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

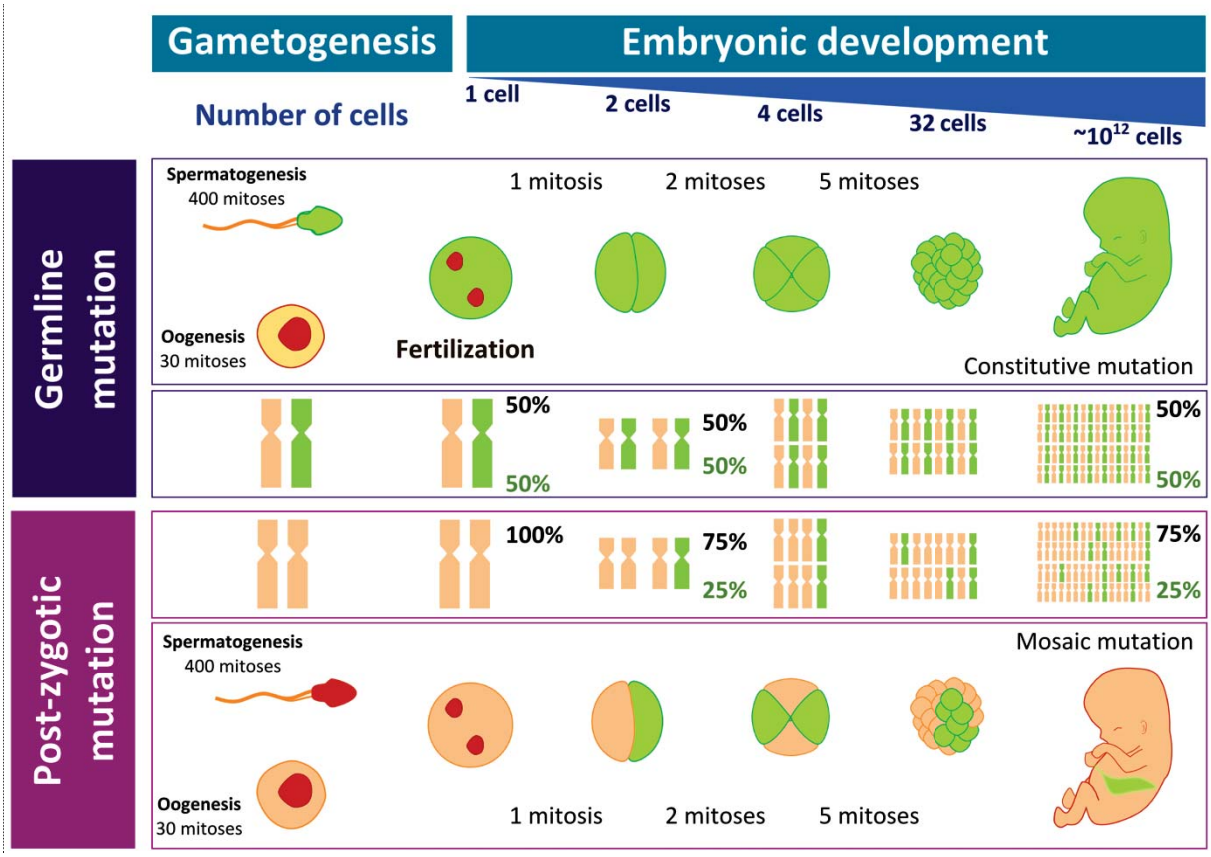

Figure S1. Overview of de novo mutations in germline and post-zygotic status. Germline *de novo* mutations are present in all cells and are therefore truly heterozygous, with an equal distribution of the wild type and the mutated allele (50:50, see top panel "germline mutation"). Somatic *de novo* mutations 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

