

Long read sequencing enhances pathogenic and novel variation discovery in patients with rare diseases

Corresponding Author: Professor Ahmad Abou Tayoun

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors describe a filtering strategy for identifying pathogenic variants from long-read sequencing data, significantly reducing the number of structural and copy number variants. After testing it on a cohort with known variants, the authors demonstrate that the applied strategy can diagnose patients with negative NGS analysis.

As was the case years ago with second-generation sequencing methods, the clinical application of long-read sequencing is held back by the huge amount of data to manage and variant calls from which to select the one or few disease-causative variants. Therefore, effective filtering strategies are needed to definitively translate technologies with enormous diagnostic potential into clinical practice.

The manuscript is very well written, and the work is certainly of interest, but it does not describe innovative applications or strategies compared to those already published in the literature with respect to long-read sequencing (e.g. studies listed in the paper by Mastroianni et al., 2023). The main novelty concerns the identification of a methylation signature specific for SMA and useful for its diagnosis. Perhaps this is the finding to be highlighted most emphatically although it should be tested on a larger number of patients with molecular diagnosis.

Following are some doubts that require clarification and some suggestions:

- Within the optimization cohort, the applied filtering strategy reduced the number of CNVs and SVs by 98 and 99.9% respectively. What about short variants (SNVs and INDELs)? Related data are reported in the supplementary table 2 but they deserve a comment within the manuscript.
 - Why is the methylation within the 15q11.2 region reduced in the OXN-068 sample which is not positive for Angelman syndrome, as confirmed by MS-MLPA? Was this simply a false positive (although with a not completely altered profile)?
 - The OXN-020 sample of the optimization cohort has a splicing variant in the HBB gene. Supplementary Table 2 shows that the filtering strategy for splicing variants was not applied in the optimization cohort. Why? How was the variant identified from the LRS data?
- Moreover, why not analyze the exonic SNVs as well? LRS could identify variants missed by exome analysis because located in GC-rich regions or characterized by highly homologous sequences.
- The two deletions prioritized from OXN-033 LRS data and filtered out after manual inspection were false positive calls not confirmed by CMA or they were considered non-clinically relevant? I think it is important that the authors also report for newly diagnosed patients the variants prioritized by the filtering strategy and then discarded, explaining in detail why.
 - For the OXN-048 sample, the size of the del16q23.3 should be reported.
 - The 3.6 kb MPLKIP deletion of patient OXN-027 should be conventionally considered among CNVs. Was it called by Sniffles and not by QDNAseq?
 - The authors suggest that MPLKIP overexpression could underlie the patient's phenotype but reported causative MPLKIP variants cause protein deficiency. How could the overexpression be responsible for the disease?
 - The authors state that the application of LRS led to the diagnosis of patients who had inconclusive testing with traditional methods. Generally, CMA is performed as a first-tier analysis on NDD patients, and in this cohort, if performed before LRS, it would likely have identified most of the CNVs detected by LRS. Why CMA was not performed on these patients, particularly those with NDD? The authors should discuss this point
 - The choice to consider only variants involving the coding region of disease-causing genes probably reduced sensitivity, especially with regard to SVs that can be causative even if the breakpoints fall in noncoding regions, through various mechanisms (e.g., position effect, alteration of TADs...). The authors should discuss this point
 - In some parts of the text ddPCR is described as double digital PCR, in others as digital droplet PCR. The correct name is droplet digital PCR.

Reviewer #2

(Remarks to the Author)

The authors present a novel genetic diagnostic framework to address the limitations of existing long-read sequencing methods. While its clinical utility is promising, there are significant issues regarding its scientific validity.

Major Concerns

1. The authors have included 10 positive control cases and 39 undiagnosed genetic cases, but the sample size is too small. Moreover, since the majority of subjects are of Arab ethnicity, there are substantial concerns about the reproducibility and generalizability of the data.
2. The authors claim that the "funnel down" filtering method efficiently reduces the number of candidate variants. However, the rationale behind the filtering criteria is unclear, raising doubts about the validity of this approach. It is essential to assess how many potentially pathogenic variants were excluded at each filtering stage and to develop an appropriate filtering strategy.

