Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation

Rocio Acuna-Hidalgo,[1] Tan Bo,[2] Michael P. Kwint,[1] Maartje van de Vorst,[1] Michele Pinelli,[3] Joris A. Veltman,[1,4] Alexander Hoischen,[1,5,*] Lisenka E.L.M. Vissers,[1,5] and Christian Gilissen[1,5]

De novo mutations are recognized both as an important source of genetic variation and as a prominent cause of sporadic disease in humans. Mutations identified as de novo are generally assumed to have occurred during gametogenesis and, consequently, to be present as germline events in an individual. Because Sanger sequencing does not provide the sensitivity to reliably distinguish somatic from germline mutations, the proportion of de novo mutations that occur somatically rather than in the germline remains largely unknown. To determine the contribution of post-zygotic events to de novo mutations, we analyzed a set of 107 de novo mutations in 50 parent-offspring trios. Using four different sequencing techniques, we found that 7 (6.5%) of these presumed germline de novo mutations were in fact present as mosaic mutations in the blood of the offspring and were therefore likely to have occurred post-zygotically. Furthermore, genome-wide analysis of "de novo" variants in the proband led to the identification of 4/4,081 variants that were also detectable in the blood of one of the parents, implying parental mosaicism as the origin of these variants. Thus, our results show that an important fraction of de novo mutations presumed to be germline in fact occurred either post-zygotically in the offspring or were inherited as a consequence of low-level mosaicism in one of the parents.

Introduction

In humans, DNA replication is estimated to entail one error in every 10^8 base pairs, giving rise to 30–100 genome-wide de novo mutations in each new generation.[1–3] Whereas neutral or benign de novo point mutations contribute to normal genetic variation, single detrimental de novo mutations have been established to cause a number of rare developmental disorders[4–6] and are increasingly recognized as a major contributor to common sporadic disorders, such as intellectual disability (ID) and autism.[7,8] De novo mutations are thought to occur predominantly in the egg or sperm cell and thus result in an embryo with a constitutive mutation. However, de novo mutations can also appear post-zygotically, leading to embryonic mosaicism, a state in which two or more genetically distinct cell populations in an individual develop from a single fertilized egg.

Several reports have shown a high frequency of mosaicism for copy-number variations (CNVs) from cleavage-stage embryos[9] to fully differentiated tissues.[10–12] Similarly, there is increasing evidence of a high prevalence of mosaicism for single-nucleotide variants (SNVs) as a result of mutations appearing from early embryogenesis onward[13,14] and throughout adult life.[15,16] Currently, post-zygotic de novo mutations receive growing attention in developmental diseases.[17–19] The timing of the event plays a key role in the clinical phenotype by determining not only the proportion of affected cells in the organism but also the type of tissues involved.[18] Despite its pervasiveness, however, the true extent of mosaicism for SNVs remains unclear. This is largely a consequence of the technological limitations to accurately detecting these mutations; on one side, mutations with low levels of mosaicism are often below the threshold of sensitivity and specificity for automated and systematic detection of traditional sequencing methods,[20] and on the other hand, mutations with a higher percentage of affected cells are easily detected by traditional sequencing methods, but it remains technically challenging to differentiate them from germline de novo mutations. Indeed, to discriminate post-zygotic from germline de novo mutations by sequencing DNA, it is crucial to distinguish biologically relevant allele imbalances from technical artifacts.

To gain insight into the frequency of post-zygotic events among de novo mutations, we performed a systematic evaluation of de novo mutations identified by trio-based whole-genome sequencing (WGS) of 50 individuals with severe ID and their parents. Previous analysis of WGS data from this cohort recently pointed to germline de novo mutations as the major cause of ID in the affected individuals.[21] Additionally, these data indicated the presence of de novo mutations of somatic origin.[21] By systematically assessing allelic ratios by various sequencing techniques, we show here that a proportion of previously reported de novo mutations did not occur during gametogenesis but, in fact, arose as post-zygotic events in the proband or were present as low-level somatic mutations in one of the parents.

[1]Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Institute of Neuroscience, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands; [2]State Key Laboratory of Medical Genetics, Central South University, 110 Xiangya Road, Changsha, Hunan 410078, China; [3]Telethon Institute of Genetics and Medicine, Pozzuoli, 80078 Naples, Italy; [4]Department of Clinical Genetics, Maastricht University Medical Centre, Universiteitssingel 50, 6229 ER Maastricht, the Netherlands
[5]These authors contributed equally to this work
*Correspondence: alexander.hoischen@radboudumc.nl
http://dx.doi.org/10.1016/j.ajhg.2015.05.008.

Material and Methods

Defining a Set of De Novo Mutations from WGS of Parent-Proband Trios

This study was performed in accordance with the ethical standards of the Medical Ethics Committee of the Radboud University Medical Center. All participants or their legal representatives gave informed consent. WGS of 50 parent-proband trios and subsequent de novo mutation detection were performed as described previously.[21] In brief, trio-based WGS was performed by Complete Genomics (CG) at 80-fold coverage. Sequence reads were mapped to the reference genome (UCSC Genome Browser hg19), and variants were called with CG software v.2.4. De novo mutations were called with CG's cgatools calldiff program, which detects the differences between the genotypes of two samples and assigns a somatic score on the basis of sequencing quality and comparison of paired samples. Mutations whose scores comparing offspring to each parent were ≥ 5 were called as high-confidence de novo mutations (a total of 4,081 were detected in the 50 trios). The original report identified a set of 127 de novo mutations affecting either genome-wide coding sequence or specifically the non-coding sequence of known ID-associated genes.[21] This set served as the starting point for the current study.

Sequencing Methods Used for Assessing the Post-zygotic State of De Novo Mutations

PCR amplicons for amplicon-based deep sequencing (ADS) and Sanger sequencing were generated according to standard PCR protocols. ADS was performed on an Ion Torrent Personal Genome Machine (Life Technologies) as described previously.[21] In brief, raw sequencing reads were mapped to the reference genome with the Burrows-Wheeler Aligner (BWA), and the alignment files were then analyzed in the Integrative Genomics Viewer (IGV).[22] For Sanger sequencing, PCR products were sequenced after enzymatic clean up.

Sequencing using single-molecule molecular inversion probes (smMIPs) was performed according to previously published protocols.[23] In brief, smMIPs targeting the selected de novo mutation and a total of 112 bp of surrounding sequence were designed in house and ordered from Integrated DNA Technologies. The smMIPs were pooled and phosphorylated, after which the genomic regions of interest were captured with the probes and amplified. Sequencing was performed on the NextSeq 500 Desktop Sequencer (Illumina), and the reads were aligned with our in-house bioinformatics pipeline for analysis of molecular inversion probes. Through the use of molecular barcodes, we were able to remove PCR duplicates. Read counts for the positions of interest were extracted from the alignment files through IGV.

Assessment of the Allelic-Ratio Distribution of True Heterozygous Variants

To define the parameters of technical variation in WGS, ADS, and Sanger sequencing, we determined for each technology the allelic ratio of inherited SNVs as a proxy for true heterozygous mutations. The allelic ratio was defined as the proportion of variant reads from the total number of sequencing reads covering a given base pair and is expressed here as a percentage. We established the distribution of the allelic ratio for true heterozygous variants in WGS data by determining the allelic ratio of 115 inherited SNVs (coding, synonymous variants absent from dbSNP138 or present at a frequency below 1.5%) from WGS data of a single individual. To minimize the risk of false-positive variant calls, we used a second independent set of 109 inherited SNVs to determine the distribution of the allelic ratio in ADS and Sanger sequencing. This set was randomly selected from a larger set of 442 rare, coding variants inherited from either parent in ten probands, and variants were selected to have a coverage $\geq$ 20-fold in WGS and an allelic ratio between 40% and 60%. Variants on the X chromosome and/or located in established disease-associated genes were excluded. For ADS experiments, after mapping with the BWA, variants were visualized with the IGV, and allelic ratios were determined by assessment of the number of total reads and each respective base at this position. For Sanger sequencing, the chromatogram trace files were visualized with Vector NTI (Life Technologies), and intensities per dye per variant base were used for calculating the allelic ratio.

Identification of Post-zygotic Events in Probands

A set of 127 de novo mutations identified by WGS were re-sequenced by ADS and Sanger sequencing. For 107 (84%) of these variants, allelic ratios could be determined for all three sequencing techniques. We calculated the individual Z score per method for each mutation by using the values from sequencing heterozygous variants with each sequencing method as a reference. To calculate Z scores, we first obtained the difference between the allelic ratio and the mean allelic ratio and then divided that by the SD for heterozygous variants on that sequencing technique. Subsequently, we combined these scores into a single Z score for each de novo mutation by summing the individual Z scores and dividing this total by the square root of the number of scores. The critical value for statistical significance was established as 0.05 after Benjamini-Hochberg correction for multiple testing. To exclude amplification bias as the cause of a deviation in the allelic ratio, we re-sequenced de novo SNVs with a statistically significant combined Z score by ADS with a second independent primer pair. Finally, we used smMIPs as an independent technique to validate the presence of these variants as mosaic mutations (a set of seven heterozygous mutations served as a reference).

Identification of Parental Mosaicism in WGS Data

To detect low-level parental mosaicism for SNVs mimicking germline de novo mutations in the child, we re-analyzed the WGS data of the 50 parent-offspring trios. To this end, we used all 4,081 high-confidence candidate de novo mutations identified in the probands, because these have previously been shown to have a de novo validation rate of 78%.[21] We then filtered for de novo variants for which at least two reads carrying the same mutation in the raw sequencing data were found in either one of the parents. We sequenced the position of interest by ADS in the DNA of the transmitting parent to validate parental mosaicism for the remaining 11 mutations. We established the position-specific sequencing error rate by sequencing the same position by ADS in the DNA of the non-transmitting parent in an independent sequencing run to avoid any contamination or barcode bleed-through. Then, the fraction of reads showing a non-reference allele at the corresponding base pair was calculated. The presence of the variant as a mosaic mutation in the transmitting parent was confirmed if the proportion of variant reads for the position and nucleotide of interest was significantly higher than the sequencing error established for that base-pair position from the non-transmitting parent.