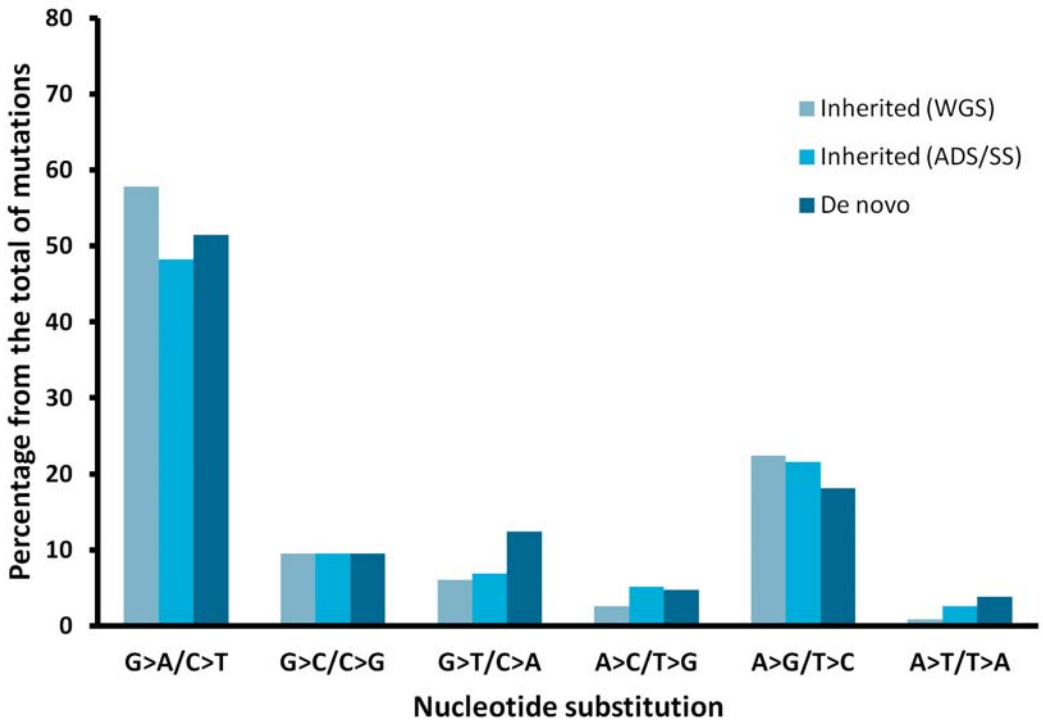

Figure S2. Frequency of nucleotide substitutions per class of variants. The frequency for all nucleotide substitutions was determined for each of the class of variants tested, including 115 inherited variants studied by WGS, 109 inherited variants studied by ADS and Sanger sequencing (SS) and 107 de 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

