Identification of *de novo* variants from parent-proband duos via long-read sequencing

Leandros Boukas[1, 2], Emmanuèle C. Délot[2, 3], Georgia Pitsava[2, 3], Christine Lambert[4], Cairbre Fanslow[4], Primo Baybayan[4], Sami Belhadj[5], Bojan Losic[5], John Harting[5], Krista Bluske[5], Jonathan LoTempio[2, 3], Huda Al-Kouatly[6], Rachid Karam[5], William Rowell[4], Changrui Xiao[2, 3], Eric Vilain[2, 3], and Seth I. Berger[1, 2, 7*]

[1]Children's National Hospital, Washington, DC

[2]UCI – GREGoR

[3]University of California, Irvine, CA

[4]PacBio, Menlo Park, CA

[5]Ambry Genetics, Aliso Viejo, CA

[6]Thomas Jefferson University, Philadelphia, PA

[7]The George Washington University, Washington, DC

[*]Correspondence to sberger@childrensnational.org (SIB)

Abstract

While *de novo* variants cause many Mendelian disorders, their detection currently requires sequencing of the proband and both biological parents. This is not feasible when only one parent is available, a limitation for millions of families. We developed *duoNovo*, which identifies *de novo* variants from single parent-proband duos using long-read sequencing followed by haplotype reconstruction and detection of identical-by-descent haplotype blocks. We sequenced 40 trios and applied *duoNovo* to each of the 80 duos constructed by masking one parent, classifying over 20 million variants. We evaluated *duoNovo*'s performance against classifications obtained using the full trios (which included over 1,900 *de novo* variants), demonstrating very high precision and low error rate, and perfect accuracy among variants absent from gnomAD. *duoNovo* is freely

available to the community as an R package, and may represent an example where long-read sequencing provides clear diagnostic benefit over short-read sequencing.

Introduction

As genetic testing becomes standard of care for numerous clinical indications, it is crucial to consider social determinants that can lead to inequities in result interpretation and diagnostic outcomes. Diversity in family structure creates common scenarios where variant interpretation is hindered by the unavailability of samples from both biological parents. This is particularly relevant for *de novo* variants, which are present in affected probands but not (constitutively) present in the parents. *De novo* variants are recognized as underlying a substantial fraction of Mendelian disorders (Brunet et al., 2021; Ligt et al., 2012; Posey et al., 2017; Vissers et al., 2010; Wright et al., 2023; Yang et al., 2014). These disorders are typically highly penetrant, and as a result, the American College of Medical Genetics (ACMG) guidelines use a variant's *de novo* status as a strong criterion for pathogenicity (Richards et al., 2015).

Currently, *de novo* status can only be determined through sequencing of the proband and both biological parents. However, various factors can prevent both parents from participating in genetic testing. US Census statistics indicate that there are approximately 10 million single-parent families in the United States alone (predominantly single-mother families). Family structures are further complicated by factors such as misattributed parentage, children conceived through sperm or egg donation, or parental death. Consequently, probands in these families face a higher risk of non-diagnostic results (Rehm et al., 2023; Richards et al., 2015; Wojcik et al., 2024). Given that ACMG guidelines advise against using VUS in treatment plan determinations, this can lead to missed opportunities to benefit from precision treatment and management options. Importantly, the magnitude of the reduction in diagnostic yield due to the unavailability of parental samples correlates with ancestry, indicating the potential for further exacerbation of existing health disparities (Wright et al., 2023).

Long-read sequencing (LRS) is emerging in the clinical setting, although higher cost combined

with uncertainty regarding diagnostic benefit over short-read sequencing (SRS) currently limit its widespread adoption (Wojcik et al., 2023). To characterize the potential diagnostic benefit, initial investigations have largely focused on the ability of LRS to detect large/complex structural variants and other variants missed by SRS (Cohen et al., 2022; Mastrorosa et al., 2023; Merker et al., 2018; Miller et al., 2021; Mizuguchi et al., 2021; Negi et al., 2025; Xie et al., 2020); an additional studied benefit is the concurrent detection of epigenetic variation (Vollger et al., 2025). However, recent improvements in LRS technologies, such as the development of PacBio HiFi LRS, now allow one to obtain reads with very high basecall accuracy (Gustafson et al., 2024; Hiatt et al., 2024; Kolmogorov et al., 2023; Kucuk et al., 2023; Mastrorosa et al., 2023; Wenger et al., 2019), enabling the identification of small variants including single-nucleotide variants. One unique advantage of LRS is that it couples the identification of such variants with their reliable, read-backed phasing (Holt et al., 2024; Martin et al., 2016). This is especially important for potentially disease-causing variants evaluated for *de novo* status in the clinical diagnostic setting, which are very rare or even unique to an affected individual and thus not amenable to phasing methods that rely on information about haplotype frequencies in the population (Browning and Browning, 2011; Guo et al., 2024).

Here, we leverage these features of LRS to develop a method - which we call *duoNovo* - to identify *de novo* variants from duos (proband and one biological parent), with the potential to open more diagnostic opportunities for these families.

Results

duoNovo: overview of approach

duoNovo uses haplotype blocks consisting of phased single nucleotide variants across hundreds of kilobases in order to determine whether a candidate variant arose on the haplotype inherited from the sequenced parent – and is thus a *de novo* variant – or not (Methods). It achieves this by evaluating the sequence similarity between pairs of haplotype blocks. Each pair consists of a proband haplotype block and a parental haplotype block. If the proband haplotype block containing the variant being tested for *de novo* status exhibits perfect similarity (excluding the candidate

variant) with one of the two parental haplotype blocks, then these haplotype blocks are inferred to be identical-by-descent, which in turn implies that the candidate variant is *de novo* (Figure 1A; Methods). If, on the other hand, it is the other proband haplotype (the one not containing the candidate variant) that exhibits perfect similarity with one of the two parental haplotype blocks, then *duoNovo* infers that the haplotype containing the candidate variant was inherited from the non-sequenced biological parent (Figure 1A). While in this latter case the candidate variant may still be *de novo*, this is not possible to determine from the available duo.

duoNovo has high positive and negative predictive value

To assess the performance of *duoNovo*, we sequenced 40 trios using PacBio HiFi LRS to an average depth of $\sim$ 33-fold across all individuals, obtaining approximately 17.7 kb-long reads on average (Methods).

We started by applying *duoNovo* to each of the 40 father-proband duos, treating the mother as the non-sequenced parent (Figure 1B). We identified an average of 1,455,871 candidate variants across these duos (heterozygous phased variant call in the proband; homozygous reference call in the father). We subjected each candidate variant to strict quality control (QC; including sequencing depth, genotype quality, and regional quality) and tested each variant that passed our QC filters for *de novo* status (Methods). Across all 40 duos, an average of 24 variants (range 14-47) were classified as *de novo* (Figure 2A). Reassuringly, the number of *de novo* classifications positively correlates with father's age (Supplemental Figure S1; Poisson regression $p = 7.04 \cdot 10^{-9}$). To evaluate the accuracy of the classifications, we examined the maternal sequence. For each variant classified as *de novo*, we determined that the classification is correct if the variant is absent in the mother (Figure 1B). We note that this a conservative assessment, as a variant can still be *de novo* even if it is present in the mother (for example, at recurrently mutated sites). We found that the average positive predictive value of *duoNovo* across duos is 98% (Figure 2A; range 81 - 100%). While the negative predictive value is expected to be high because *de novo* variants are extremely rare, we found that *duoNovo* always outperforms the naive baseline (calling every variant as non-*de novo*) (Figure 2B).

We next turned our attention to the mother-proband duos, this time treating the fathers as the non-sequenced parents. As expected given the elevated contribution of the paternal germline to the pool of *de novo* variants (Jónsson et al., 2017; Kong et al., 2012; Sasani et al., 2019), *duoNovo* always detected more *de novo* variants in father-proband duos compared to mother-proband duos (with the median ratio of paternally to maternally derived *de novo* classified variants being 4.2; Supplemental Figure S2). This can explain why, while the positive predictive value in mother-proband duos remains high, it is somewhat lower compared to father-proband duos (Figure 2C; average 92%; range 60-100%). As in father-proband duos, *duoNovo* always has higher negative predictive value compared to the naive baseline (Figure 2D).

In both father-proband and mother-proband duos, we observed no difference in the positive predictive value when the proband was of European vs non-European ancestry (Supplemental Figure S3; Methods)

duoNovo has perfect accuracy when applied to candidate variants absent from gnomAD

In the previous section, we quantified the positive predictive value of *duoNovo* among all variants classified as *de novo*. In a realistic diagnostic setting, the potential pathogenicity of candidate variants evaluated for *de novo* status is supported by additional lines of evidence, an important one being their absence (or very low frequency) in control population databases. To evaluate the performance of *duoNovo* in such a setting, we focused on candidate *de novo* variants that are absent from gnomAD (Karczewski et al., 2020), and therefore have a higher prior probability of being *de novo* and pathogenic. *duoNovo* achieved perfect accuracy. Among 629 variants classified as *de novo* from father-proband duos and 142 variants classified as *de novo* from mother-proband duos, the positive predictive value is 100% in all but one of father-proband duos, and all mother-proband duos (Figure 2E).

duoNovo very rarely classifies reference alleles as *de novo*

Our results so far focus on candidate alternative alleles (1|0 or 0|1 proband genotype; 0/0 parental genotype). As an additional check of *duoNovo*'s performance, we compared the percentage of

candidate alternative alleles classified as *de novo* to the corresponding percentage for candidate reference alleles (1|0 or 0|1 proband genotype; 1/1 parental genotype). Collectively, alternative candidate alleles were $\sim$ 97 times more likely to be classified as *de novo* compared to reference candidate alleles in father-proband duos, and $\sim$ 11 times in mother-proband duos. In 36 out of the 40 father-proband duos and 30 out of the 40 mother-proband duos, no reference alleles were classified as *de novo* (Supplemental Figure S5). This further supports the accuracy of our *de novo* classifications, since the reference allele is common in the population and thus candidate reference alleles have a very low prior probability of being *de novo*.

duoNovo has very low false positive rate

To directly quantify the genome-scale false positive rate of *duoNovo*, we subsequently restricted our attention to candidate variants which we inferred – based on the full trio information – to have been transmitted from the non-sequenced parent (reliable heterozygous phased variant call in the proband and reliable heterozygous or homozygous variant call in the non-sequenced parent; reliable homozygous reference call in the sequenced parent). We examined the fraction of these variants that are classified as *de novo* (false positive rate; equivalent to the specificity), and found that *duoNovo* has an average false positive rate across father-proband and mother-proband duos equal to 1.2e-6, and 1.1e-6, respectively (Figure 3A, B). The majority (close to 60%) of these variants were correctly classified as present on the haplotype inherited from the non-sequenced parent, with slightly more than 40% labeled as uncertain (Figure 3A). In 23 of the 40 father-proband duos, and 24 of the 40 mother-proband duos, there were zero false positives (Figure 3B); the maximum false positive rate across all duos was 6.4e-6. As noted in the previous section, this is a conservative assessment since some of these variants may in fact be *de novo* even though they are present in the non-sequenced parent.

