Title: Leveraging long-read sequencing to identify *de novo* variants from parent-proband duos

Authors: L. Boukas, E. C. Délot, G. Pitsava, C. Lambert, C. Fanslow, P. Baybayan, S. Belhadj, B. Losic, J. Harting, K. Bluske, J. LoTempio, H. B. Al-Kouatly, R. Karam, W. J. Rowell, C. Xiao, E. Vilain, & S. I. Berger.

Abstract: *De novo* status can often upgrade a variant from VUS to pathogenic. However, asserting *de novo* status currently requires sequencing of the proband and both biological parents. As a result, millions of single-parent families are less likely to receive a genetic diagnosis and benefit from precision treatment and management options.

To address this limitation, we developed *duoNovo*, which identifies *de novo* variants from parent-proband duos, i.e. using only one biological parent. *duoNovo* leverages long-read sequencing followed by haplotype reconstruction and detection of identical-by-descent (IBD) haplotype blocks. *duoNovo* can determine whether a candidate variant of interest is *de novo* without knowledge of the missing parent's genotype, by testing whether the variant is present on an IBD haplotype shared by the proband and the sequenced parent.

To evaluate *duoNovo*, we sequenced 40 trios with the PacBio HiFi technology. To the best of our knowledge, this is one of the largest trio cohorts sequenced with long-read sequencing to date. We applied *duoNovo* to each of the 80 duos constructed by masking one parent, classifying over 20 million single-nucleotide variants/indels. We assessed *duoNovo*'s performance against ground truth classifications obtained from the full trios (which included over 1900 *de novo* variants), demonstrating very high precision (average 95%) and low error rate (average 1.15e-6). Notably, *duoNovo* had perfect accuracy (100% positive predictive value) when tested on variants absent from gnomAD, which are enriched for rare pathogenic alleles and are thus the most likely to be clinically relevant.

A practical issue is that most single-parent families are single-mother families, while most *de novo* variants are transmitted from the paternal germline. To address this, we show that unaffected siblings can serve as surrogates for the missing father (since they share 50% of their genome with the father), allowing *duoNovo* to detect a significant fraction of paternally transmitted *de novo* variants.

In summary, *duoNovo* has the potential to transform the diagnostic yield of single-parent genetic testing, and represents an example where long-read sequencing provides clear benefit over short-read sequencing even for single nucleotide variants. We have implemented *duoNovo* as an R package, aiming to facilitate its use and adoption by the community.