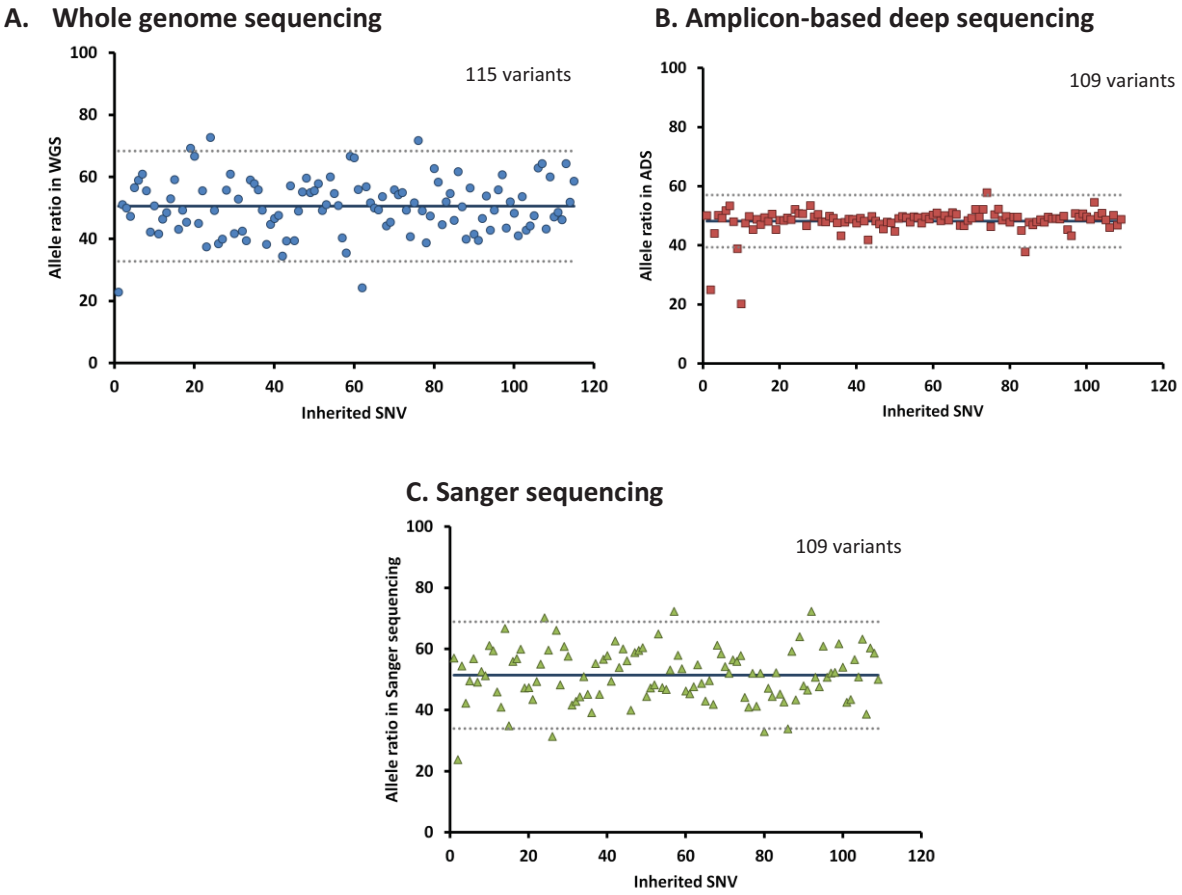

Figure S3. Allele ratio of inherited heterozygous variants sequenced using different sequencing techniques. Results include whole genome sequencing (panel A), amplicon-based deep sequencing (panel B) and Sanger sequencing (panel C). Allele ratios were calculated as the percentage of variant reads from 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