As an orthogonal assessment of the false positive rate, we applied *duoNovo* to swapped duos consisting of the father and the mother from each trio (that is, the proband was swapped with the parent previously treated as the non-sequenced one). From each swapped duo, we calculated the percentage of candidate variants that were classified as "*de novo*", or as "present on the non-

sequenced parent haplotype". Either classification requires the detection of a haplotype block shared identical-by-descent between the two members of the duo. We would thus expect the percentage of classified variants from these swapped duos to be very low, since fathers and mothers share a much smaller percentage of their genome identical-by-descent compared to father-proband or mother-proband pairs. Indeed, we found that the median percentage across swapped duos is 0.03%, compared to 65% across true parent-proband duos (Supplemental Figure S6). One swapped duo stands out as an outlier, with 10.6% of variants receiving a classification (Supplemental Figure S6); this is substantially higher than all the rest, though still well below the percentage from true parent-proband duos (Supplemental Figure S6). Upon closer examination, we found that the father and mother from that swapped duo had the highest relatedness score among all father/mother pairs (Methods; Supplemental Figure S6), which could explain why *duoNovo* detected a larger fraction of identical-by-descent haplotype blocks.

duoNovo's sensitivity reflects known parent-of-origin effect

Finally, we sought to characterize the sensitivity of *duoNovo*, based on a set of *de novo* variants detected using the entire trio (hereafter referred to as trio-*de novo* variants; Methods), which we treated as the ground truth. Consistent with the aforementioned fact that most *de novo* variants are contributed by the fathers, we found that *duoNovo* collectively classified $\sim$ 54% of trio-*de novo* variants as *de novo* from the father-proband duos, and only $\sim$ 13% as *de novo* from the mother-proband duos (Figure 3C). Conversely, approximately 13% were classified as present on the maternally inherited haplotype from the father-proband duos, whereas about 47% were classified as present on the paternally inherited haplotype from the mother-proband duos (Figure 3C). 33% and 40% of variants were labeled as uncertain from the father-proband and mother-proband duos, respectively. Looking at the total fraction of trio-*de novo* variants that are classified as *de novo* (either from the father-proband or from the mother-proband duo; sensitivity), we found that the average sensitivity across all trios is 57% (Figure 3D). The remaining were all variants that were either labeled as uncertain or did not pass our QC filters; none of the trio-*de novo* variants were misclassified as non-*de novo* (Figure 3D).

Variants classified as *de novo* fall into expected mutation subtypes

In agreement with previously reported mutational patterns among *de novo* variants (Kessler et al., 2020; Shojaeisaadi et al., 2024), we found that the single nucleotide variants *duoNovo* classified as *de novo* are mostly C>T and T>C substitutions (Supplemental Figure S4), with the C>T substitutions occurring both within and outside the CpG context. There was no statistically significant difference in the proportion of mutation types among the *de novo* classifications from father-proband duos (915 single nucleotide variants), and those from mother-proband duos (236 single nucleotide variants) (Supplemental Figure S4; chi-squared $p = 0.4$). We also found no statistically significant difference in the proportion of indels classified as *de novo* from father-proband versus mother-proband duos (Fisher's exact test, $p = 0.39$).

Discussion

We have developed and evaluated a method that enables the identification of *de novo* variants from parent-proband duos. Our approach is simple, and leverages the unique ability of LRS to both accurately detect and phase variants across large genomic segments. While it is established that LRS can detect variants that are undetectable with SRS, *duoNovo* illustrates that LRS can also enable the interpretation of variants that can be detected with SRS, but are hard-to-interpret.

Conceptually, *duoNovo* is based on the recognition that phasing can provide information about haplotype transmission, and thus reveal variant inheritance, without requiring access to both biological parent genotypes (see also Steyaert et al. (2024)). As such, it has the potential to improve variant interpretation in cases where sequencing both parents is not feasible, addressing a critical unmet need in clinical genetic testing. An inherent limitation of our approach is that it can only definitively classify *de novo* variants derived from the germline of the sequenced available parent. Given that the paternal germline contributes approximately 4 times as many *de novo* single nucleotide variants than the maternal germline (Jónsson et al., 2017; Kong et al., 2012; Sasani et al., 2019), this will limit the yield in assigning *de novo* status to variants identified clinically. However, regardless of whether it is applied to a father-proband or mother-proband duo, *duoNovo* can aid

in the classification of clinically identified variants whose interpretation was otherwise limited by the unavailability of biologic family members.

While we have focused on single nucleotide variants and indels here, we anticipate that *duoNovo* is going to be similarly useful for additional variant types which LRS is capable of reliably detecting, such as structural variants and short tandem repeats (Beyter et al., 2021; Chaisson et al., 2015; Figueroa et al., 2024; Logsdon et al., 2020; Sedlazeck et al., 2018; Smolka et al., 2024). Finally, although we used PacBio HiFi LRS in this study, we note that *duoNovo* is generally applicable to any sequencing platform that generates accurate variant calls and enables read-backed haplotype reconstruction.

In summary, *duoNovo* is the first systematic method that can identify *de novo* variants from duos with high accuracy at the genome scale. It has the potential to increase the diagnostic yield of genetic testing for millions of families with diverse family structures, and may represent an example where long-read sequencing provides a clear benefit compared to short-read sequencing in the clinical diagnostic setting. We anticipate that its implementation and free availability as an R package will facilitate its use and adoption by the community.

Methods

Participants

We sequenced 122 individuals, from 38 trios and two quad families. Each of the two quad families was split into two trios, thus yielding a total of 42 trios. Two of these trios were then excluded from the analysis because more than 75% of candidate variants detected from the mother-proband duo failed QC due to low sequencing depth. Thus, all analyses were conducted with the remaining 40 trios. All individuals consented to provide samples for genetic testing to the Pediatrics Mendelian Genomics Research Center (part of the GREGoR consortium). The study was approved by Children's National Hospital IRB (IRB #PRO00015852). Age (at childbirth) and genetic ancestry of the parents are provided in Supplemental Table 1.

Long-read sequencing

Samples were prepared for LRS following the standard operating procedure (SOP) available at PACB.com. ("Preparing whole genome and metagenome libraries using SMRTbell prep kit 3.0"). Sequencing was performed on the PacBio Revio system with the Revio polymerase Kit, following the Revio SMRT Link setup. Genomic DNA quality and concentration were assessed using the FEMTO Pulse (Agilent Technologies) and Qubit dsDNA HS reagents Assay kit (Thermo Fisher Scientific). SMRTbell libraries were constructed on the Hamilton Microlab Star and VENUS 5 system following the SOP available on PACB.com under "Preparing whole genome and metagenome libraries using SMRTbell prep kit 3.0". For library characterization, post size-selected SMRTbell Library samples were quantified using the Qubit DNA HS assay and DNA size was estimated using the FEMTO Pulse. The libraries were loaded at an On-Plate Concentration of 90 pM.

Raw sequencing reads were subsequently processed as follows. Circular consensus sequences were generated from subreads using CCS (version 7.0.0) with default settings, yielding demultiplexed HiFi reads. These reads were then aligned to the hg38 reference genome (https://github.com/PacificBiosciences/reference_genomes/blob/main/reference_genomes/human_GRCh38_no_alt_analysis_set/README.md) using PBMM2 (version 1.10.0), with default parameters.

Variant calling

Variant calling following LRS was performed using DeepVariant (version 1.5.0; (Poplin et al., 2018; Yun et al., 2021)). For each duo as well as for each trio, gvcfs were first generated and joint variant calls were subsequently produced for each duo and trio separately with GLnexus (version 1.4.1) with the preset "DeepVariant_unfiltered". All our analyses and numbers reported are based on candidate *de novo* variants that were called both from a duo and its corresponding trio and thus were candidates we could evaluate.

For the evaluation of *duoNovo*'s performance, genotype calls from the non-sequenced parent at positions with PHRED genotype quality (GQ) at least 40 were used. Specifically, *de novo* classifications were determined to be correct if the non-sequenced parent did not have the candidate

variant allele at that position (that is, if the non-sequenced parent was either homozygous for a non-variant allele or heterozygous with two different non-variant alleles). Similarly, non-*de novo* classifications were determined to be correct if the non-sequenced parent had the candidate variant allele at that position.

To obtain *de novo* variants from the entire trio, we first identified positions with a heterozygous variant in the proband and a homozygous reference call in both parents. Out of these positions, we then retained those with a minimum GQ of 40 for each sample, and a minimum depth of 20 reads for each sample. Multi-allelic variant sites were excluded. From the resulting set of *de novo* variants, we excluded variants clustered within the same phasing haplotype block, reasoning that these are likely to represent artifacts (see also below).

Phasing

Following LRS, phasing was conducted using HiPhase (version 1.4.0) (Holt et al., 2024) with default parameters on the jointly called vcf from each duo. HiPhase assigned each phased variant to a phasing set. Variants in the same phasing set had the same phase relative to one another, enabling the resolution of those variants into haplotypes.

Estimation of haplotype similarity and classification of candidate *de novo* variants with *duoNovo*

Following phasing, we imported the vcf files containing the phased variant calls into R with the VariantAnnotation package (Obenchain et al., 2014). We then only retained variant positions where sequencing depth was at least 20. With respect to GQ, we initially retained all positions with GQ at least 30. After obtaining classifications based on parent-proband sequence similarities computed using these positions (see next paragraph for details), we only retained classified variants with GQ at least 40 to ensure our results are not affected by genotyping errors. As stated in the Results, we defined candidate *de novo* variants as variants heterozygous in the proband and absent in the parent.