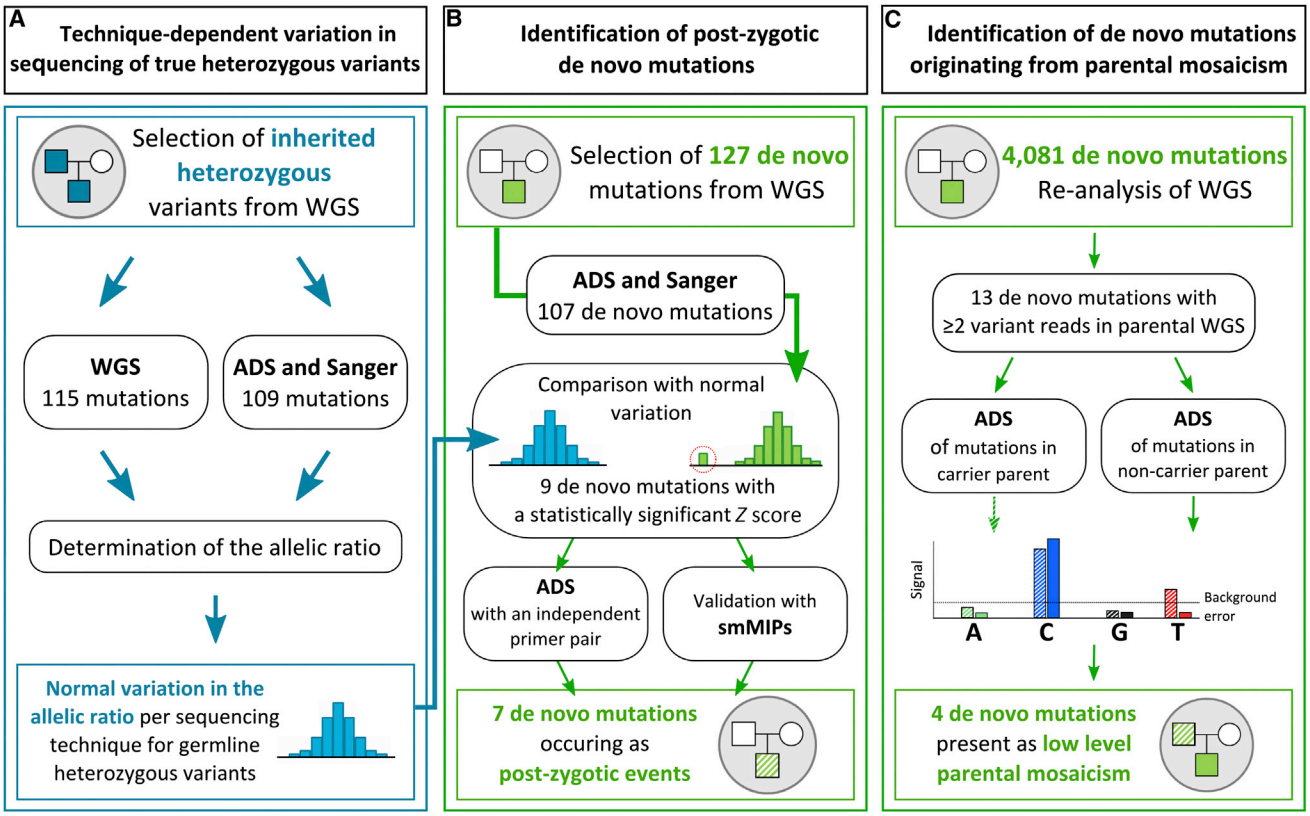

Figure 1. Workflow for the Detection of Mosaic Mutations among a Subset of Apparently De Novo Mutations
(A) Assessment of technique-dependent variation in sequencing of two groups of heterozygous germline variants (in blue) for determining the distribution of allelic ratios for three different techniques (WGS, ADS, and Sanger sequencing).
(B) Previously identified de novo mutations were re-sequenced by ADS and Sanger sequencing for determining the variant ratio. With the use of the combined *Z* score, nine putative somatic variations were identified. They were then validated by ADS with a second independent primer pair and smMIPs. Seven of nine were confirmed to deviate in allelic ratio, suggesting a non-germline event.
(C) Identification of de novo mutations originating from parental mosaicism. Of 4,081 high-confidence de novo mutations identified by WGS, 13 were identified to have two or more variant reads in parental DNA. With the use of ADS data from the non-carrier parent for correcting for the background sequencing error, four mutations appearing as de novo in the child were identified as low-level mosaicism in one of the parents.

Computational Modeling of Sequencing Coverage for the Identification of Mosaicism

To assess the ability of identifying mosaic variants from sequencing data, we simulated the effect of sequencing coverage on variant identification for different levels of mosaicism. To distinguish low-level mosaicism from sequencing artifacts, we assumed that automated variant-calling algorithms require the variant to be present in ≥ 5 sequencing reads and constitute $\geq 5\%$ of the total number of reads at the position of interest. We used a binomial distribution to calculate the probability of reaching both these requirements for different depths of coverage and various levels of mosaicism. Assuming that a mosaic variant is identified, we also modeled the deviation of the allelic ratio from 50% (representing true heterozygosity), which is necessary for distinguishing a mosaic from a germline variant. Reads for heterozygous variants at different sequencing depths were simulated (n = 10,000) on the basis of a binomial distribution. We calculated the SD of this distribution and the level of mosaicism at which a mosaic variant could be reliably distinguished from a heterozygous variant for different thresholds of significance. Lastly, we determined the sequence coverage required for identifying low-level parental mosaicism. In this case, the position of interest was readily identified because the offspring presented with an apparently de novo mutation at this position. For this, we considered that at least two variant reads were sufficient for distinguishing the variant from a background sequencing error. Finally, we applied a binomial model for different sequencing depths and levels of mosaicism to calculate the probability of obtaining two variant reads in the sequencing data.

Results

Determining the Technical Variation for WGS, ADS, and Sanger Sequencing

In this study, we set out to distinguish mosaic mutations from true germline de novo mutations (Figure S1) by sequencing. To gain insight into the sensitivity of WGS, ADS, and Sanger sequencing, we re-sequenced two different sets of inherited germline mutations as a proxy for true heterozygosity (Figure 1A and Figure S2). We subsequently determined the distribution of the allelic ratios per technology (Table S1 and Figure S3). With an allelic ratio of 48.2 ± 4.4% (average ± SD), ADS showed to be the

Table 1. De Novo Mutations Occurring as Post-zygotic Events in Offspring

Gene	OMIM Accession Number	Mutation at gDNA Level (hg19)	Location	Predicted Mutation at cDNA Level (GenBank Accession Number)	Predicted Protein Substitution	p Value[a]	Average Allelic Ratio
KANSL2	615488	chr12:49072911C>A	exon 4	c.453G>T (NM_017822.3)	p.(=)	6.94E−21	20.8%
CREBL2	603476	chr12:12788868G>C	exon 2	c.173G>C (NM_001310.2)	p.Arg58Pro	6.40E−19	21.0%
PIAS1	603566	chr15:68468014T>A	exon 10	c.1209T>A (NM_016166.1)	p.Asp403Glu	1.84E−18	22.9%
PNKP	605610	chr19:50367525C>T	intron 5	c.579−32G>A (NM_007254.3)	NA	7.05E−17	22.7%
HIVEP2	143054	chr6:143092683C>T	exon 5	c.3193G>A (NM_006734.3)	p.Ala1065Thr	2.20E−14	25.2%
DPYD	274270	chr1:97588236C>T	intron 21	c.2623−24048G>A (NM_000110.3)	NA	3.17E−10	29.7%
NEK1	604588	chr4:170359295T>G	exon 27	c.2703A>C (NM_001199397.1)	p.Lys901Asn	3.67E−08	29.4%

The following abbreviation is used: NA, not applicable.
[a]p values were corrected by Benjamini-Hochberg for multiple testing. The level of the mutation was calculated as the average variant ratio for each mutant from all sequencing methods.

most precise technique for identifying true heterozygosity. In comparison, WGS showed an allelic ratio of 50.5 ± 8.9%, and Sanger sequencing had a ratio of 51.4 ± 8.7% (Table S2). On the basis of the obtained distributions for the allelic ratio, we determined that de novo mutations with an allelic ratio below 32.8% for WGS, 39.3% for ADS, and 33.9% for Sanger sequencing had a statistically significant deviation from the expected ratio for true heterozygous mutations and might, as such, reflect mosaic mutations.

Identification of Post-zygotic De Novo Mutations in Probands

Our next objective was to determine the proportion of post-zygotic events among a subset of de novo mutations in our cohort. For this, we studied a pre-defined set of 107 de novo mutations by using WGS, ADS, and Sanger sequencing (Figure 1B).[21] As we did for the inherited variants, we determined each mutation's allelic ratio for each sequencing technique. After calculation of the mean allelic ratio across the three sequencing techniques, nine de novo mutations showed a statistically significant deviation from the expected ratio for true germline heterozygosity (Figures S4 and S5). To exclude technical artifacts resulting from biased allele amplification during PCR, which would thereby falsely suggest the presence of mosaicism, we generated a second independent amplicon with different PCR primers to re-sequence all nine mutations by ADS (Tables 1, S1, and S3). This analysis confirmed a statistically significant deviation in the allelic ratio for eight out of nine de novo mutations. Of note, three of these mutations had been previously reported as possible mosaic mutations.[21]

To validate these findings with an independent test, we set out to sequence the eight candidate mosaic mutations by using smMIPs for increased depth and accuracy. By sequencing germline mutations within the same assay, we first established for this technique the average and SD of the allelic ratio for true heterozygosity—this was shown to be 47.1 ± 3.3%. Unique smMIPs could be designed for all but one candidate mosaic event, located in an intron of *SETBP1* (OMIM: 611060). The remaining seven mutations were tested and confirmed to be present as mosaic events with allelic ratios between 20.8% and 29.7%. Translating these allelic ratios into percentages of cells carrying the mutation predicted that the mutations must be present in 41.6%–59.4% of the cells in blood. Thus, our results indicate that at least 7/107 (6.5%) de novo mutations detected in our cohort did not occur in the germline of the parent but instead arose post-zygotically in the offspring.