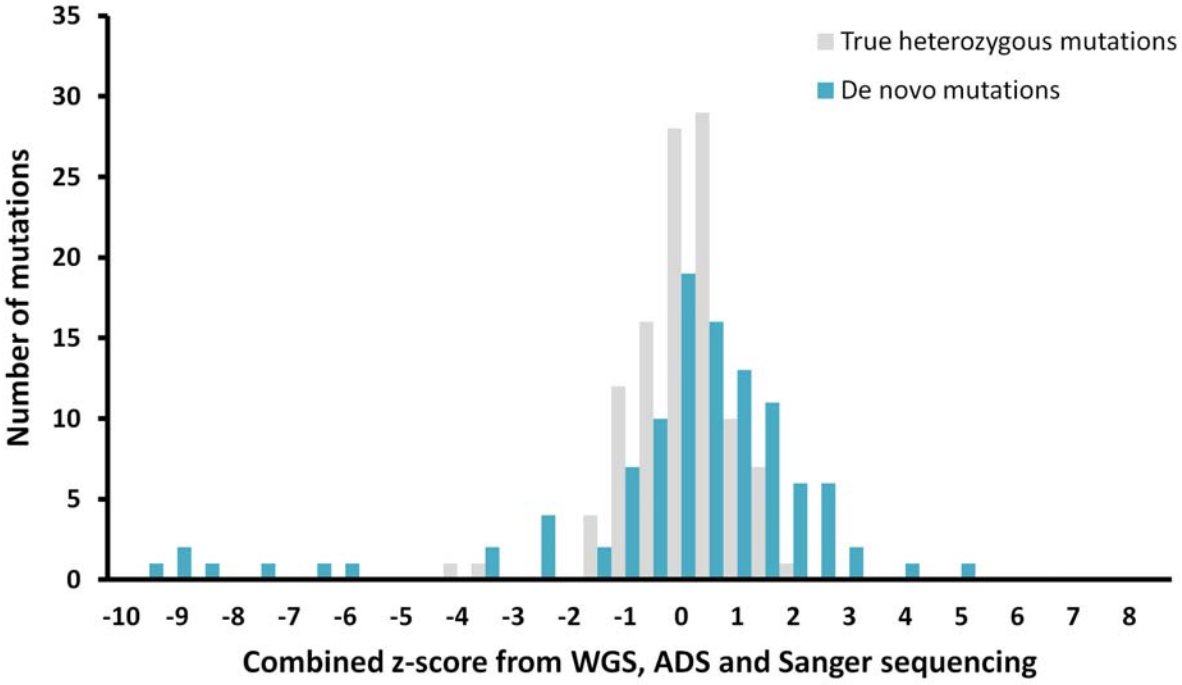

Figure S4. Distribution of de novo versus true heterozygous variants. Inherited heterozygous variants (n=109) are visualized in grey whereas de novo mutations (n=107) are reprented in blue, with combined z-scores for all sequencing techniques on the x-axis and the number of mutations on the y-axis. Putative 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

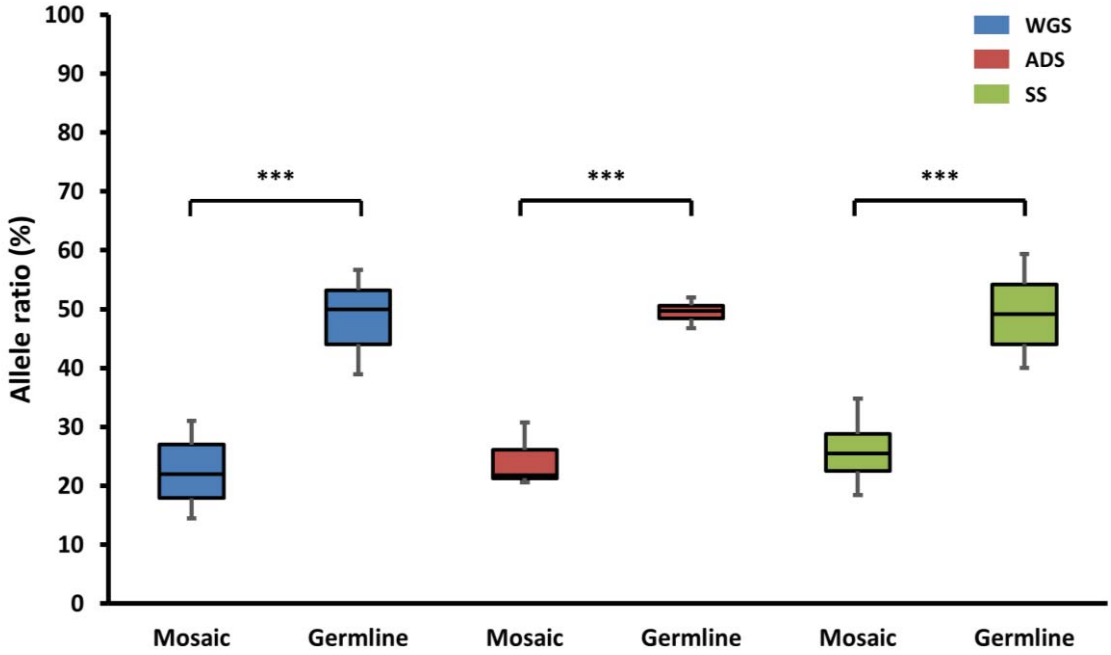

Figure S5. Distribution of the allele ratio per sequencing technique of 7 *de novo* mutations identified as post-zygotic events compared to 100 germline *de novo* mutations in the proband. Per class, the median, 10th, 25th, 75th and 90th percentile are depicted. The difference between the variant ratio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