We then developed *duoNovo* to obtain classifications for these candidate variants. *duoNovo* first defines haplotype blocks as genomic regions within which: a) proband variants are assigned to the same phasing set; and b) parent variants are assigned to the same phasing set. *duoNovo* then evaluates candidate variants, separately in each of the two phasing orientations. It starts by discarding candidate variants that are either in haplotype blocks smaller than 10 kb, or at the boundaries of the remaining blocks (2 kb from the start/end coordinates). These boundary variants are discarded because we found that when they are classified as *de novo*, they tend to be false positives, especially in mother-proband duos (see below). Subsequently, *duoNovo* sequentially examines each of the proband haplotype blocks containing candidate *de novo* variants. Each of these haplotype blocks is compared to each of the two corresponding parental haplotype blocks using the Hamming distance. In addition, *duoNovo* performs the same comparisons for the other proband haplotype (not containing the candidate variant). In all cases, the Hamming distance between a pair of haplotype blocks is calculated after representing each haplotype block as a binary string (with 0 for the reference allele and 1 for the variant allele) and excluding candidate variants in the same phasing orientation as the candidate variant being evaluated. After performing these comparisons, *duoNovo* produces classifications based on the following criteria:

(i) If the proband haplotype block containing the candidate variant has a Hamming distance of 0 with only one of the two parental haplotype blocks, while the other proband haplotype block has a Hamming distance greater than 40 with both parental haplotype blocks, the candidate variant is classified as *de novo* (see below for an examination of the impact of the specific Hamming distance threshold).

(ii) If the proband haplotype block containing the candidate variant has a Hamming distance greater than 40 with both parental haplotype blocks, while the other proband haplotype block has a Hamming distance of 0 with only one of the two parental haplotype blocks, the candidate variant is classified as inherited from the other parent.

(iii) If none of the two above conditions were satisfied (e.g. due to no proband-parent haplotype pair having a Hamming distance of 0), the candidate variant does not receive a classification

and is labeled as uncertain.

To minimize false positive classifications due to genotyping errors, from the resulting classifications we excluded variants falling within regions stratified as problematic by the Genome-in-a-Bottle consortium (Dwarshuis et al., 2024). Furthermore, as mentioned earlier, we labeled candidate variants with GQ less than 40 in either proband or parent as uncertain.

Finally, if two or more of the variants classified as *de novo* are present within the same haplotype block, *duoNovo* discards all *de novo* classifications in that block. While these could represent true *de novo* events, we have observed that they tend to be false positives (see section "Examining sources of false positive classifications").

Interpretation of Hamming distance equal to 0 between a proband-parent haplotype block pair

Intuitively, a Hamming distance equal to 0 between a proband-parent haplotype block pair indicates that this haplotype is shared identical-by-descent. This intuition can be made more precise when we consider the genotypes of the different positions that determine the Hamming distance.

1. **Positions where the proband is heterozygous and the parent is homozygous.** These positions provide direct information about haplotype transmission, because only one of the two proband haplotypes could have been transmitted from the sequenced parent. We note that, without read-backed phasing, candidate variants would always be classified as inherited from the non-sequenced parent, since they are heterozygous in the proband and absent in the sequenced parent. Hence, a candidate variant can only be classified as *de novo* using read-backed phasing. In turn, this implies that in order for the classification to be accurate, it is critical to ensure that the read-based phasing of variants in the region is accurate. We accomplish this using the two types of positions below.

2. **Positions where both proband and parent are heterozygous**. When considered in isolation, these positions do not provide direct information about haplotype transmission, because each of the two proband haplotypes could be identical-by-descent with the corresponding

parental haplotype. However, when considered in the context of the haplotype inferred to have been transmitted from the sequenced parent based on 1 above, these positions serve as a quality control for the accuracy of the phasing in the region, in both the proband and the parent.

3. **Positions where the proband is homozygous and the parent is heterozygous**. Like 2, these do not provide direct information about haplotype transmission in isolation, because each of the two proband haplotypes could be identical-by-descent with the parental haplotype that harbors the allele for which the proband is homozygous. However, when considered in the context of the haplotype inferred to have been transmitted from the sequenced parent based on 1 above, they serve as a quality control for the accuracy of phasing in the parent, thus lending further support to the accuracy of phasing in the proband as described in 2.

Consequently, when the Hamming distance between a pair of proband-parent haplotype blocks is equal to 0, we can infer that: a) the proband inherited this haplotype block from the sequenced parent, and not from the other (non-sequenced) parent; b) the phasing of the candidate variant based on which we have determined whether the variant is *de novo* or not is accurate.

Parameter sensitivity analysis

We examined the impact of *duoNovo*'s tuning parameters on the positive predictive value (PPV) and the number of *de novo* classifications, across all 80 duos.

First, we varied the threshold for the Hamming distance used to determine if a pair of proband-parent haplotype blocks is dissimilar and is thus not shared identical-by-descent (our default being 40). We found that a threshold of 0 always yields more *de novo* classifications compared to our default (as expected, since it is more lenient), but almost always yields a lower PPV, indicating that these classifications are enriched for false positives (Supplemental Figure S7, S8). Generally, we observed slight improvement when increasing thresholds up to 40, but no net gain when using thresholds above 40.

Second, we varied the distance threshold from the boundaries of haplotype blocks. By default,

duoNovo excludes variants within 2kb of the start/end coordinates of haplotype blocks. We found that using a more lenient threshold tends to yield more false positives, especially in mother-proband duos (potentially reflecting a drop in phasing accuracy at these boundaries; Supplemental Figure S9). The impact on the number of *de novo* classifications is minimal, as expected given that this threshold affects the inclusion of a comparatively small number of variants (Supplemental Figure S10).

We then varied the thresholds for GQ and sequencing depth. One would naturally expect that using more lenient thresholds for candidate variants themselves would lead to an inflation of false positives false positives. However, the impact of these thresholds when focusing on the positions surrounding candidate variants – which determine the Hamming distance between each proband-parent haplotype block pair – is less obvious. We found that, while using more lenient GQ thresholds (10 or 20 instead of 30) appears to increase the PPV (Supplemental Figure S11), this is because a much smaller number of variants get classified as *de novo*. This is likely due to the fact that a lower GQ causes haplotype block pairs that in reality are identical-by-descent to not have a Hamming distance of 0 due to sequencing errors, thus leading to missed *de novo* variants. Finally, a more lenient sequencing depth threshold (10 instead of 20) had an overall minimal impact (Supplemental Figure S12).

Examining sources of false positive classifications

As described above, from the resulting classifications *duoNovo* excludes variants clustered in the same haplotype block, and variants within regions annotated as problematic by the Genome-in-a-Bottle consortium (Dwarshuis et al., 2024). This is because we found that including these variants leads to an increased rate of false positive *de novo* classifications. Specifically, after pooling counts across all father-proband duos, we found that including variants within Genome-in-a-Bottle problematic regions reduced the precision from 98% to 90% in father-proband duoos, and from 91% to 71% in mother-proband duos. Moreover, including variants classified as *de novo* that are clustered in the same haplotype block reduced the precision to 70% in father-proband duos, and to 46% in mother-proband duos.

We also saw that, when only restricting to variants within genes (exonic and intronic regions either in coding or non-coding genes), there was no appreciable difference in the precision (97% in father-proband duos; 88% in mother-proband duos).

Variant Annotation

We annotated all variants using ANNOVAR (Wang et al., 2010). For each variant, annotations included genomic compartments (e.g. exonic, intronic, intergenic), CpG vs non-CpG context, gnomAD version 4.1 allele frequencies, and whether the variant falls into a GIAB-problematic region (Dwarshuis et al., 2024).

Relatedness and genetic ancestry calculation

Across all pairs of samples, relatedness was calculated using the "relate" function from Somalier (Pedersen et al., 2020) with default parameters. Genetic ancestry for each sample was predicted using the Somalier "ancestry" function using labeled data from the 1000 genomes project.

Mutation types

We calculated the number of occurrences of the different types of single nucleotide variants and indels per duo with the MutationalPatterns R package (Manders et al., 2022). For Supplemental Figure S4, percentages were calculated after pooling the counts across all father-proband duos, and all mother-proband duos.

Data and code availability

Sequencing data are available on AnVIL as part of the GREGoR Consortium data release (`https://gregorconsortium.org/data`). *duoNovo* is freely available for installation as an R package at `https://github.com/sbergercnmc/duonovo`.

Acknowledgements

Sample sequencing was performed through the UCI-GREGoR center, funded by NIH grant U01HG011745. LB was funded though the Children's National Hospital Pediatric Residency Research, Education, and Advocacy for Children's Health (REACH) Program.

Author contributions

SIB conceived of the study. LB and SIB designed and implemented *duoNovo*. LB and SIB processed and analyzed data. LB visualized data. CL, CF, PB, SB, BL, JH, KB, RK and WR performed the long-read sequencing and initial post-sequencing processing. LB, GP and SIB organized participant information. LB wrote the paper, which was then revised and edited by LB and SIB, with critical input from ED, GP, WR, JL, HAK, CX, EV. All authors reviewed the data and approved of the final manuscript.

Disclosure declaration

CL, CF, PB, and WR are employees and shareholders of Pacific Biosciences. SB, BL, JH, KB, and RK are employees of Ambry Genetics.