Parental Mosaicism as a Source of Seemingly De Novo Mutations

Gonadal mosaicism in a healthy parent can lead to the transmission of disease-causing mutations and recurrence of disorders with seemingly de novo ocurrence.[24] In some cases, mosaicism might not be restricted to the germ cells; it was recently shown that healthy individuals with gonadal mosaicism for disease-causing CNVs, revealed by recurrence of the disease in the offspring, carried low levels of mosaicism for this CNV in blood.[25] Following this idea, we aimed to determine whether any of the seemingly germline de novo events in our cohort of 50 probands had actually occurred as somatic mutations in one of the parents (Figure 1C). For this, we re-analyzed all 4,081 high-confidence de novo mutations previously detected by WGS in the probands and selected those de novo mutations in which two or more variant reads could be detected in the raw sequence data in one of the respective parents. Thirteen such mutations were identified, but two could not be amplified by PCR and were excluded from further analysis. We performed ADS on the remaining 11 mutations to determine whether we could detect the variant in DNA from the carrier parent. After stringent correction for the background sequencing error, four of these mutations were confirmed to be present in the blood of one of the parents. These low-level parental mosaic mutations showed an average allelic ratio of 3.54% (range 0.22%—6.15%; Tables 2 and S4). Of note, these low-level

Table 2. De Novo Mutations Originating from Parental Mosaicism

Genomic Location	Gene	OMIM Accession Number	Gene Location	Origin	Total Reads (ADS)	Variant Reads	p Value[a]
chr13:78303535A>T	*SLAIN1*	610491	intron	father	31,470	6.15%	<0.001
chr18:25210178C>T	–	–	intergenic	father	34,149	2.56%	<0.001
chr5:11327458C>T	*CTNND2*	604275	intron	mother	12,754	5.25%	<0.001
chr5:147855052G>A	*HTR4*	602164	intron	father	20,927	0.22%	<0.05

[a]p values were corrected for multiple testing by Bonferroni correction.

parental mosaic mutations, of which three were transmitted by the father and one by the mother, were not detected in the parental DNA by Sanger sequencing (Figure S6).

Modeling the Effect of Sequence Coverage on the Detection of Mosaic Mutations

Evidently, sufficient sequencing coverage is required for reliably identifying mosaic mutations. To investigate the impact of coverage on the detection of mosaic mutations, we modeled the probability of detecting both post-zygotic mutations in a proband and low-level parental mosaicism given different sequencing coverage.

The detection of post-zygotic de novo mutations requires two essential steps: calling the variant in the proband and identifying a significant deviation of the allelic distribution. Modeling under the assumption that ≥ 5 variant reads are required for variant calling and that these constitute $\geq 5\%$ of the total number of sequence reads indicates that at least 100-fold coverage is required for calling 90% of mosaic variants with an allelic ratio equal to 10% or higher (Figure S7). Increased sequencing coverage decreases the SD in the allelic ratio, which reduces technical variation (Figure S8) and allows for better discrimination between true heterozygosity and mosaicism. Provided that a post-zygotic mutation is called, we also modeled the required deviation in the allelic ratio of a mosaic variant for it to be reliably distinguished from a heterozygous variant (Figure S9). Our model indicated that at least 100-fold coverage is required for distinguishing mosaic mutations with allelic ratios $< 40\%$ from germline mutations with 95% probability.

The analysis for parental mosaicism for de novo mutations identified in a proband requires a different approach; the identification of parental mosaicism for a seemingly de novo mutation in the offspring is guided by the presence of the variant in the proband. As a consequence, the only requirement for the identification of parental mosaicism is to distinguish the variant reads in the parent from the background sequencing error at the respective genomic location. Under the assumption that two variant reads in the parent are sufficient for this, we modeled the coverage required for identifying low-level parental mosaicism (Figure S10), which showed that at least 140-fold coverage is needed for detecting low-level mosaicism of $\geq 5\%$ with $\geq 95\%$ probability.

Discussion

The aim of our study was to investigate the presence of non-germline events among de novo mutations. Our results show that 6.5% (7/107) of a subset of de novo mutations were present as mosaic mutations in the blood of the proband, strongly suggestive of a post-zygotic origin. Extrapolating our results to published genome-wide de novo mutation rates[3,21] suggests that each individual carries at least two to seven de novo mutations of post-zygotic origin. Additionally, from a group of 4,081 mutations presumed to be de novo in the offspring, we detected four mutations that were in fact inherited from one of the parents in whom the mutation was present as a low-level mosaic mutation. Although this represents only 0.1% of all high-quality de novo mutations, parental mosaicism for a seemingly de novo mutation in the offspring was observed in 4 out of 50 trios. On the basis of the stringent criteria that we used to validate variants as mosaics and our modeling data, we anticipate that our results are most likely an underestimation of the true number of mosaic mutations present in blood.

Our initial selection of potential mosaic variants was based on results obtained with relatively high-coverage (80-fold) WGS. We have shown that, for trio-based WGS, 80-fold sequencing coverage is sufficient for identifying post-zygotic events among de novo mutations. However, statistical modeling of the probability of detecting mosaicism given various sequencing depths showed that, with this coverage, there is only an 80% probability of obtaining sufficient reads for identifying mosaicism present in $\geq 10\%$ of the alleles (corresponding to $\geq 20\%$ of the cells studied; Figure S7). Similarly, with this coverage, we were only able to reliably distinguish somatic events with allelic ratios below 39% from germline mutations (Figure S9). This suggests that post-zygotic variants with allelic ratios at either extreme in the proband could have gone unidentified in our study. On the other hand, the probability of obtaining at least two sequence reads for identifying $\geq 5\%$ parental mosaicism is only 78% with 80-fold sequencing coverage, suggesting that the identification of these mutations can also be optimized by higher-sequencing coverage (Figure S10). Indeed, the low-level parental mosaic variants identified in our study had a significantly higher sequencing depth in the carrier parent than did the other de novo or post-zygotic mutations studied (Figure S11). Our results and statistical modeling highlight the

importance of high sequencing coverage in the design of trio-based WGS studies. Currently, most WGS studies are performed at 30-fold coverage.[13,26,27] If we assume that sequence quality is comparable to that of our study, this entails that fewer than 20% of mosaic variants with an allelic ratio between 10% and 33% can be identified with 30-fold sequencing coverage. Additionally, at this sequencing coverage, only mosaic mutations with an allelic ratio below 35% can be reliably distinguished from true heterozygous variants. Furthermore, our modeling suggests that there is less than a 20% probability of identifying parental mosaicism with an allelic ratio of less than 5% with WGS at 30-fold coverage. Given these results, our findings underline the need for increased sequencing coverage in WGS for the accurate identification of mosaicism.

Despite the aforementioned limitations, we have shown that WGS is a powerful method for genome-wide discovery of mosaic events. In this study, we used three additional techniques to confirm mosaicism of SNVs. After identifying de novo mutations by WGS, we first evaluated their status as post-zygotic events by ADS and Sanger sequencing. A limitation of both of these techniques is that they might show an allelic imbalance as a result of biased amplification of one allele over the other.[28] For the most part, significant deviations in the allelic ratio secondary to technical artifacts observed in Sanger sequencing and ADS were method-specific rather than reproducible PCR artifacts (Figure S12). We have attempted to remedy this problem by using smMIPs, which provide targeted high sequence coverage and the ability to identify individual captured molecules[23] and thus prevent any allelic-ratio deviations resulting from PCR amplification bias.

The presence of parental gonosomal mosaicism as the cause of a sporadic disorder in a family places the subsequent offspring at higher risk for recurrence of the disease than when the mutation is caused by a germline de novo mutation.[29] Considering this, the presence of parental mosaicism in 4 out of 50 individuals of our cohort stresses the importance of a thorough follow-up in families affected by a disorder due to a de novo mutation.[30] Notably, the lower limit of detection by Sanger sequencing has been reported to be close to only 10%,[25] whereas the highest level of parental mosaicism here detected was only 6.15% and could not be identified by Sanger sequencing (Figure S6). Because Sanger sequencing is commonly used in diagnostics, parental mosaicism below the threshold of detection of this method could account for recurrence of de novo disorders within families[24,31] and explain unsolved pedigrees with an apparently recessive inheritance of disorders otherwise known to be dominant.[32] Under these circumstances, high-coverage next-generation sequencing should be favored over Sanger sequencing for the detection of low-level parental mosaicism and might even be warranted as a standard follow-up test for each pathogenic de novo mutation. Related to this, the frequent detection of mosaic events might partially explain the occurrence of known dominant pathogenic mutations within large-scale variant databases of healthy individuals, such as the NHLBI Exome Sequencing Project Exome Variant Server. This point needs to be taken into account when these databases are used for clinical interpretation of possible pathogenic mutations. Also, previous studies have shown that certain mutations found as true heterozygous events in one tissue could be detected at low levels or be completely absent in another.[33] Clearly, further studies of mosaic mutations and their impact on phenotypic variation require an in-depth analysis of different tissues.