Minor Concerns

1. Page 7, Line 116: "SLC38A3" should be "SLC38A8".
2. Page 9, Line 191: "parametes" should be "parameters".

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors clarified all the doubts and answered the points I had raised. I make one final note with regard to the figures, particularly the supplementary ones, whose small size does not allow for adequate appreciation of the details

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have sufficiently addressed all the concerns raised in the initial review. I commend their efforts in revising the manuscript, and I have no further concerns regarding the study.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors describe a filtering strategy for identifying pathogenic variants from long-read sequencing data, significantly reducing the number of structural and copy number variants. After testing it on a cohort with known variants, the authors demonstrate that the applied strategy can diagnose patients with negative NGS analysis. As was the case years ago with second-generation sequencing methods, the clinical application of long-read sequencing is held back by the huge amount of data to manage and variant calls from which to select the one or few disease-causative variants. Therefore, effective filtering strategies are needed to definitively translate technologies with enormous diagnostic potential into clinical practice.

The manuscript is very well written, and the work is certainly of interest, but it does not describe innovative applications or strategies compared to those already published in the literature with respect to long-read sequencing (e.g. studies listed in the paper by Mastroianni et al., 2023). The main novelty concerns the identification of a methylation signature specific for SMA and useful for its diagnosis. Perhaps this is the finding to be highlighted most emphatically although it should be tested on a larger number of patients with molecular diagnosis. Following are some doubts that require clarification and some suggestions:

- Within the optimization cohort, the applied filtering strategy reduced the number of CNVs and SVs by 98 and 99.9% respectively. What about short variants (SNVs and INDELs)? Related data are reported in the supplementary table 2 but they deserve a comment within the manuscript.

We thank the reviewer for their feedback. We have now performed additional analysis on SNV/INDELs detected from long read sequencing. We analyzed noncoding short variants on all positive and negative cases, while for coding SNVs we focused on those identified by LRS but missed by exome sequencing in the negative or inconclusive cases. The details of the method and results are mentioned in the manuscript (page 3, lines 122-144; page 7, lines 285-292), Supplementary Table 3, Supplementary Fig. 3, and Extended Methods (page 3, lines 88-117).

- Why is the methylation within the 15q11.2 region reduced in the OXN-068 sample which is not positive for Angelman syndrome, as confirmed by MS-MLPA? Was this simply a false positive (although with a not completely altered profile)?

We thank the reviewer for the question. Although OXN-068, exhibited reduced methylation at the 15q11.2 loci, it had significantly higher methylation levels compared to the Angelman positive sample (Supplementary Fig. 1e). MS-MLPA testing confirmed that OXN-068 is negative for Angelman Syndrome. Interestingly, OXN-68 is SMA positive; a recent study by Heimdörfer et al (Am J Hum Genet, 2024) has linked SMN1 to the Padre-Willi/Angelman locus, which we speculate can explain the altered methylation profile. However, we did not see the similar methylation profile in other SMA patients. Although an interesting case in itself, understanding the mechanism, functional and phenotype impact of the altered methylation in this region is out of scope for this study. We have now added this analysis as part of the Supplementary Fig. 1e-f and have explained in the legend for this figure.

- The OXN-020 sample of the optimization cohort has a splicing variant in the *HBB* gene. Supplementary Table 2 shows that the filtering strategy for splicing variants was not applied in the optimization cohort. Why? How was the variant identified from the LRS data? Moreover, why not analyze the exonic SNVs as well? LRS could identify variants missed by exome analysis because located in GC-rich regions or characterized by highly homologous sequences.

We thank the reviewer for the feedback. The *HBB* splice variant for OXN-020 was indeed identified through the filtering strategy for splice variants. We have now added the detailed information regarding splicing in the Supplementary Table 3 and the filtering strategy in Supplementary Fig 3a (right panel) for the positive cohort as well.