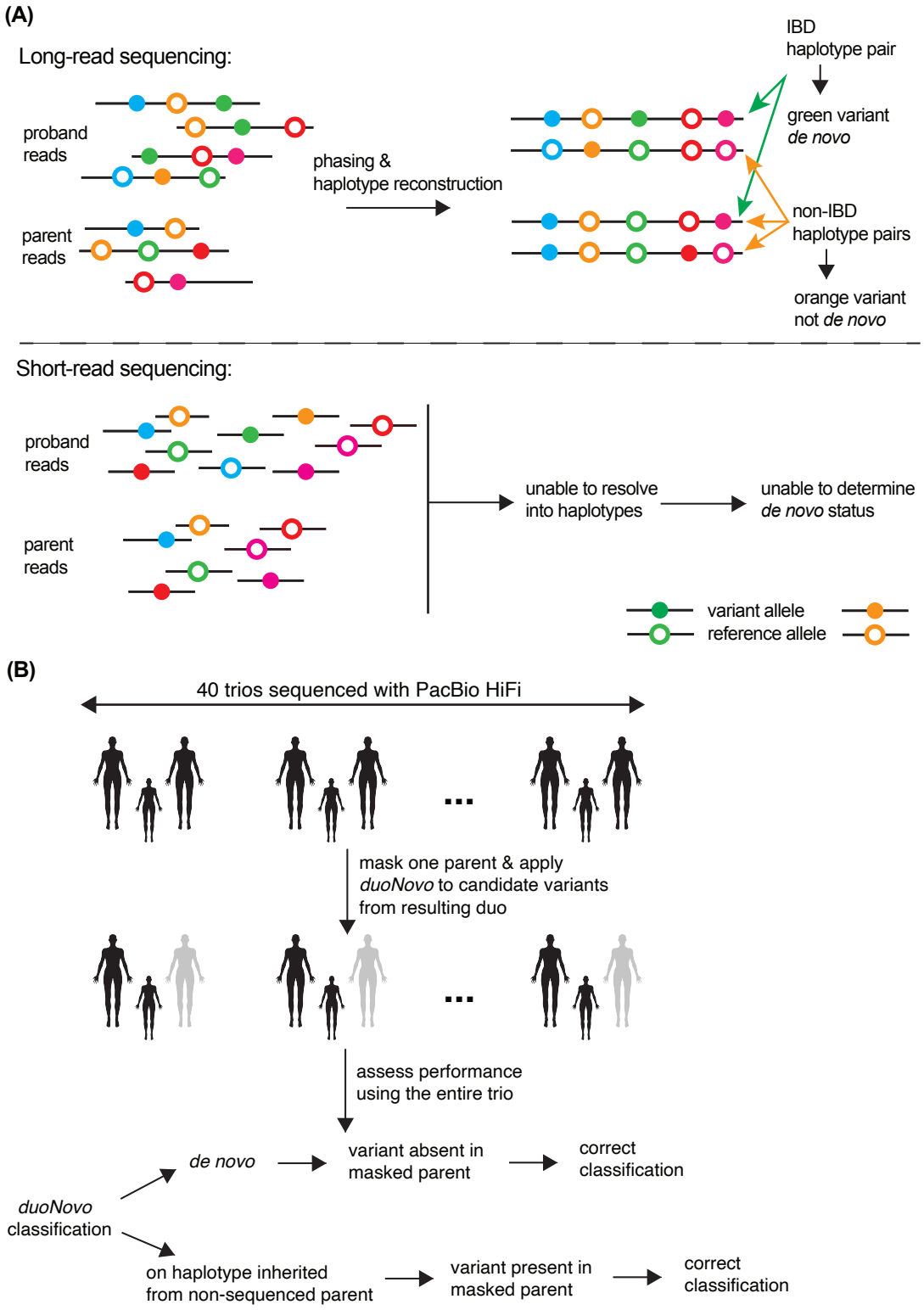


Figure 1. Detection of *de novo* variants from duos with *duoNovo*: conceptual basis and overview of strategy for performance testing.

Figure 1. Detection of *de novo* variants from duos with *duoNovo*: conceptual basis and overview of strategy for performance testing.

(A) The long reads produced by PacBio HiFi sequencing enable the read-backed phasing of variants and the subsequent reconstruction of haplotypes. *duoNovo* leverages this to identify pairs of proband-parent haplotypes that are identical-by-descent based on their sequence similarity. It then classifies candidate variants (heterozygous in the proband; absent in the parent) on such identical-by-descent haplotypes as *de novo* (orange variant in the cartoon figure; see also Methods). On the other hand, candidate variants on haplotypes not identical-by-descent with any of the two parent haplotypes (green variant in the cartoon figure) are inferred to have arisen on a haplotype inherited from the non-sequenced parent, in which case ascertaining their *de novo* status is not possible. This approach would not be feasible with short-read sequencing, as the short length of reads would preclude phasing and haplotype reconstruction. **(B)** We sequenced 40 trios with PacBio HiFi sequencing (Methods), and constructed father-proband and mother-proband duos by masking the mothers or the fathers, respectively. *duoNovo* was then applied to candidate variants from each duo, and the resulting variant classifications were evaluated based on the classifications that one would have obtained by having access to the entire trio.

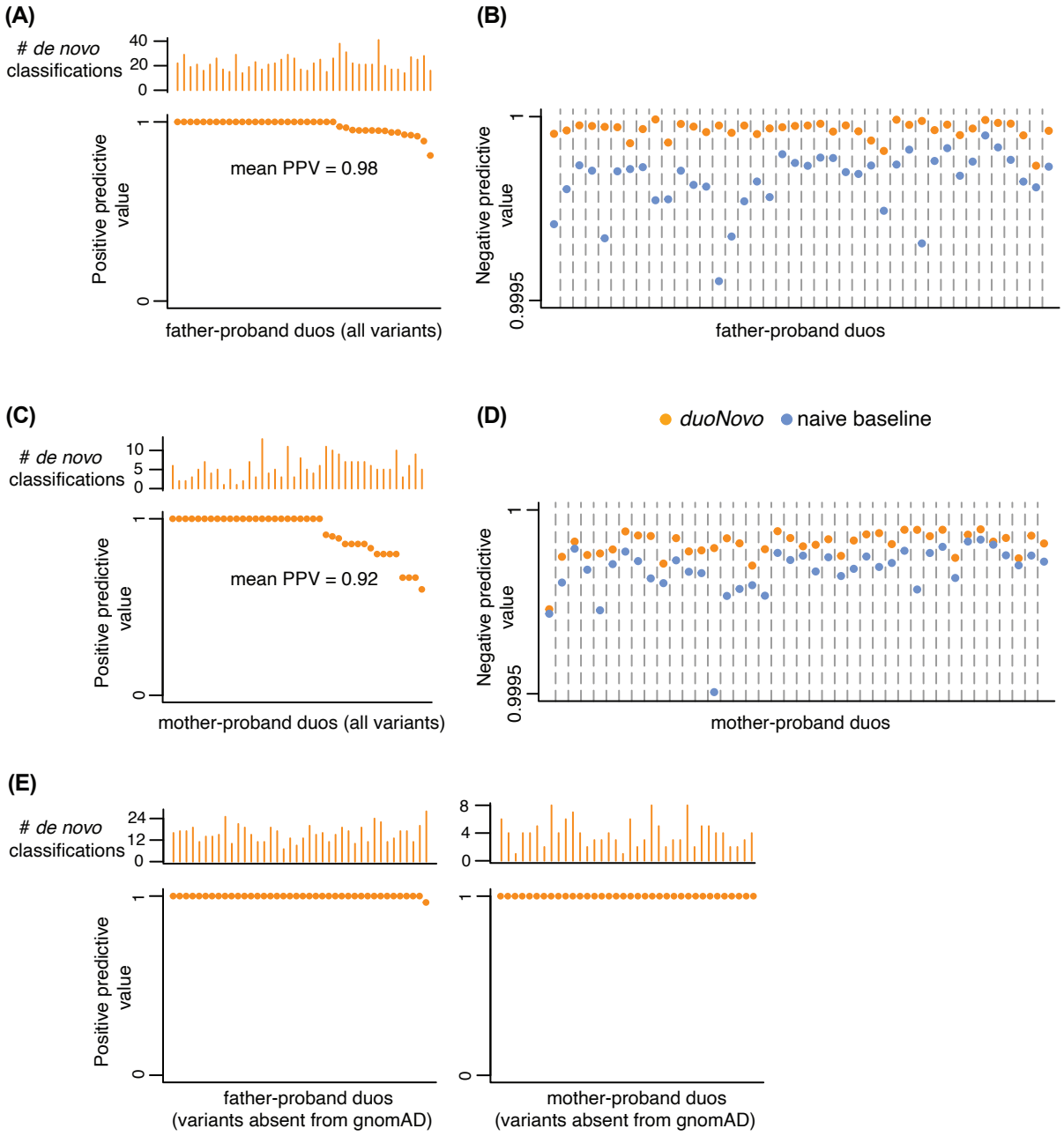


Figure 2. *duoNovo* has high positive and negative predictive value.

Figure 2. *duoNovo* has high positive and negative predictive value.

(A) The positive predictive value of *duoNovo* (bottom panel y axis) and the number of variants classified as *de novo* (top panel y axis axis). Each point and its associated vertical bar above correspond to a father-proband duo. **(B)** The negative predictive value (y axis) of *duoNovo* (orange points) and of the naive baseline approach (blue points) for each father-proband duo (indexed on the x axis). The naive baseline approach consists of classifying every candidate variant as non-*de novo*. **(C)** and **(D)** Like **(A)** and **(B)**, but for mother-proband duos. Note that in **(A)** and **(C)**, duos are ordered on the x axis in decreasing positive predictive value, and thus the order is different between **(A)** and **(C)**. **(E)** Like **(A)** and **(C)**, but restricting to variants that are absent from gnomAD 4.1. In 4 mother-proband duos, there were no *de novo* classifications among variants absent from gnomAD, hence these duos are not visualized.

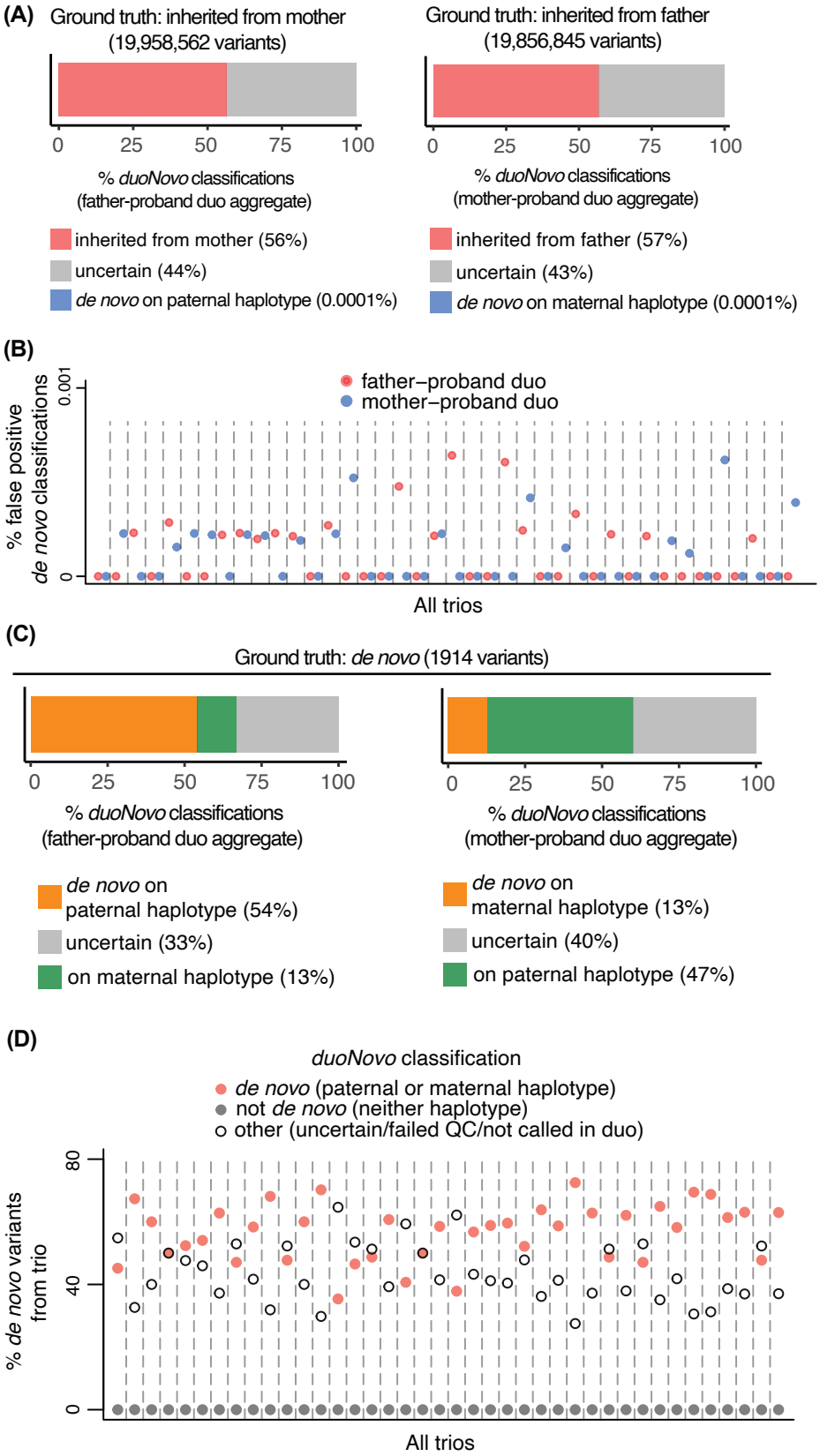


Figure 3. *duoNovo* has low error rate.