In summary, our results show that a proportion of de novo mutations presumed to be germline actually either occurred post-zygotically in the offspring or were inherited from low-level mosaicism in one of the parents. This indicates that de novo mutations do not arise solely during gametogenesis but also as post-zygotic mutations, suggesting that our genomes might be much more dynamic than previously considered. As the contribution of de novo mutations to human disease becomes increasingly apparent, this conclusion might very well have clinical implications. Pathogenic variants in the mosaic state require particular attention as to their detection via sequencing methods. Furthermore, their influence on the risk of recurrence of a disease underlines the importance of identifying mosaicism to offer accurate genetic counseling in sporadic disorders caused by de novo mutations.

Supplemental Data

Supplemental Data include 12 figures and 4 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.05.008.

Acknowledgments

We thank Drs. Bregje van Bon, Marjolein Willemsen, Bert de Vries, Tjitske Kleefstra, and Han Brunner from the Department of Human Genetics of the Radboud University Medical Center for the inclusion of affected individuals. We also thank Richard Leach, Robert Klein, and Rick Tearle from Complete Genomics for whole-genome sequencing. This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (916-14-043 to C.G., 916-12-095 to A.H., and SH-271-13 to C.G. and J.A.V.) and the European Research Council (ERC Starting Grant DENOVO 281964 to J.A.V.). R.A.-H. was supported by a Radboud University Medical Center grant.

Received: December 26, 2014
Accepted: May 11, 2015
Published: June 5, 2015

Web Resources

The URLs for data presented herein are as follows:

Integrated Genomics Viewer, http://www.broadinstitute.org/igv/
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
OMIM, http://www.omim.org/
UCSC Genome Browser, https://genome.ucsc.edu/

References

1. Nachman, M.W.M., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297–304.
2. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., 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*, 471–475.
3. Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al.; 1000 Genomes Project (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714.
4. Hoischen, A., van Bon, B.W.M., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., de Vries, P., de Reuver, R., Wieskamp, N., Mortier, G., et al. (2010). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. *42*, 483–485.
5. Rivière, J.-B., van Bon, B.W.M., Hoischen, A., Kholmanskikh, S.S., O'Roak, B.J., Gilissen, C., Gijsen, S., Sullivan, C.T., Christian, S.L., Abdul-Rahman, O.A., et al. (2012). De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. Nat. Genet. *44*, 440–444, S1–S2.
6. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790–793.
7. Vissers, L.E.L.M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nat. Genet. *42*, 1109–1112.
8. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. *43*, 585–589.
9. Vanneste, E., Voet, T., Le Caignec, C., Ampe, M., Konings, P., Melotte, C., Debrock, S., Amyere, M., Vikkula, M., Schuit, F., et al. (2009). Chromosome instability is common in human cleavage-stage embryos. Nat. Med. *15*, 577–583.
10. O'Huallachain, M., Karczewski, K.J., Weissman, S.M., Urban, A.E., and Snyder, M.P. (2012). Extensive genetic variation in somatic human tissues. Proc. Natl. Acad. Sci. USA *109*, 18018–18023.
11. McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R.S., Vermeesch, J.R., Hall, I.M., and Gage, F.H. (2013). Mosaic copy number variation in human neurons. Science *342*, 632–637.
12. Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M.S., Tomasini, L., Ferrandino, A.F., Rosenberg Belmaker, L.A., Szekely, A., Wilson, M., et al. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature *492*, 438–442.
13. Dal, G.M., Ergüner, B., Sağıroğlu, M.S., Yüksel, B., Onat, O.E., Alkan, C., and Özçelik, T. (2014). Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. J. Med. Genet. *51*, 455–459.
14. Huang, A.Y., Xu, X., Ye, A.Y., Wu, Q., Yan, L., Zhao, B., Yang, X., He, Y., Wang, S., Zhang, Z., et al. (2014). Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. Cell Res. *24*, 1311–1327.
15. Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478.
16. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.
17. Lindhurst, M.J., Sapp, J.C., Teer, J.K., Johnston, J.J., Finn, E.M., Peters, K., Turner, J., Cannons, J.L., Bick, D., Blakemore, L., et al. (2011). A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N. Engl. J. Med. *365*, 611–619.
18. Shirley, M.D., Tang, H., Gallione, C.J., Baugher, J.D., Frelin, L.P., Cohen, B., North, P.E., Marchuk, D.A., Comi, A.M., and Pevsner, J. (2013). Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. N. Engl. J. Med. *368*, 1971–1979.
19. Kurek, K.C., Luks, V.L., Ayturk, U.M., Alomari, A.I., Fishman, S.J., Spencer, S.A., Mulliken, J.B., Bowen, M.E., Yamamoto, G.L., Kozakewich, H.P., and Warman, M.L. (2012). Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. Am. J. Hum. Genet. *90*, 1108–1115.
20. Rohlin, A., Wernersson, J., Engwall, Y., Wiklund, L., Björk, J., and Nordling, M. (2009). Parallel sequencing used in detection of mosaic mutations: comparison with four diagnostic DNA screening techniques. Hum. Mutat. *30*, 1012–1020.
21. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.
22. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.
23. Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J., and Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. *23*, 843–854.
24. Natacci, F., Baffico, M., Cavallari, U., Bedeschi, M.F., Mura, I., Paffoni, A., Setti, P.L., Baldi, M., and Lalatta, F. (2008). Germline mosaicism in achondroplasia detected in sperm DNA of the father of three affected sibs. Am. J. Med. Genet. A. *146A*, 784–786.
25. Campbell, I.M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M.E., Nagamani, S.C.S., Erez, A., Bartnik, M., Wiśniowiecka-Kowalnik, B., et al. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am. J. Hum. Genet. *95*, 173–182.
26. Petersen, B.S., Spehlmann, M.E., Raedler, A., Stade, B., Thomsen, I., Rabionet, R., Rosenstiel, P., Schreiber, S., and Franke, A. (2014). Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. BMC Genomics *15*, 564.
27. Nemirovsky, S.I., Córdoba, M., Zaiat, J.J., Completa, S.P., Vega, P.A., González-Morón, D., Medina, N.M., Fabbro, M., Romero, S., Brun, B., et al. (2015). Whole genome sequencing reveals a de novo SHANK3 mutation in familial autism spectrum disorder. PLoS ONE *10*, e0116358.

28. Veal, C.D., Freeman, P.J., Jacobs, K., Lancaster, O., Jamain, S., Leboyer, M., Albanes, D., Vaghela, R.R., Gut, I., Chanock, S.J., and Brookes, A.J. (2012). A mechanistic basis for amplification differences between samples and between genome regions. BMC Genomics *13*, 455.
29. Campbell, I.M., Stewart, J.R., James, R.A., Lupski, J.R., Stankiewicz, P., Olofsson, P., and Shaw, C.A. (2014). Parent of origin, mosaicism, and recurrence risk: probabilistic modeling explains the broken symmetry of transmission genetics. Am. J. Hum. Genet. *95*, 345–359.
30. Faivre, L., Williamson, K.A., Faber, V., Laurent, N., Grimaldi, M., Thauvin-Robinet, C., Durand, C., Mugneret, F., Gouyon, J.-B., Bron, A., et al. (2006). Recurrence of SOX2 anophthalmia syndrome with gonosomal mosaicism in a phenotypically normal mother. Am. J. Med. Genet. A. *140*, 636–639.
31. Elalaoui, S.C., Kraoua, L., Liger, C., Ratbi, I., Cavé, H., and Sefiani, A. (2010). Germinal mosaicism in Noonan syndrome: A family with two affected siblings of normal parents. Am. J. Med. Genet. A. *152A*, 2850–2853.
32. Schinzel, A., and Giedion, A. (1978). A syndrome of severe midface retraction, multiple skull anomalies, clubfeet, and cardiac and renal malformations in sibs. Am. J. Med. Genet. *1*, 361–375.
33. Huisman, S.A., Redeker, E.J., Maas, S.M., Mannens, M.M., and Hennekam, R.C. (2013). High rate of mosaicism in individuals with Cornelia de Lange syndrome. J. Med. Genet. *50*, 339–344.

The American Journal of Human Genetics

Supplemental Data

Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation

Rocio Acuna-Hidalgo, Tan Bo, Michael P. Kwint, Maartje van de Vorst, Michele Pinelli, Joris A. Veltman, Alexander Hoischen, Lisenka E.L.M. Vissers, and Christian Gilissen