As suggested by the reviewer, we have performed new analysis focusing on exonic variants identified by LRS but missed by whole exome sequencing in the negative cohort. The results did not lead to any positive diagnosis in these patients (although ruled out a splicing variant). These results and methods are reported in manuscript (page 3, lines 122-144; page 7, lines 285-292), Supplementary Table 3, Supplementary Fig. 3b-c, and Extended Methods (page 3, lines 102-117).

- The two deletions prioritized from OXN-033 LRS data and filtered out after manual inspection were false positive calls not confirmed by CMA or they were considered non-clinically relevant? I think it is important that the authors also report for newly diagnosed patients the variants prioritized by the filtering strategy and then discarded, explaining in detail why.

We thank the reviewer for raising this point, indeed only one was classified as pathogenic and clinically relevant. Using the new larger dataset, only one other deletion was filtered out. This heterozygous deletion was real, as confirmed by CMA, but it was inherited from a healthy parent and affects one gene known to cause an autosomal recessive immunodeficiency disorder, so it was not considered as contributory to this patient's disease. We have now discussed this in detail in the manuscript (page 4, lines 155-159).

- For the OXN-048 sample, the size of the del16q23.3 should be reported.

Thank you. We have now reported the size of the deletion (page 4, line 162).

- The 3.6 kb *MPLKIP* deletion of patient OXN-027 should be conventionally considered among CNVs. Was it called by Sniffles and not by QDNAseq?

The deletion in *MPLKIP* was detected from SV calling methods Sniffles and CuteSV and not from CNV calling method QDNAseq. We have mentioned this now in the manuscript (page 4 line, 169).

- The authors suggest that *MPLKIP* overexpression could underlie the patient's phenotype but reported causative *MPLKIP* variants cause protein deficiency. How could the overexpression be responsible for the disease?

Thank you for the question. Indeed, biallelic loss of function of the *MPLKIP* gene has been associated with learning disabilities and brittle hair, while overexpression of the protein has not been linked to disease. However, given the phenotypic overlap between *MPLKIP*-related

disease and our patient's phenotype, and that MPLKIP protein localizes to the centrosome and plays roles in regulation of mitosis and maintenance of cell cycle integrity, we speculate that its optimal function might be dependent on its expression levels, and that overexpression might equally interrupt cell cycle maintenance leading to disease. However, as we now clarify in the manuscript (page 4, lines 175-181), functional evidence is needed to confirm this possibility.

- The authors state that the application of LRS led to the diagnosis of patients who had inconclusive testing with traditional methods. Generally, CMA is performed as a first-tier analysis on NDD patients, and in this cohort, if performed before LRS, it would likely have identified most of the CNVs detected by LRS. Why CMA was not performed on these patients, particularly those with NDD? The authors should discuss this point.

That is a great comment. Actually, a consensus statement in 2019 recommended that whole exome sequencing be used as the first-tier testing for patients with neurodevelopmental disorders (NDD), given the superior diagnostic yield of exome relative to CMA (PMID: 31182824). Since then, we have worked with our ordering providers to change their workflow where exome sequencing is first ordered for NDD patients. Of course, not all patients will undergo CMA (or any additional testing) after a negative exome for several reasons (insurance reimbursement, follow up issues, etc.). Nevertheless, as we clarify in manuscript (page 3, lines 116-118), 41% of patients in our cohort had multiple genetic testing, 86% of whom had CMA. Our hope is that LRS might become a single unified assay for the detection of SNVs, INDELs, CNVs, SVs and methylation profiles in routine clinical genetic testing.