Figure 3. *duoNovo* has very low error rate.

(A) The percentage of candidate variants present in the non-sequenced parent (and thus inferred to be non-*de novo*) that are classified by *duoNovo* as present on the haplotype inherited from the masked parent, *de novo* or uncertain. Percentages (x axis) were computed after pooling counts across all father-proband or mother-proband duos. Variants that failed QC are excluded. **(B)** The false positive rate (fraction of non-*de novo* variants misclassified as *de novo*; y axis) of *duoNovo*, calculated separately for each father-proband and mother-proband duo. **(C)** The percentage of *de novo* variants detected from the entire trio (Methods) that received a classification from *duoNovo* (after QC filters) and were labeled as *de novo*, present on the haplotype inherited from the non-sequenced parent, or uncertain. Percentages (x axis) were computed after pooling counts across all father-proband or mother-proband duos. Variants that failed QC are excluded. **(D)** The percentage of trio-*de novo* variants classified by *duoNovo* as *de novo* (either in the father-proband or in the mother-proband duo), non-*de novo* (from both the father-proband and the mother-proband duo), or did not receive a classification. Each percentage is calculated separately for each trio.

References

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**.6: 779–786. DOI: 10.1038/s41588-021-00865-4.

Browning SR and Browning BL (2011). Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**.10: 703–714. DOI: 10.1038/nrg3054.

Brunet T, Jech R, Brugger M, Kovacs R, Alhaddad B, Leszinski G, Riedhammer KM, Westphal DS, Mahle I, Mayerhanser K, et al. (2021). De novo variants in neurodevelopmental disorders-experiences from a tertiary care center. *Clin. Genet.* **100**.1: 14–28. DOI: 10.1111/cge.13946.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**.7536: 608–611. DOI: 10.1038/nature13907.

Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, Bansal L, Bartik L, Baybayan P, Belden B, et al. (2022). Genomic answers for children: Dynamic analyses of ¿1000 pediatric rare disease genomes. *Genet. Med.* **24**.6: 1336–1348. DOI: 10.1016/j.gim.2022.02.007.

Dwarshuis N, Kalra D, McDaniel J, Sanio P, Alvarez Jerez P, Jadhav B, Huang WE, Mondal R, Busby B, Olson ND, et al. (2024). The GIAB genomic stratifications resource for human reference genomes. *Nat. Commun.* **15**.1: 9029. DOI: 10.1038/s41467-024-53260-y.

Figueroa KP, Gross C, Buena-Atienza E, Paul S, Gandelman M, Kakar N, Sturm M, Casadei N, Admard J, Park J, et al. (2024). A GGC-repeat expansion in ZFHX3 encoding polyglycine causes spinocerebellar ataxia type 4 and impairs autophagy. *Nat. Genet.* **56**.6: 1080–1089. DOI: 10.1038/s41588-024-01719-5.

Guo MH, Francioli LC, Stenton SL, Goodrich JK, Watts NA, Singer-Berk M, Groopman E, Darnowsky PW, Solomonson M, Baxter S, et al. (2024). Inferring compound heterozygosity from large-scale exome sequencing data. *Nat. Genet.* **56**.1: 152–161. DOI: 10.1038/s41588-023-01608-3.

Gustafson JA, Gibson SB, Damaraju N, Zalusky MP, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, De Coster W, et al. (2024). High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res.*: gr.279273.124. DOI: 10.1101/gr.279273.124.

Hiatt SM, Lawlor JMJ, Handley LH, Latner DR, Bonnstetter ZT, Finnila CR, Thompson ML, Boston LB, Williams M, Rodriguez Nunez I, et al. (2024). Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders. *Genome Res.* DOI: 10.1101/gr.279227.124.

Holt JM, Saunders CT, Rowell WJ, Kronenberg Z, Wenger AM, and Eberle M (2024). HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40**.2. DOI: 10.1093/bioinformatics/btae042.

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, et al. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**.7673: 519–522. DOI: 10.1038/nature24018.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**.7809: 434–443. DOI: 10.1038/s41586-020-2308-7.

Kessler MD, Loesch DP, Perry JA, Heard-Costa NL, Taliun D, Cade BE, Wang H, Daya M, Ziniti J, Datta S, et al. (2020). De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc. Natl. Acad. Sci. U. S. A.* **117**.5: 2560–2569. DOI: 10.1073/pnas.1902766117.

Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, Asri M, Alvarez Jerez P, Malik L, Dewan R, et al. (2023). Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat. Methods* **20**.10: 1483–1492. DOI: 10.1038/s41592-023-01993-x.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**.7412: 471–475. DOI: 10.1038/nature11396.

Kucuk E, Sanden BPGH van der, O'Gorman L, Kwint M, Derks R, Wenger AM, Lambert C, Chakraborty S, Baybayan P, Rowell WJ, et al. (2023). Comprehensive de novo mutation discovery with HiFi long-read sequencing. *Genome Med.* **15**.1: 34. DOI: 10.1186/s13073-023-01183-6.

Ligt J de, Willemsen MH, Bon BWM van, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, Vries P de, Gilissen C, et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**.20: 1921–1929. DOI: 10.1056/NEJMoa1206524.

Logsdon GA, Vollger MR, and Eichler EE (2020). Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**.10: 597–614. DOI: 10.1038/s41576-020-0236-x.

Manders F, Brandsma AM, Kanter J de, Verheul M, Oka R, Roosmalen MJ van, Roest B van der, Hoeck A van, Cuppen E, and Boxtel R van (2022). MutationalPatterns: the one stop shop for the analysis of mutational processes. *BMC Genomics* **23**.1: 134. DOI: 10.1186/s12864-022-08357-3.

Martin M, Patterson M, Garg S, O Fischer S, Pisanti N, Klau GW, Schöenhuth A, and Marschall T (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv*. DOI: 10.1101/085050.

Mastrorosa FK, Miller DE, and Eichler EE (2023). Applications of long-read sequencing to Mendelian genetics. *Genome Med.* **15**.1: 42. DOI: 10.1186/s13073-023-01194-3.

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, et al. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* **20**.1: 159–163. DOI: 10.1038/gim.2017.86.

Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, Lewis AP, Fuerte EPA, Paschal CR, Walsh T, et al. (2021). Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**.8: 1436–1449. DOI: 10.1016/j.ajhg.2021.06.006.

Mizuguchi T, Okamoto N, Yanagihara K, Miyatake S, Uchiyama Y, Tsuchida N, Hamanaka K, Fujita A, Miyake N, and Matsumoto N (2021). Pathogenic 12-kb copy-neutral inversion in

syndromic intellectual disability identified by high-fidelity long-read sequencing. *Genomics* **113**.1 Pt 2: 1044–1053. DOI: 10.1016/j.ygeno.2020.10.038.

Negi S, Stenton SL, Berger SI, Canigiula P, McNulty B, Violich I, Gardner J, Hillaker T, O'Rourke SM, O'Leary MC, et al. (2025). Advancing long-read nanopore genome assembly and accurate variant calling for rare disease detection. *Am. J. Hum. Genet.* DOI: 10.1016/j.ajhg.2025.01.002.

Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, and Morgan M (2014). VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**.14: 2076–2078. DOI: 10.1093/bioinformatics/btu168.

Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, Bronner MP, Underhill HR, and Quinlan AR (2020). Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**.1: 62. DOI: 10.1186/s13073-020-00761-2.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**.10: 983–987. DOI: 10.1038/nbt.4235.

Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, Walkiewicz M, Bi W, Xiao R, Ding Y, et al. (2017). Resolution of disease phenotypes resulting from multilocus genomic variation. *N. Engl. J. Med.* **376**.1: 21–31. DOI: 10.1056/NEJMoa1516767.

Rehm HL, Alaimo JT, Aradhya S, Bayrak-Toydemir P, Best H, Brandon R, Buchan JG, Chao EC, Chen E, Clifford J, et al. (2023). The landscape of reported VUS in multi-gene panel and genomic testing: Time for a change. *Genet. Med.* **25**.12: 100947. DOI: 10.1016/j.gim.2023.100947.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**.5: 405–424. DOI: 10.1038/gim.2015.30.

Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, and Quinlan AR (2019). Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife* **8**. DOI: 10.7554/eLife.46922.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Haeseler A von, and Schatz MC (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**.6: 461–468. DOI: 10.1038/s41592-018-0001-7.

Shojaeisaadi H, Schoenrock A, Meier MJ, Williams A, Norris JM, Palmer ND, Yauk CL, and Marchetti F (2024). Mutational signature analyses in multi-child families reveal sources of age-related increases in human germline mutations. *Commun. Biol.* **7**.1: 1451. DOI: 10.1038/s42003-024-07140-2.

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. (2024). Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**.10: 1571–1580. DOI: 10.1038/s41587-023-02024-y.

Steyaert W, Sagath L, Demidov G, Yépez VA, Esteve-Codina A, Gagneur J, Ellwanger K, Derks R, Weiss M, Ouden A den, et al. (2024). Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing. *medRxiv*. DOI: 10.1101/2024.05.03.24305331.

Vissers LELM, Ligt J de, Gilissen C, Janssen I, Steehouwer M, Vries P de, Lier B van, Arts P, Wieskamp N, Rosario M del, et al. (2010). A de novo paradigm for mental retardation. *Nat. Genet.* **42**.12: 1109–1112. DOI: 10.1038/ng.712.

Vollger MR, Korlach J, Eldred KC, Swanson E, Underwood JG, Bohaczuk SC, Mao Y, Cheng YHH, Ranchalis J, Blue EE, et al. (2025). Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition. *Nat. Genet.* DOI: 10.1038/s41588-024-02067-0.