Figure S6

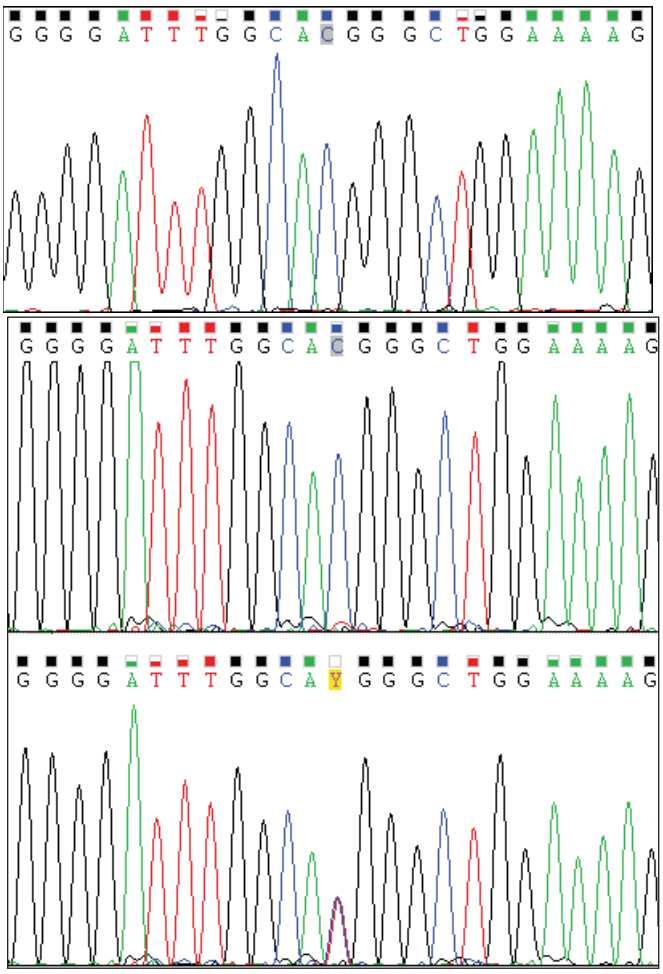

Figure S6. Sanger sequencing traces of de novo mutation chr5:11327457 C>T in *CTNDD2* in the non-carrier father (top), carrier mother (middle) and proband (bottom). This mutation was originally identified by trio-based WGS and later confirmed in DNA derived from maternal blood by ADS, with a variant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

Figure S7

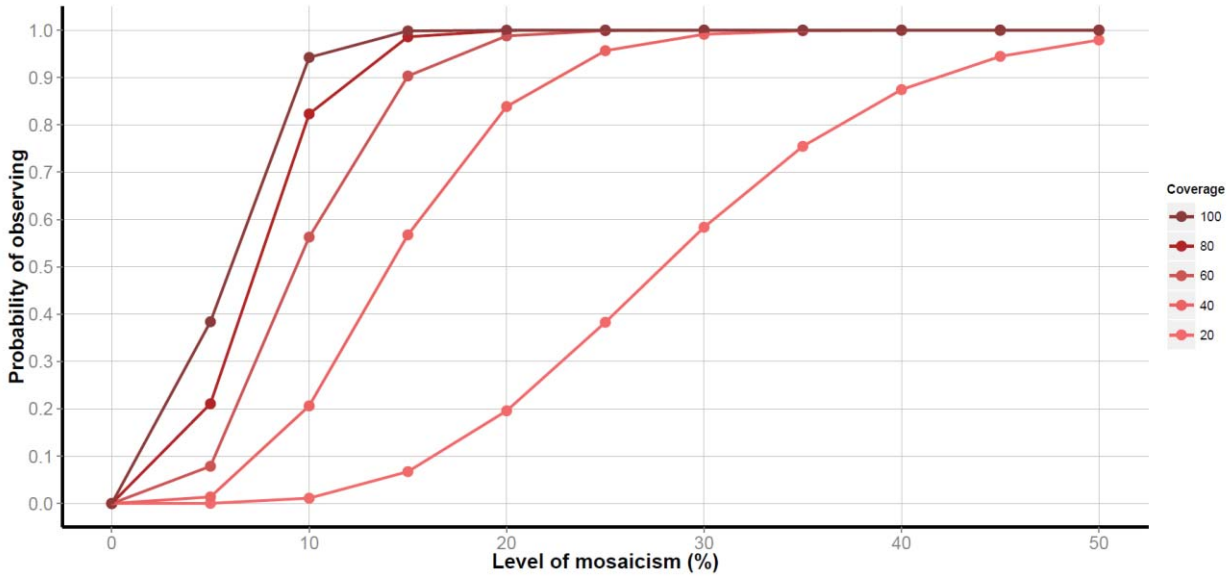

Figure S7. Modeling of the probability of identifying variants with different levels of mosaicism given different sequencing coverage. We assumed that automated variant calling algorithms require a variant to be present in at least 5 sequencing reads which constitute at least 5% of the total number of 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