- The choice to consider only variants involving the coding region of disease-causing genes probably reduced sensitivity, especially with regard to SVs that can be causative even if the breakpoints fall in noncoding regions, through various mechanisms (e.g., position effect, alteration of TADs...). The authors should discuss this point

We thank the reviewer for the feedback. We agree with the reviewer that, by restricting ourselves to disease-causing genes, we are probably reducing the sensitivity. However, the purpose of this study is to understand the diagnostic utility of the method in a clinical setting where variants can be interpreted using existing guidelines such as those by ACMG or ClinGen. In this context, variants in noncoding regions, with possible pathogenic effects as correctly pointed out by the reviewer, will be missed as they might need additional analyses to characterize their functional effects. As requested by reviewer, we have now discussed this point in the manuscript (page 5, line 209-215).

- In some parts of the text ddPCR is described as double digital PCR, in others as digital droplet PCR. The correct name is droplet digital PCR.

All the text referring to droplet digital PCR has been appropriately changed in the manuscript and Extended methods.

Reviewer #2 (Remarks to the Author):

The authors present a novel genetic diagnostic framework to address the limitations of existing long-read sequencing methods. While its clinical utility is promising, there are significant issues regarding its scientific validity.

Major Concerns

1. The authors have included 10 positive control cases and 39 undiagnosed genetic cases, but the sample size is too small. Moreover, since the majority of subjects are of Arab ethnicity, there are substantial concerns about the reproducibility and generalizability of the data.

Thank you for this comment. We have made significant changes to address this concern:

- 1- We have now increased our sequenced sample size to a total of 70 samples, with 51 undiagnosed cases and 17 diagnosed cases representing the “Negative” and “Positive” cohorts, respectively, as well as 2 healthy SMA carriers. All positive cases were correctly identified.**
- 2- In addition, we have now evaluated the methylation profile matching module (Epimarker) against the epi-signatures from 34 Mendelian neurodevelopment disorders (MNDD) in an independent long read sequencing dataset of 57 samples and observed 100% sensitivity in correctly classifying the disease samples (N=17). None of the 40 controls from this dataset were incorrectly classified as a MNDD demonstrating 100% specificity (Supplementary Table S1, extended methods and manuscript).**
- 3- For spinal muscular atrophy analysis, we increased our sample size to a total of 17 samples with 6 positive patients, 2 carriers, and 9 non-carriers (negative) individuals, and we performed further analysis, including droplet digital PCR and LRS read deconvolution, to show reproducible methylation tagging and to confirm bi-allelic *SMN1* loss in all patients.**
- 4- Using our filtering strategy on the two cohorts positive (N=17) and negative (N=51) cohorts, we observed similar percent of reduction for CNV, SNV and SV (Figure 1b, 2a, Supplementary Fig S1c, S3a, and Supplementary Table S3). We hope these new changes and results support the reproducibility of the study.**

We agree with the reviewer that majority of human subjects in our study are of Arab ethnicity, which is consistent with the patient population we serve in the United Arab Emirates. Given the underrepresentation of Arabs and the Middle East in global genomic studies and databases, and the nature of our study, aiming to assess the utility of LRS and additional variation (CNVs, SVs, methylation) in clinical settings which can be population-independent, we hope that the reviewer agrees that this should not affect the generalizability of the study. Instead, we hope studies from underrepresented populations, like ours, will eventually lead to equitable access to benefits in genomic advances (which is arguably skewed towards individuals of European Ancestry).

2. The authors claim that the "funnel down" filtering method efficiently reduces the number of candidate variants. However, the rationale behind the filtering criteria is unclear, raising doubts about the validity of this approach. It is essential to assess how many potentially pathogenic variants were excluded at each filtering stage and to develop an appropriate filtering strategy.

We thank the reviewer for this comment. We have now addressed the rationale in the manuscript (page 2, lines 65-75) and explained the criteria in Extended Methods (pages 1 and 2) in more detail. Number of variants selected for further investigation at each step is mentioned in Fig. 1b, Fig. 2a, supplementary Fig. 1c, and supplementary Fig. 3a.

Minor Concerns

1. Page 7, Line 116: "SLC38A3" should be "SLC38A8".

Thanks. This has been corrected

2. Page 9, Line 191: "parametes" should be "parameters".

Thank you. This is corrected.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.