Wang K, Li M, and Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**.16: e164. DOI: 10.1093/nar/gkq603.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**.10: 1155–1162. DOI: 10.1038/s41587-019-0217-9.

Wojcik MH, Lemire G, Berger E, Zaki MS, Wissmann M, Win W, White SM, Weisburd B, Wieczorek D, Waddell LB, et al. (2024). Genome sequencing for diagnosing rare diseases. *N. Engl. J. Med.* **390**.21: 1985–1997. DOI: 10.1056/NEJMoa2314761.

Wojcik MH, Reuter CM, Marwaha S, Mahmoud M, Duyzend MH, Barseghyan H, Yuan B, Boone PM, Groopman EE, Délot EC, et al. (2023). Beyond the exome: What's next in diagnostic testing for Mendelian conditions. *Am. J. Hum. Genet.* **110**.8: 1229–1248. DOI: 10.1016/j.ajhg.2023.06.009.

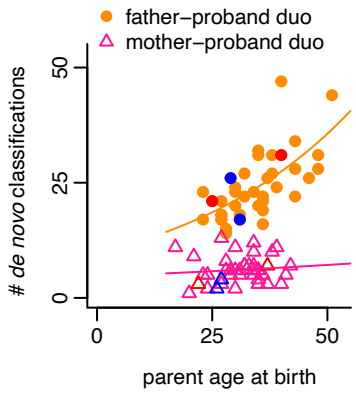
Wright CF, Campbell P, Eberhardt RY, Aitken S, Perrett D, Brent S, Danecek P, Gardner EJ, Chundru VK, Lindsay SJ, et al. (2023). Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N. Engl. J. Med.* **388**.17: 1559–1571. DOI: 10.1056/NEJMoa2209046.

Xie Z, Sun C, Zhang S, Liu Y, Yu M, Zheng Y, Meng L, Acharya A, Cornejo-Sanchez DM, Wang G, et al. (2020). Long-read whole-genome sequencing for the genetic diagnosis of dystrophinopathies. *Ann. Clin. Transl. Neurol.* **7**.10: 2041–2046. DOI: 10.1002/acn3.51201.

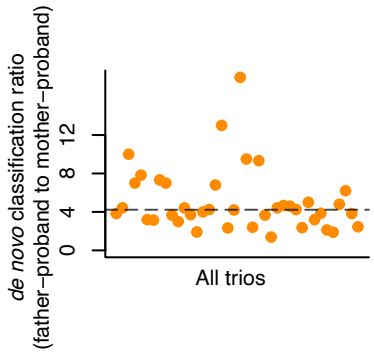
Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**.18: 1870. DOI: 10.1001/jama.2014.14601.

Yun T, Li H, Chang PC, Lin MF, Carroll A, and McLean CY (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**.24: 5582–5589. DOI: 10.1093/bioinformatics/btaa1081.

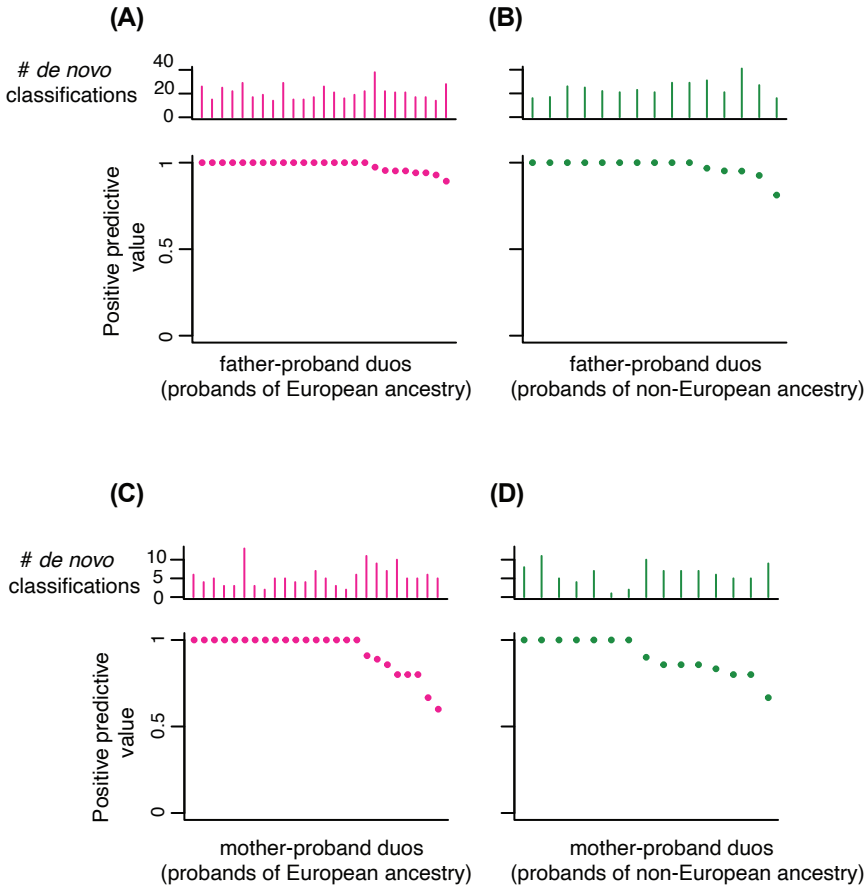
Supplemental Figures



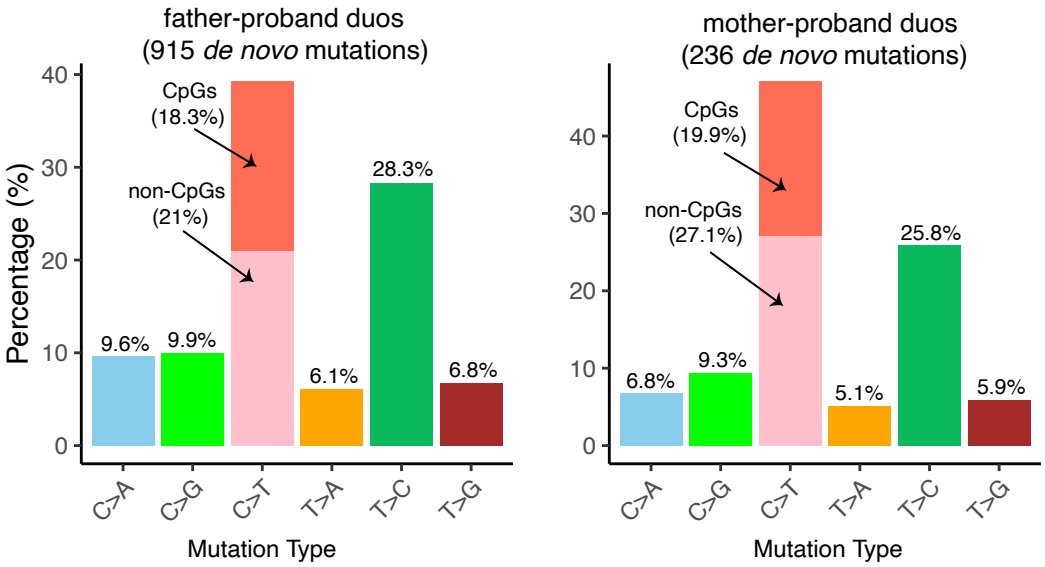
Supplemental Figure S1. Paternal age positively correlates with the number of *de novo* classifications from father-proband duos. Each point corresponds to a duo. Trend lines were plotted after fitting Poisson regression models using the log link. The red and blue points correspond to 2 father-proband and 2 mother-proband duos that share the same father and mother, respectively (from the 2 quad families that were each split into 2 trios).



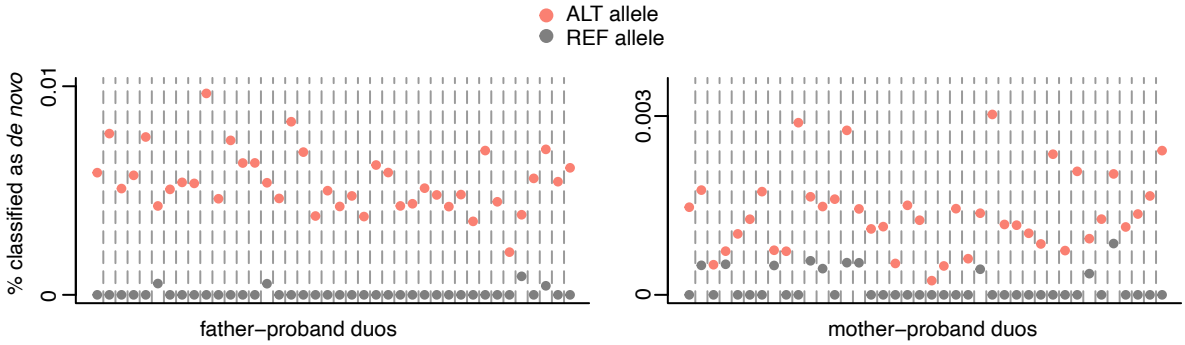
Supplemental Figure S2. *duoNovo* detects more *de novo* variants from father-proband duos compared to mother-proband duos. Each point corresponds to a trio and its position on the y axis corresponds to the ratio of the number of *de novo* classifications from the father-proband duo to the number of *de novo* classifications from the mother-proband duo. The dashed horizontal line corresponds to the median across all trios.



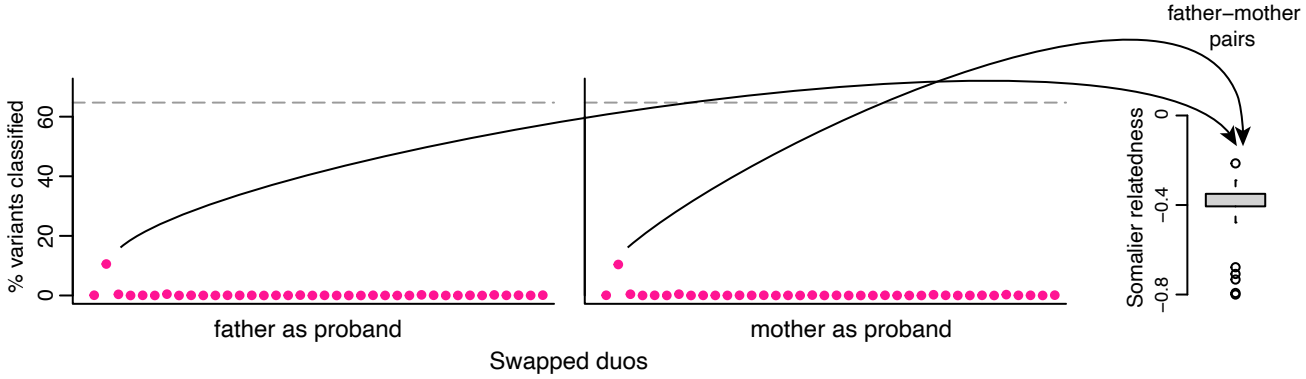
Supplemental Figure S3. No difference in the positive predictive value of *duoNovo* in probands of European vs non-European ancestry. Like Figure 2A, C, but shown separately for duos with probands of European versus non-European ancestry. Ancestry was calculated using Somalier (Pedersen et al., 2020). Non-European ancestry groups include African (2 probands), Admixed American (7 probands), East Asian (3 probands), and South Asian (3 probands).