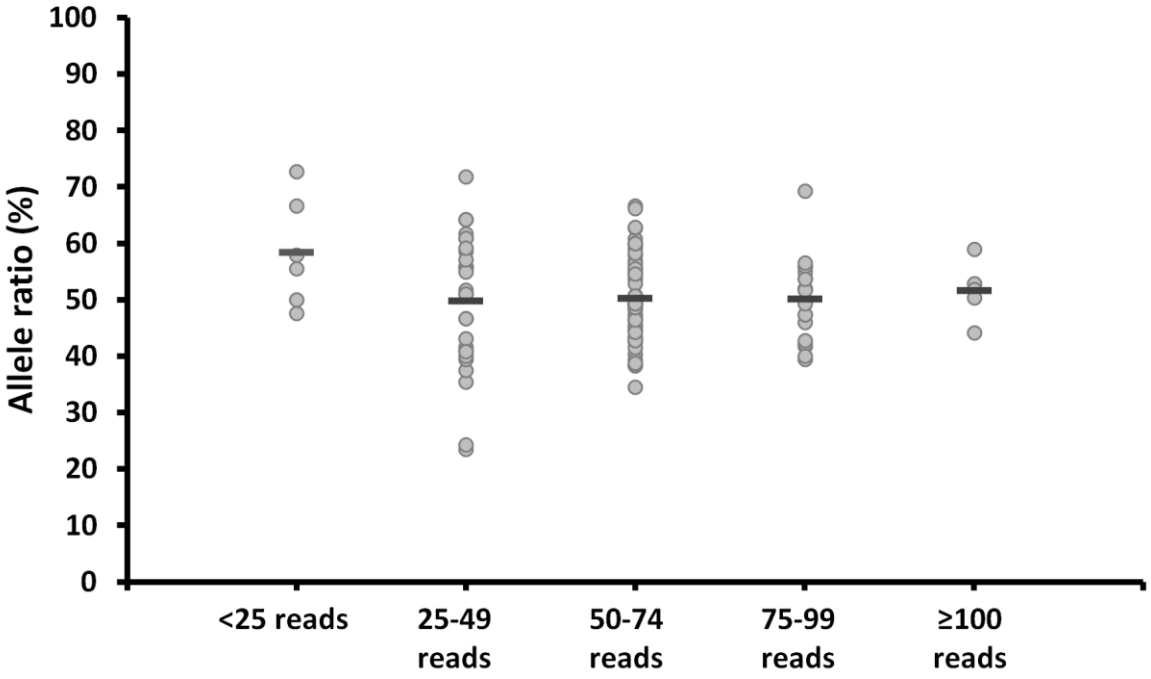

Figure S8. Allele ratio by sequencing coverage for 115 inherited heterozygous variants in WGS data. Inherited variants were classified according to sequencing depth at the respective base pair positions (in bins, increasing by 25 sequence reads each) and the allele ratio at which the variant allele was 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