Figure S1

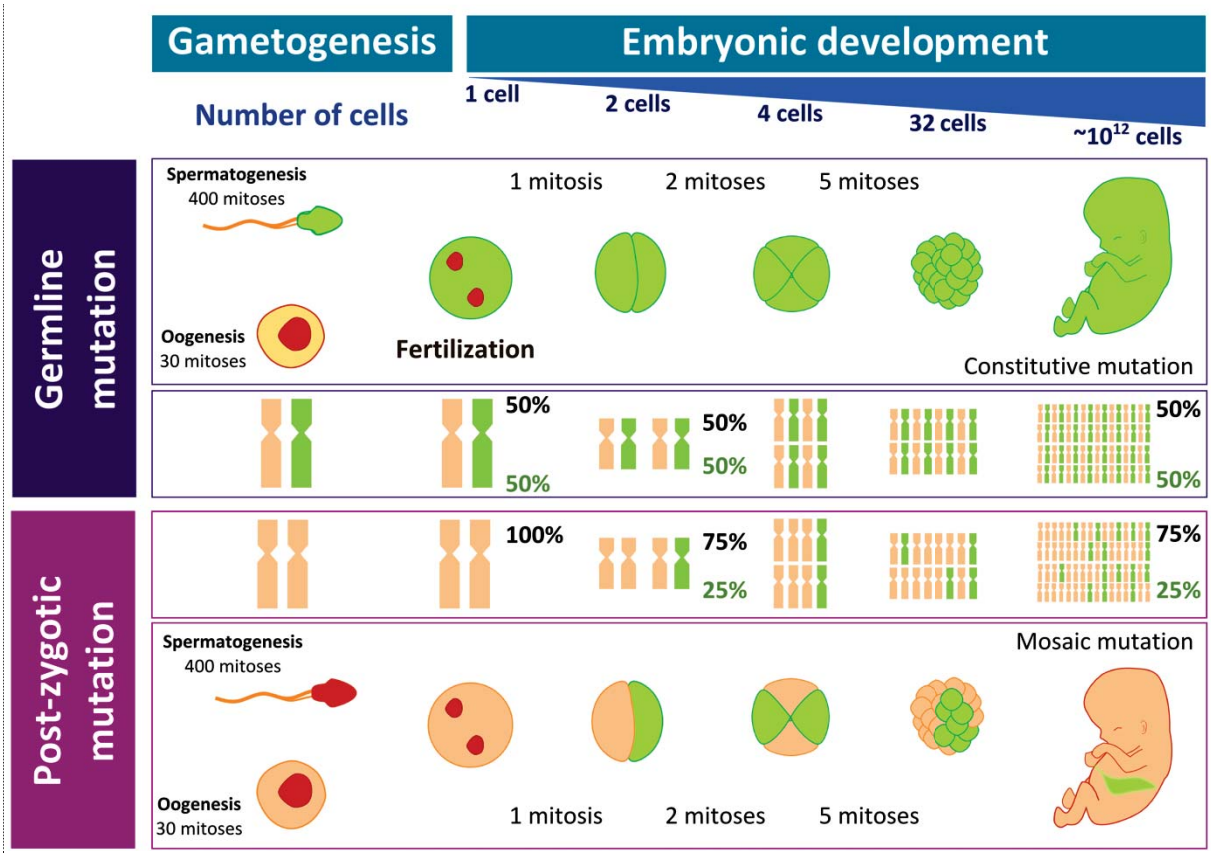

Figure S1. Overview of de novo mutations in germline and post-zygotic status. Germline *de novo* mutations are present in all cells and are therefore truly heterozygous, with an equal distribution of the wild type and the mutated allele (50:50, see top panel "germline mutation"). Somatic *de novo* mutations are not present in all cells of an organism; some cells will carry the mutation while others will not, leading to an unbalanced distribution of the mutated allele (*e.g.* 80:20, see bottom panel "post-zygotic mutation"). This unbalanced distribution of the mutated alleles over wild type alleles can be detected by sequencing as a deviation in the allele ratio (*i.e.* the signal corresponding to the mutant allele versus the signal corresponding to the reference allele). Using next generation sequencing (NGS) techniques, this is observed as lower number of reads carrying the mutated allele versus reads carrying the reference allele. For Sanger sequencing, this unbalance is detected as a difference in the intensity of the bases in the chromatogram. However, there may also be an unbalance in the distribution of the alleles as a result of technical variation. To identify mutations present in mosaic state, it is necessary to differentiate the deviation in the variant ratio that is a result of technical variation from the deviation in the variant ratio that is a consequence of a biological phenomenon. Cells and chromosomes carrying a mutation are shown in green in the figure.

Figure S2

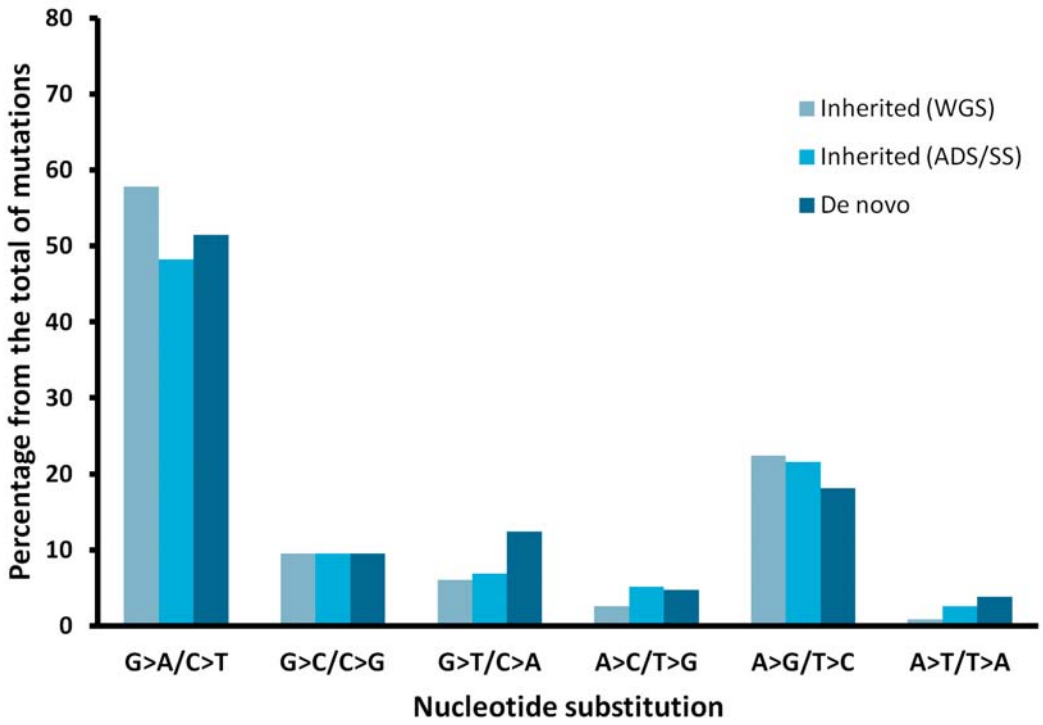

Figure S2. Frequency of nucleotide substitutions per class of variants. The frequency for all nucleotide substitutions was determined for each of the class of variants tested, including 115 inherited variants studied by WGS, 109 inherited variants studied by ADS and Sanger sequencing (SS) and 107 de novo SNV mutations. Nucleotide substitution frequencies were not significantly different between the variant classes (Chi square test, df = 6, X^2=6.113, p = 0.41).

Figure S3

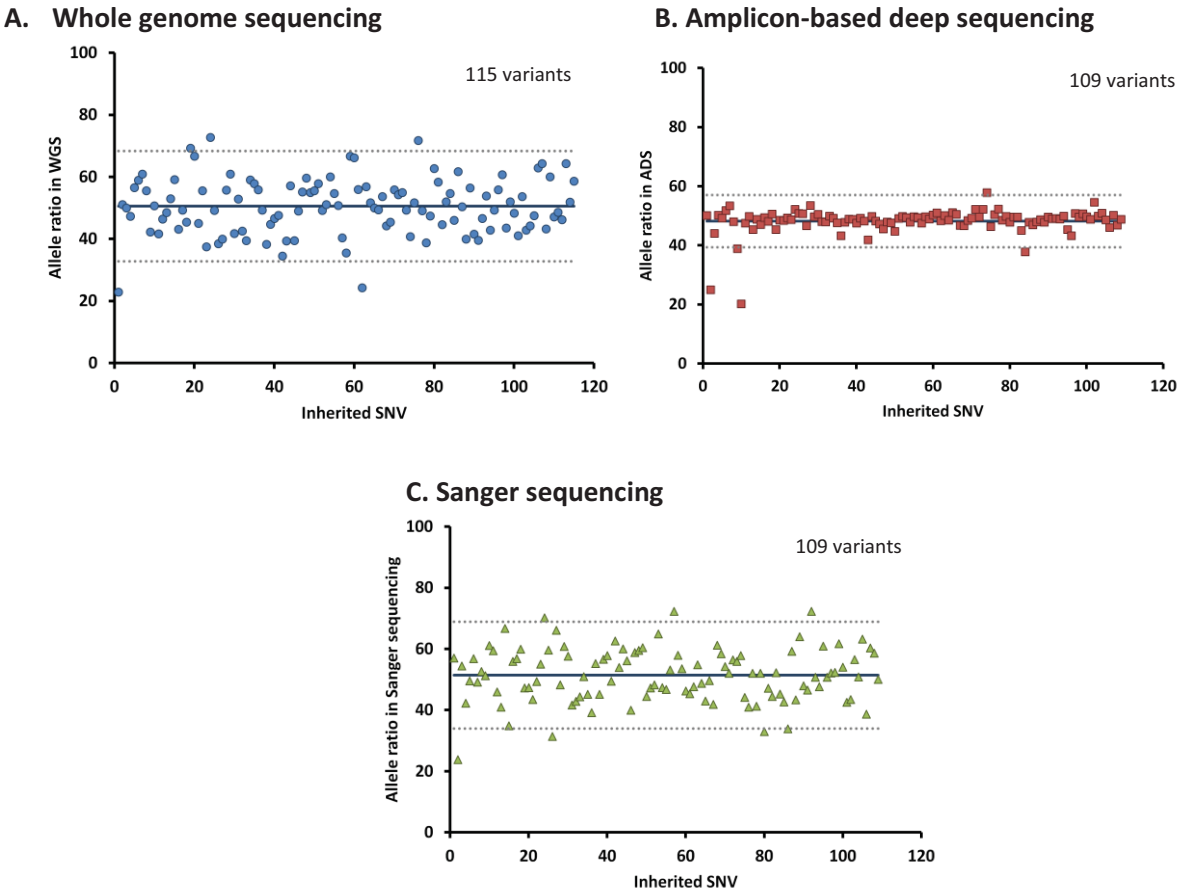

Figure S3. Allele ratio of inherited heterozygous variants sequenced using different sequencing techniques. Results include whole genome sequencing (panel A), amplicon-based deep sequencing (panel B) and Sanger sequencing (panel C). Allele ratios were calculated as the percentage of variant reads from the total of reads for NGS technologies, and as the intensity of the mutated base in the chromatogram over the sum of the intensities of the reference and the mutated bases for Sanger sequencing. The mean allele ratio is indicated by the black line; ±2 standard deviations are indicated by the dotted gray lines (see table S1 for the raw data).