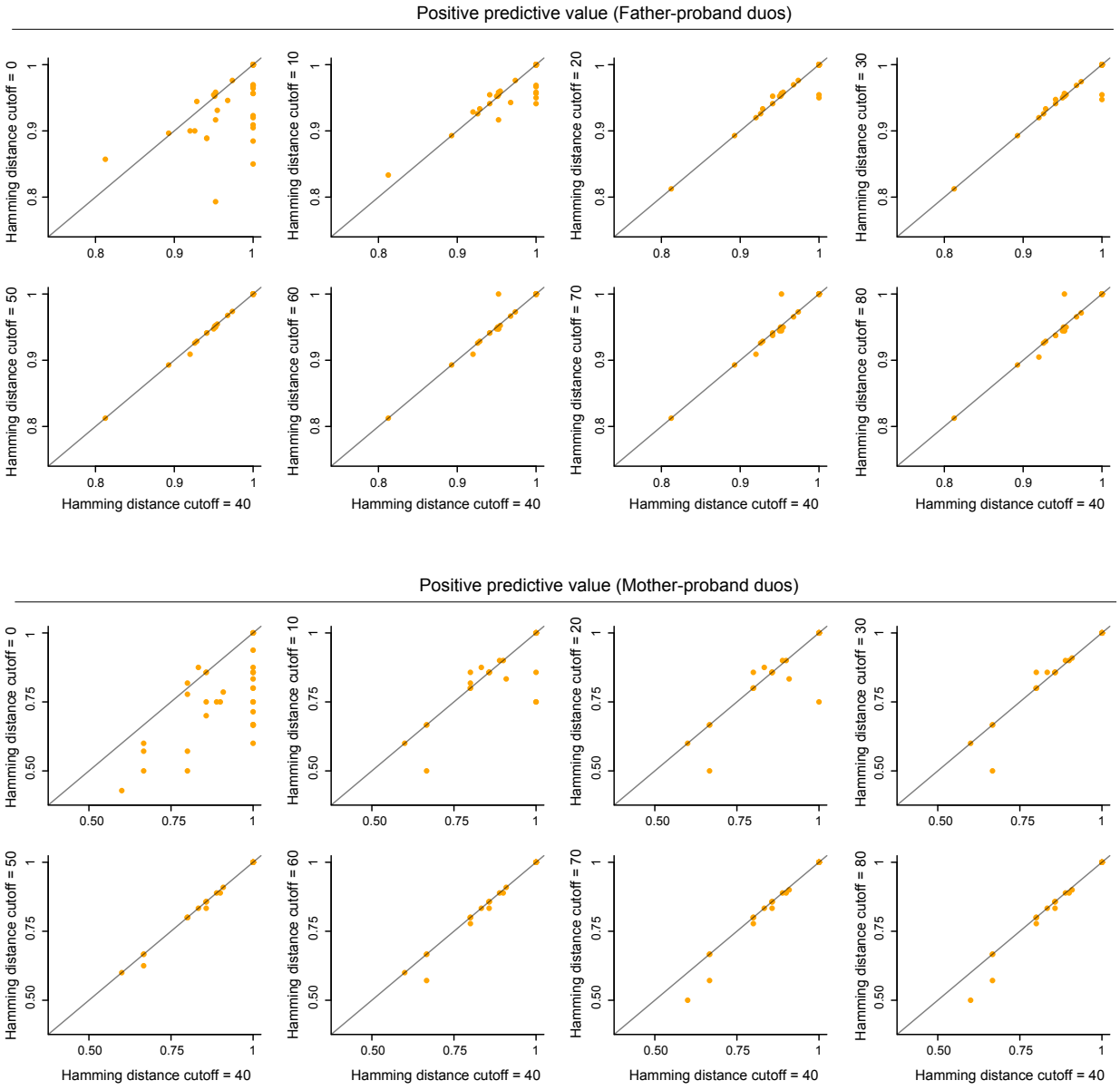
Supplemental Figure S4. The *de novo* single nucleotide variants identified by *duoNovo* fall into expected mutation subtypes. Percentages were computed after pooling all single nucleotide variants classified as *de novo* from the father-proband duos (left) and the mother-proband duos (right).



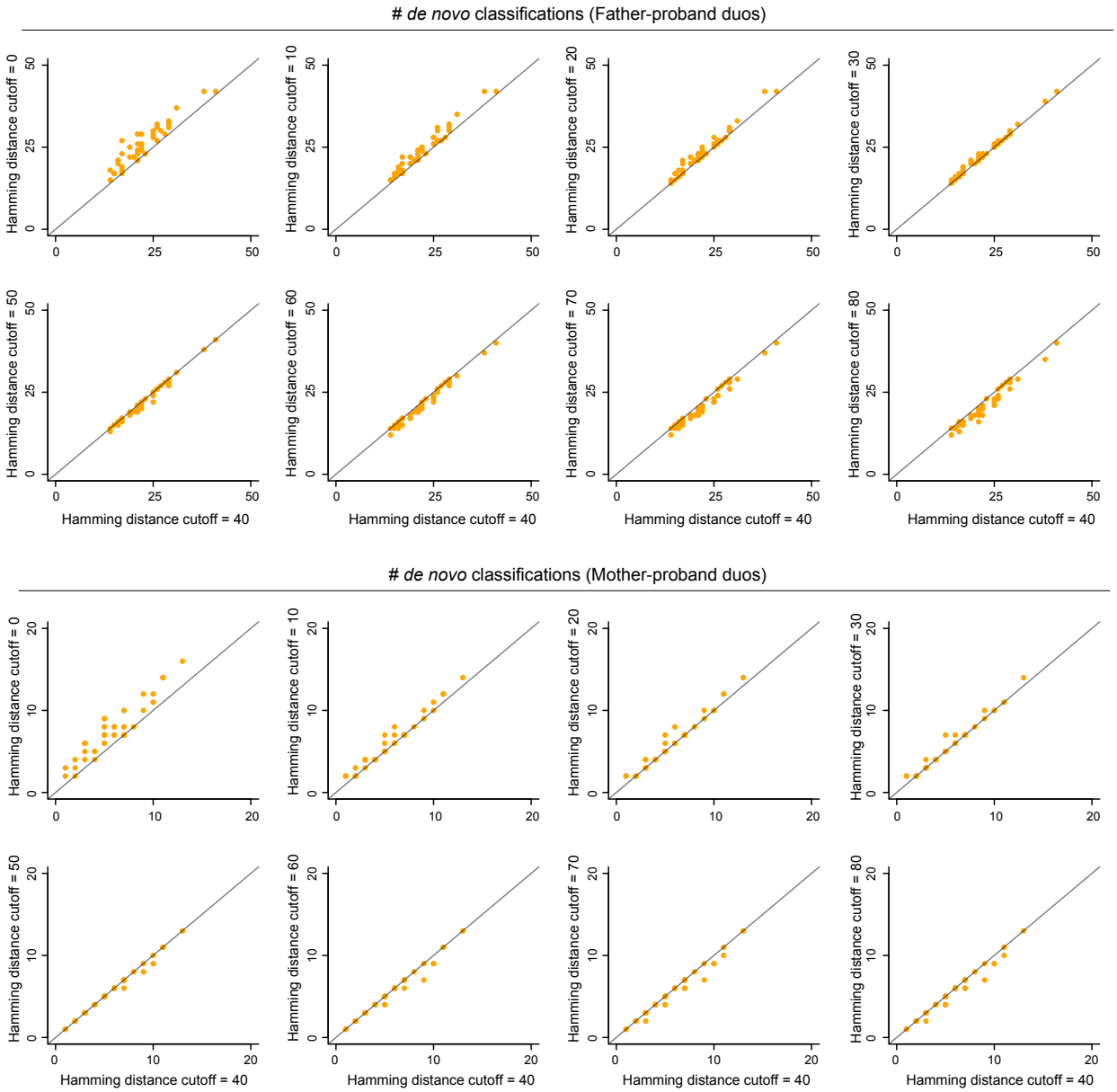
Supplemental Figure S5. *duoNovo* very rarely classifies candidate reference alleles as *de novo*. Candidate alternative (ALT) alleles were identified by finding positions where the proband had the "1|0" or "0|1" genotype and the parent had the "0/0" genotype. Conversely, candidate reference (REF) alleles were identified by finding positions where the proband had the "1|0" or "0|1" genotype and the parent had the "1/1" genotype. Positions were either the proband or the parent failed QC filters (Methods) were excluded.