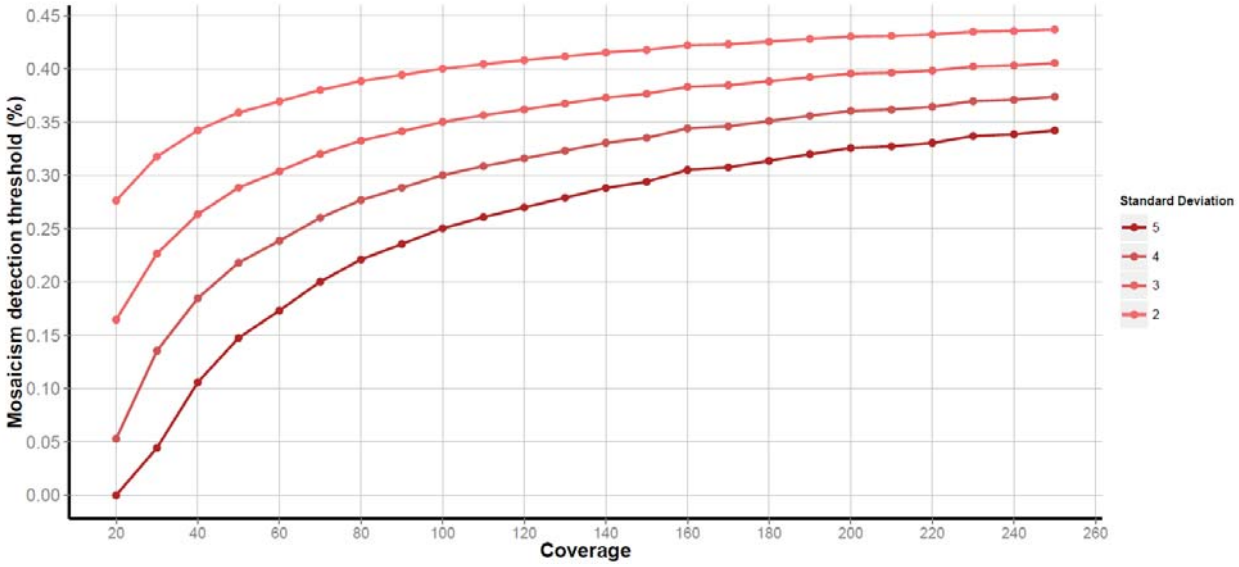

Figure S9. Simulation of the level of mosaicism which can be statistically distinguished from a heterozygous variants given different sequencing coverage and thresholds for significance. We simulated reads for heterozygous variants (n=10,000) at different sequencing coverage based on a binomial distribution. 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