Figure S4

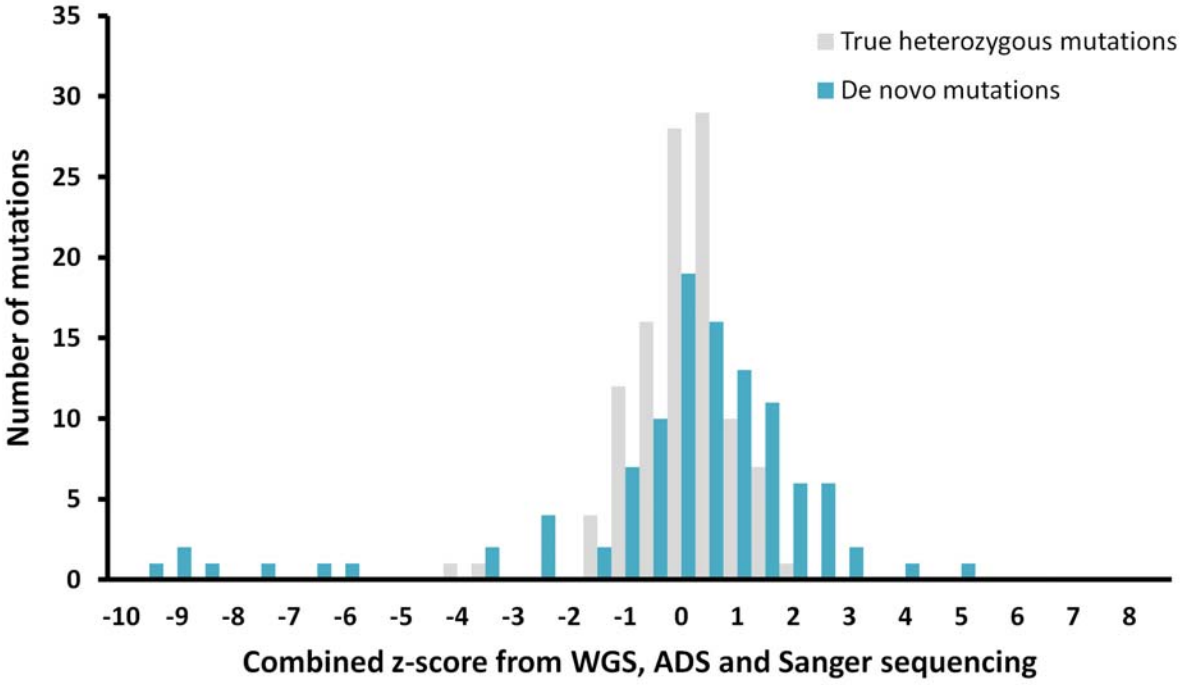

Figure S4. Distribution of de novo versus true heterozygous variants. Inherited heterozygous variants (n=109) are visualized in grey whereas de novo mutations (n=107) are reprented in blue, with combined z-scores for all sequencing techniques on the x-axis and the number of mutations on the y-axis. Putative mosaic de novo mutations are located at the left of the graph. These seven variants consistently show a lower allele ratio in different sequencing techniques as well as when sequencing was performed using independent primer pairs for amplification.

Figure S5

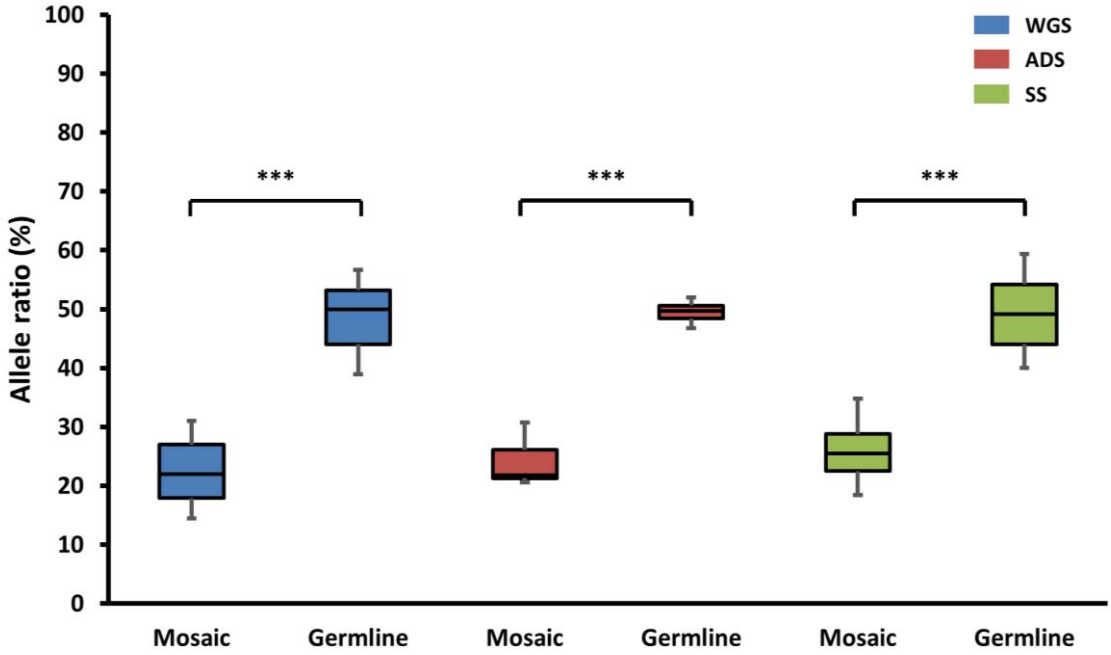

Figure S5. Distribution of the allele ratio per sequencing technique of 7 *de novo* mutations identified as post-zygotic events compared to 100 germline *de novo* mutations in the proband. Per class, the median, 10th, 25th, 75th and 90th percentile are depicted. The difference between the variant ratio of post-zygotic versus germline mutations was statistically significant for all methods (Student's T-test, *** p<0.001).

Figure S6

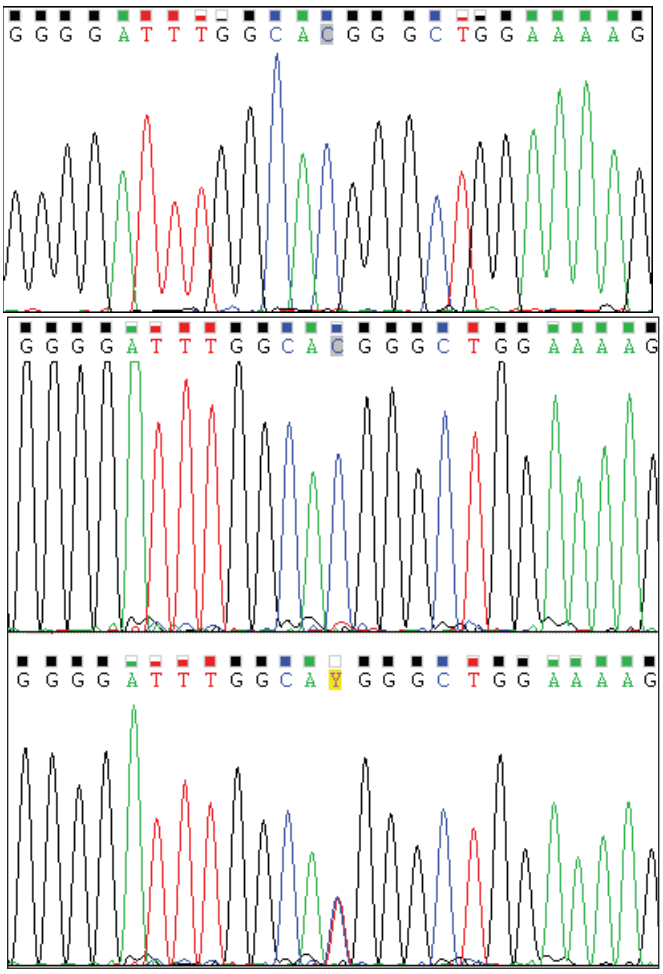

Figure S6. Sanger sequencing traces of de novo mutation chr5:11327457 C>T in *CTNDD2* in the non-carrier father (top), carrier mother (middle) and proband (bottom). This mutation was originally identified by trio-based WGS and later confirmed in DNA derived from maternal blood by ADS, with a variant ratio of 5.25%. Note that the mutated allele in the maternal DNA sample is indistinguishable from the background noise.

