanopore sequencer across a range of nucleotide bias. *Scientific reports*, 8(1):1–13, 2018.

[12] Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.

[13] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.