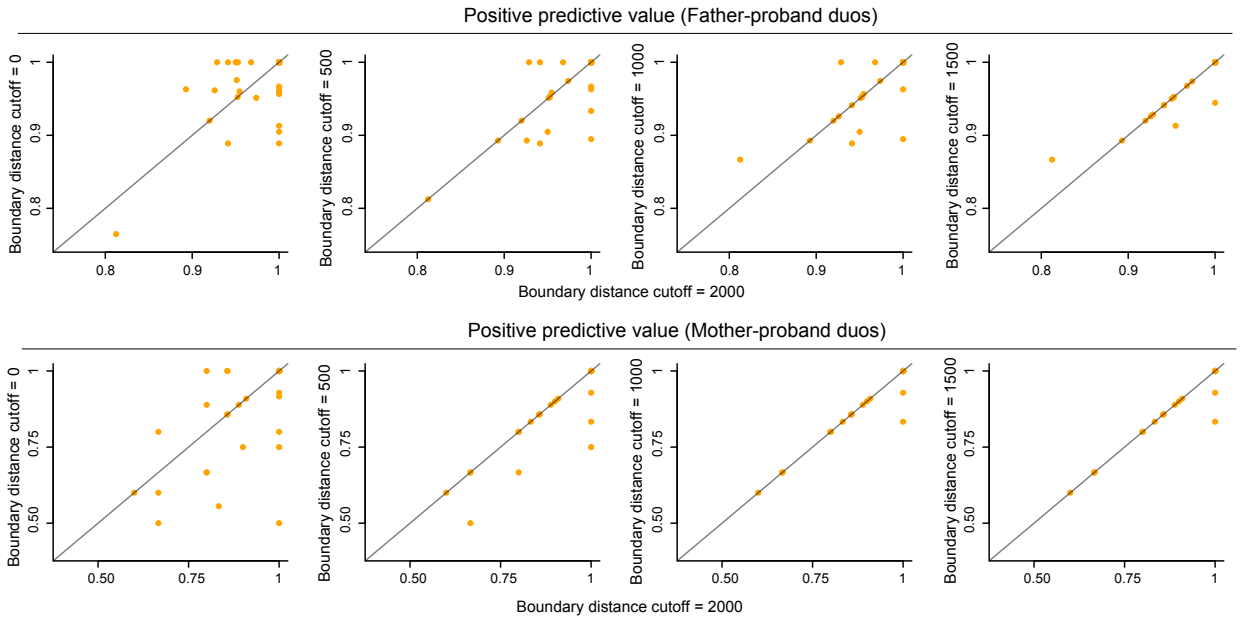
Supplemental Figure S6. *duoNovo* does not classify variants from swapped duos. Each point corresponds to a swapped duo, and its position on the y axis indicates the percentage of variants that received a classification (either *de novo* or on the non-sequenced parent haplotype). Each swapped duo consists of the two parents instead of a parent and the proband; *duoNovo* was applied to each swapped duo by either treating the father as the proband (left) or the mother as the proband (right). The dashed horizontal line corresponds to the median percentage of variants that received a classification across all regular (father-proband or mother-proband) duos. The rightmost panel depicts the distribution of relatedness coefficients between all the fathers and mothers (each point corresponds to a father-mother pair), showing that the outlier swapped duo is the one with the highest relatedness coefficient.



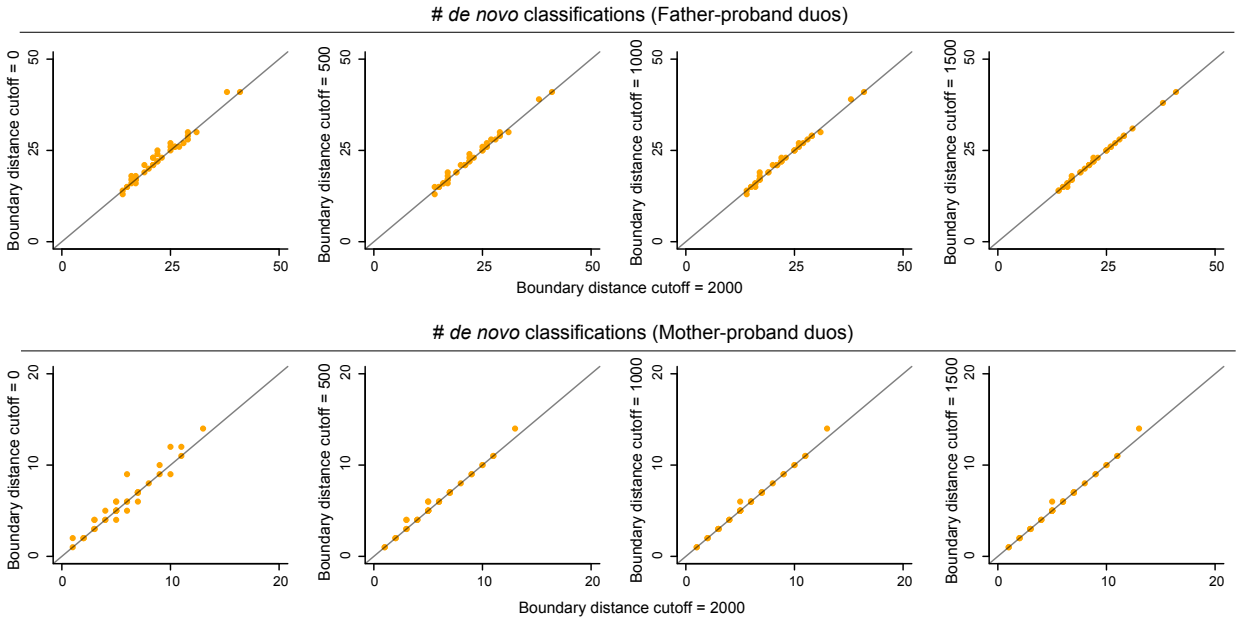
Supplemental Figure S7. The impact of the Hamming distance threshold for determining dissimilarity between a pair of haplotype blocks on the positive predictive value. Each point corresponds to a duo. The x-axis shows the positive predictive value (PPV) when using a Hamming distance threshold of 40 (default), and the y axis shows the PPV when using different Hamming distance thresholds. There is some degree of overplotting, due to duos that have the same or highly similar PPV.



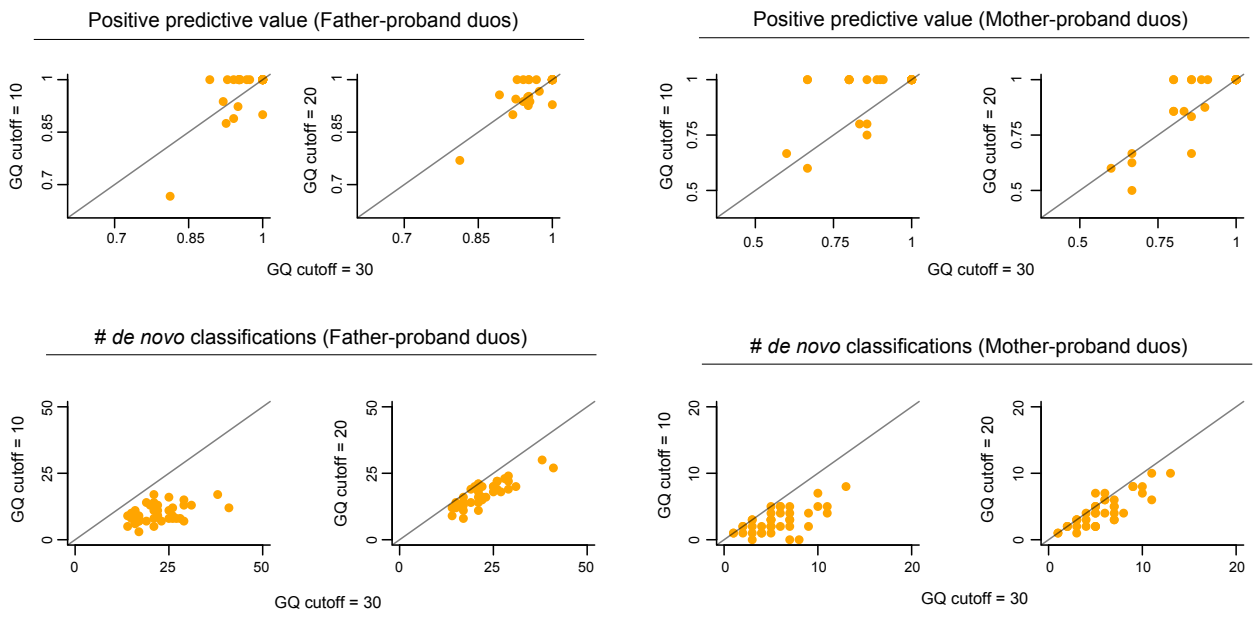
Supplemental Figure S8. The impact of the Hamming distance threshold for determining dissimilarity between a pair of haplotype blocks on the number of *de novo* classifications. Like Supplemental Figure S7, but depicting the number of *de novo* classifications instead of the PPV.



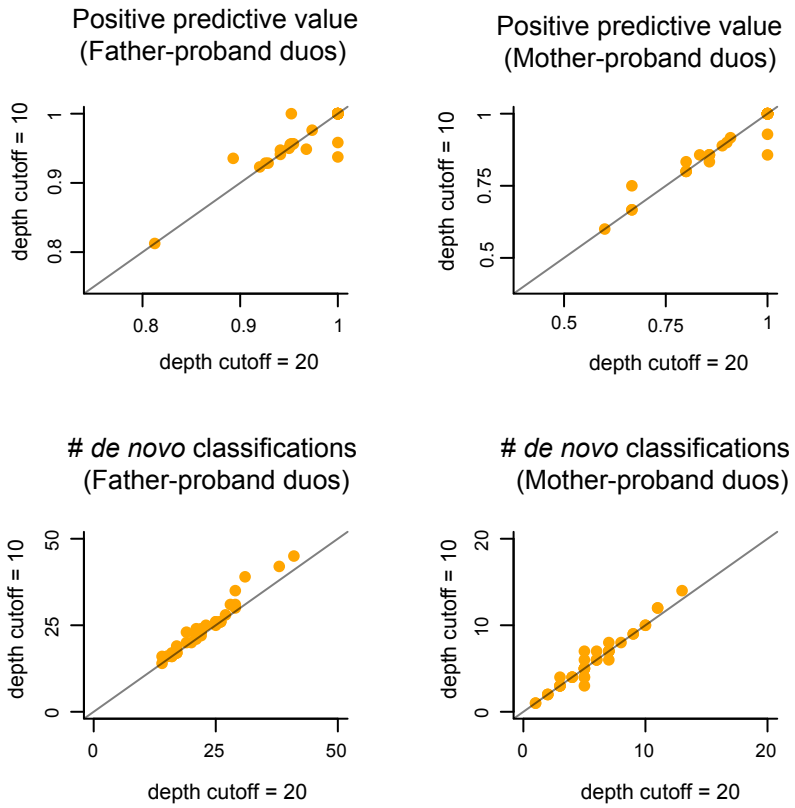
Supplemental Figure S9. The impact of the distance threshold from haplotype block boundaries on the positive predictive value. Like Supplemental Figure S7, but showing the impact of the threshold for the distance from the haplotype block boundaries (start/end coordinates) on the PPV.



Supplemental Figure S10. The impact of the distance threshold from haplotype block boundaries on the number of *de novo* classifications. Like Supplemental Figure S8, but showing the impact of the threshold for the distance from the haplotype block boundaries (start/end coordinates) on the number of *de novo* classifications.



Supplemental Figure S11. The impact of the phred quality (GQ) threshold at positions surrounding candidate variants on the positive predictive value and the number of *de novo* classifications. Like Supplemental Figure S7 and Supplemental Figure S8, but showing the impact of the GQ threshold for positions surrounding the candidate variant (that is, positions which determine the Hamming distance) on the PPV and the number of *de novo* classifications.



Supplemental Figure S12. The impact of the sequencing depth threshold at positions surrounding candidate variants on the positive predictive value and the number of *de novo* classifications. Like Supplemental Figure S11, but showing the impact of the sequencing depth threshold for positions surrounding the candidate variant (that is, positions which determine the Hamming distance) on the PPV and the number of *de novo* classifications.