Long-Read Genome Sequencing Resolves Genetic Cases Missed by Short-Read Sequencing

Greta Pitsava[1], Krista Bluske[2], Ivan De Dios[1], Rebekah Barrick[1], Sami Belhadj[2], Erica Smith[2], Miguel Almalvez[1], Arthur Ko[3], Lightning Auriga[1], UCI-GREGoR Consortium, Rachid Karam [2], Changrui Xiao[4], Emmanuèle C. Délot[1], Seth Berger[2], Eric Vilain[1]

1 Institute for Clinical and Translational Science, University of California, Irvine, CA, USA

2 Ambry Genetics, Aliso Viejo, CA, USA

3 Center for Genetic Medicine Research, Children's National Research Institute, Washington, DC, USA

4 Department of Neurology, University of California, Irvine, CA, USA

Background

Long-read genome sequencing (LR-GS) has the potential to increase the diagnostic yield of genetic testing by overcoming limitations of short-read sequencing such as the detection of variants in highly homologous regions and the detection of complex structural variants. However, the diagnostic yield of LR-GS is not well-characterized.

Objective

To evaluate the diagnostic utility of LR-GS in individuals suspected to have a genetic disorder but with no molecular diagnosis despite extensive prior workup including chromosomal microarray testing, exome sequencing, and short-read genome sequencing (SR-GS).

Methods

We performed PacBio HiFi LR-GS in 73 families: 55 trios, 10 parent-proband duos and 8 proband-only cases. Variant interpretation was carried out using the Geneyx platform combined with manual curation and filtering.

Results

We identified a diagnosis in 3 out of 73 cases, representing an incremental diagnostic yield of 4.1% (95% CI: 1.4%-11.5%) after negative SR-GS. In one case, the causative variant was a 12 kb deletion including exon 7 in *PEPD*. Both breakpoints demonstrated sequence identity, explaining why short-read copy number variant callers were unable to confidently identify the deletion. In the second case, a nonsense variant in *ACAN* was previously missed because it fell within a 1-kb “dead zone” in exon 12 where SR-GS reads align with zero mapping quality because of identical matches elsewhere in the genome. The third case was missed in the prior variant analysis pipeline because it was deep intronic. Validation via RNA-seq demonstrated the abnormal inclusion of a pseudoexon in the mature transcript.

We also identified a candidate variant in 14 additional cases. 3 of these cases (4.1%) are deemed by the study team as "probably solved” based on compelling molecular characteristics, phenotype, and mode of inheritance match, though additional confirmatory studies are needed. All 3 of these could have been detected with a reanalysis of SR-GS. Of the 11 remaining cases, two had variants undetectable with SR-GS. The first was an inversion in *trans* with a previously identified missense variant. The second case was a compound heterozygote with two missense variants, neither of which were called by SR-GS. The remaining candidate variants could all have been detected via reanalysis of SR-GS but were filtered out by the automated pipeline.

Conclusion

When focusing on variant detection, LR-GS was able to diagnose clinically important variants not detected in SR-GS through overcoming issues related to homology and detection of structural variation.