Figure S7

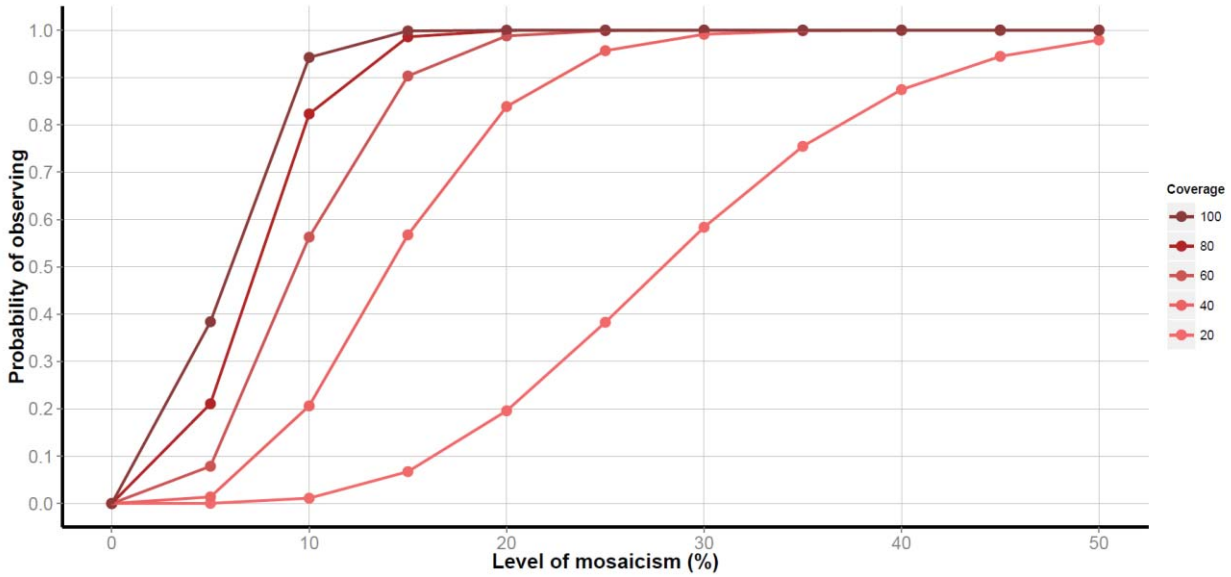

Figure S7. Modeling of the probability of identifying variants with different levels of mosaicism given different sequencing coverage. We assumed that automated variant calling algorithms require a variant to be present in at least 5 sequencing reads which constitute at least 5% of the total number of reads at a position. A binomial distribution was used to calculate the probability (*i.e.* power) of reaching both these requirements for different depths of coverage and various levels of mosaicism. The X-axis indicates the level of mosaicism (*i.e.* the true allelic ratio). The Y-axis shows the probability of identifying this mosaicism. Each line represents the results for different sequencing coverage according to the legend.

Figure S8

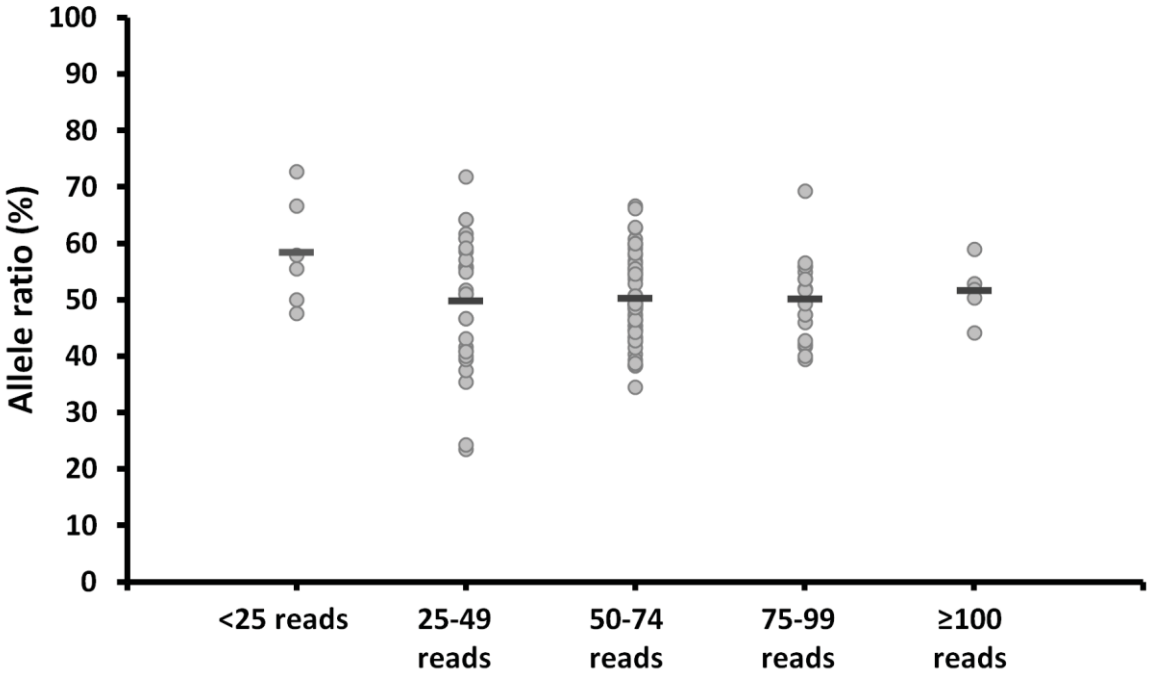

Figure S8. Allele ratio by sequencing coverage for 115 inherited heterozygous variants in WGS data. Inherited variants were classified according to sequencing depth at the respective base pair positions (in bins, increasing by 25 sequence reads each) and the allele ratio at which the variant allele was observed. Each mutation is represented by a circle, with the horizontal bar representing the average allele ratio per bin. In more detail, the classes comprise 6 (<25 reads), 28 (25-49 reads), 56 (50-74 reads), 19 (75 to 99 reads) and 5 variants (≥100 reads), respectively. Whereas the average allele ratio does not significantly differ with increasing sequence depth (49.8% with 25-49 fold coverage versus 51.7% with ≥100-fold coverage), higher sequence coverage does decrease the standard deviation (11.9 with 25-49 fold coverage versus 5.27 with coverage ≥100-fold). These data indicate that higher sequencing coverage decreases the technical variation and offers higher sensitivity for the detection of biologically relevant deviations in the variant ratio.

Figure S9

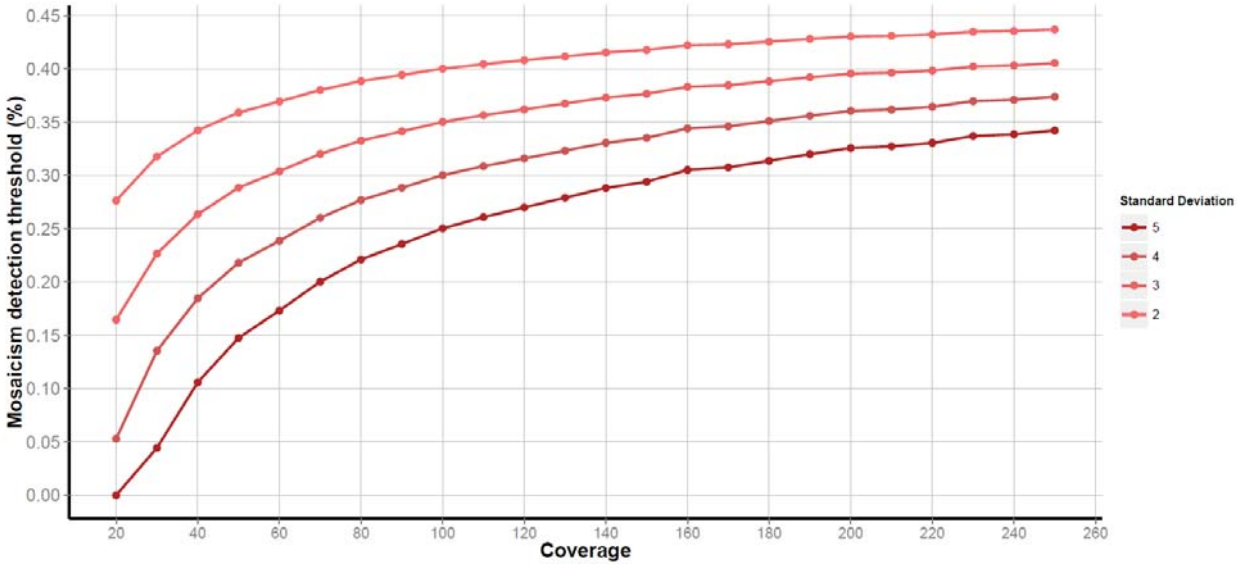

Figure S9. Simulation of the level of mosaicism which can be statistically distinguished from a heterozygous variants given different sequencing coverage and thresholds for significance. We simulated reads for heterozygous variants (n=10,000) at different sequencing coverage based on a binomial distribution. From this, we calculated the standard deviation of this distribution and, for different significance thresholds, we assessed the level of mosaicism for which the allelic ratio can reliably be distinguished from the allelic ratio of a heterozygous variant. The X-axis indicates the sequencing coverage, while the Y-axis indicates the level of mosaicism that can be distinguished from a heterozygous variant. Each line represents the significance threshold as the distance from the average in standard deviations. Note that 10% of the allelic ratio means that 20% of the cells carry the variant.

Figure S10

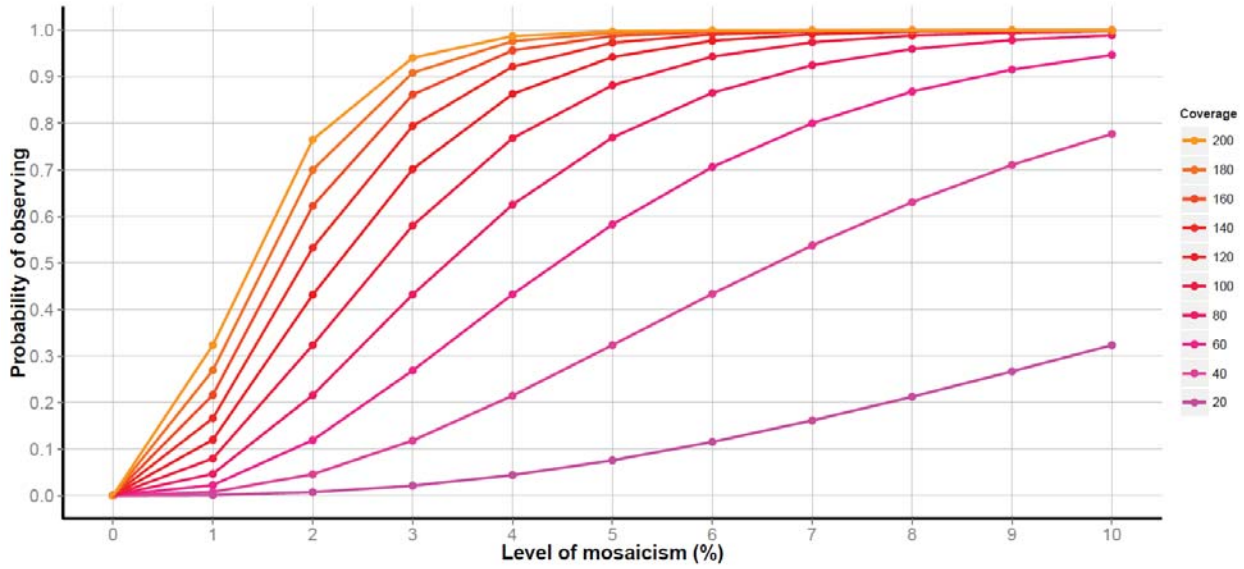

Figure S10. **Modeling of the probability of identifying different levels of mosaicism in at least two reads for different sequencing depths**. In this scenario, the position of interest is already identified, as the offspring will have a de novo mutation at this base pair. We considered 2 reads showing the mutated allele to be sufficient to distinguish the variant from background sequencing error. We applied a binomial model for different sequencing depths and levels of mosaicism to calculate the probability of obtaining 2 sequencing reads with the variant. The X-axis indicates the percentage of mosaicism as the allelic ratio, while the Y-axis indicates the probability of identifying at least 2 reads. Each line shows the result for different depths of coverage.

Figure S11

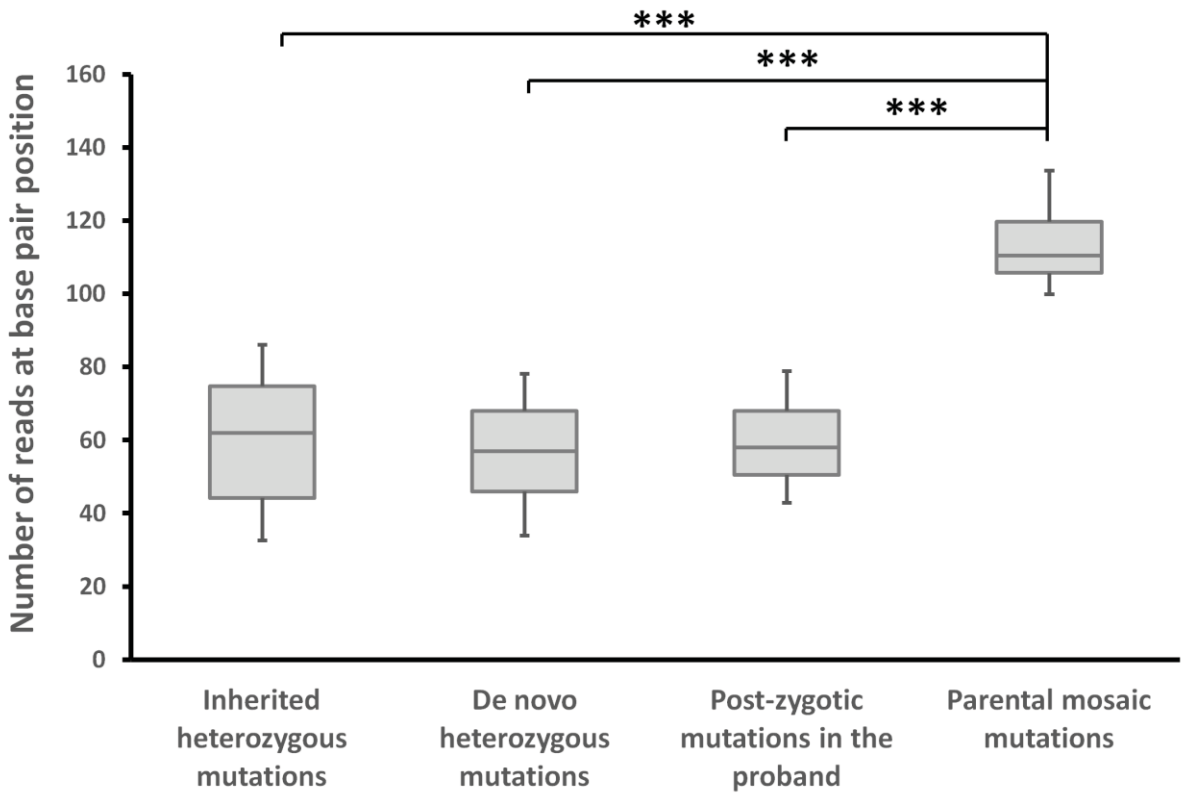

Figure S11. Sequencing coverage in WGS data per mutation category. Sequencing depth in WGS data for the evaluated mutations are presented per category, including 115 inherited heterozygous mutations (WGS from the proband), 100 de novo heterozygous mutations (WGS from the proband), 7 post-zygotic mutations in the proband (WGS from the proband) and 4 parental mosaic mutations (WGS from the parent). The median, the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile for each group are plotted, with the asterisks denoting a difference in sequencing coverage between inherited heterozygous, germline *de novo* variants and post-zygotic mutations in the proband and parental mosaic mutations (*** $p < 0.001$, Student's t-test). These data suggest that the sequencing coverage required for the detection of de novo mutations is lower than the sequencing depth necessary for the detection of low-level parental mosaicism.

Figure S12

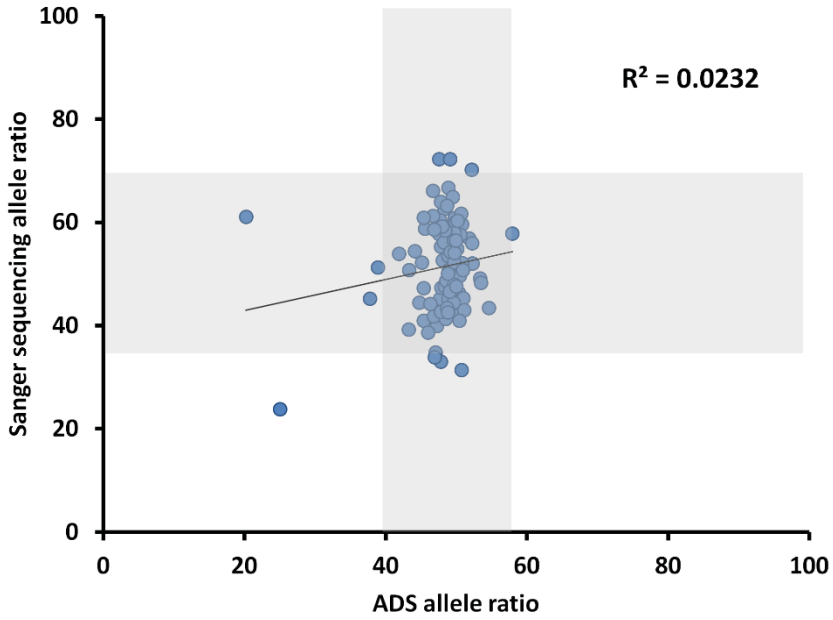

Figure S12. Comparison of the allele ratio obtained for different sequencing techniques in truly heterozygous variants. A group of 109 inherited variants was amplified using the same primer pair and sequenced both by ADS and Sanger sequencing. Each circle represents one variant, while the gray rectangles highlight the 95% confidence interval for each sequencing method. While there are several variants outside the 95% confidence interval for each method, only one SNV shows a statistically significant deviation in the allele ratio both in ADS and Sanger sequencing. Deviation in both sequencing methods may be secondary to biased allele amplification, while deviations observed in a single technique, but not reproducible in another may be caused by technical error specific to each sequencing method.

Table S2

	Whole Genome Sequencing (n = 115)	Amplicon-based deep sequencing (n = 109)	Sanger sequencing (n = 109)	smMIPs (n = 7)
	Allele ratio %	Allele ratio %	Allele ratio %	Allele ratio %
Average	50.5	48.2	51.4	48.1
Standard deviation	8.9	4.4	8.7	3.1
95% interval	32.8-68.3	39.3-57.0	33.9-68.8	41.9-54.3
Maximum observed	72.7	57.9	72	50.8
Minimum observed	22.9	20.2	24	41.8

Table S2. Technical specifications for each sequencing technique.

Table S3

		WGS		Amplicon-based deep sequencing		Sanger sequencing		Amplicon-based deep sequencing (2)		Statistical analysis			Single molecule MIPs	
Gene name	**Genomic location (hg19)**	**mutant %**	**z-score**	**mutant %**	**z-score**	**mutant %**	**z-score**	**mutant %**	**z-score**	**Combined z-score**	**p-value (BH)**	**Average mutant %**	**mutant %**	**z-score**
KANSL2	chr12:49072911C>A	21	-3.35	20	-6.32	19	-3.70	19	-6.55	-9.96	6.94E-21	20.8	24.7	-6.85
CREBL2	chr12:12788868G>C	14	-4.14	20	-6.32	31	-2.36	21	-6.09	-9.46	6.40E-19	21.0	19.4	-8.51
PNKP	chr19:50367525C>T	23	-3.13	22	-5.87	25	-2.98	23	-5.64	-8.81	7.05E-17	22.7	20.2	-8.25
PIAS1	chr15:68468014T>A	22	-3.24	22	-5.87	25	-3.08	20	-6.32	-9.25	1.84E-18	22.9	25.9	-6.50
HIVEP2	chr6:143092683C>T	31	-2.22	22	-5.87	31	-2.37	23	-5.64	-8.05	2.20E-14	25.2	19.5	-8.43
NEK1	chr4:170359295T>G	15	-4.06	32	-3.61	40	-1.26	34	-3.16	-6.05	3.67E-08	29.4	25.7	-6.79
DPYD	chr1:97588236C>T	31	-2.22	30	-4.07	27	-2.85	29	-4.29	-6.71	3.17E-10	29.7	31.9	-5.25

Table S3. Z-scores and statistical evaluation per sequencing technique. P-values are corrected for multiple testing using Benjamini-Hochberg correction.