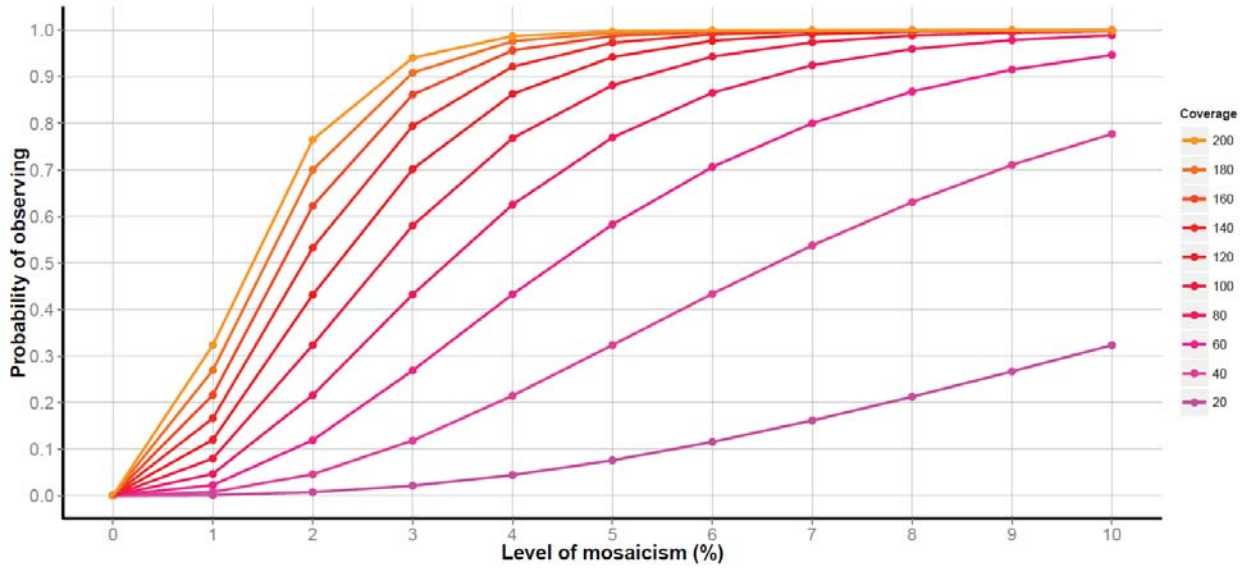

Figure S10. **Modeling of the probability of identifying different levels of mosaicism in at least two reads for different sequencing depths**. In this scenario, the position of interest is already identified, as the offspring will have a de novo mutation at this base pair. We considered 2 reads showing 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

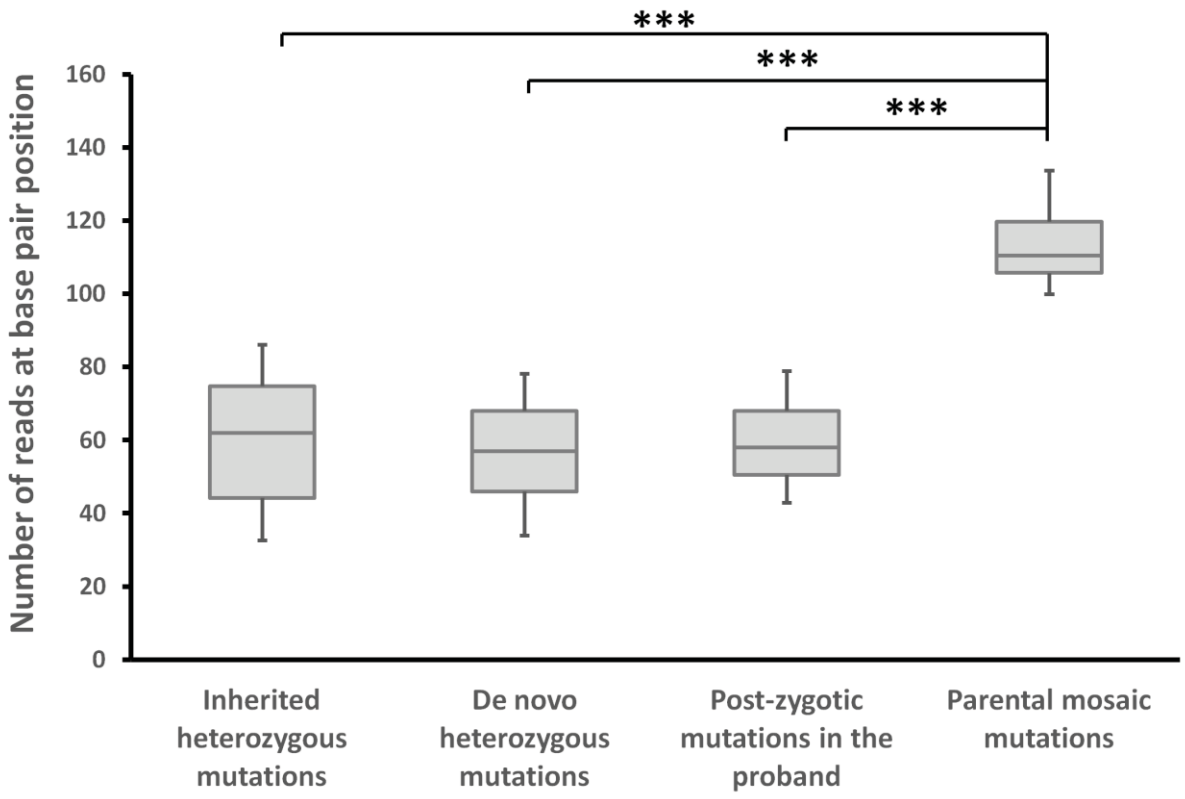

Figure S11. Sequencing coverage in WGS data per mutation category. Sequencing depth in WGS data for the evaluated mutations are presented per category, including 115 inherited heterozygous mutations (WGS from the proband), 100 de novo heterozygous mutations (WGS from the proband), 7 post-zygotic mutations 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

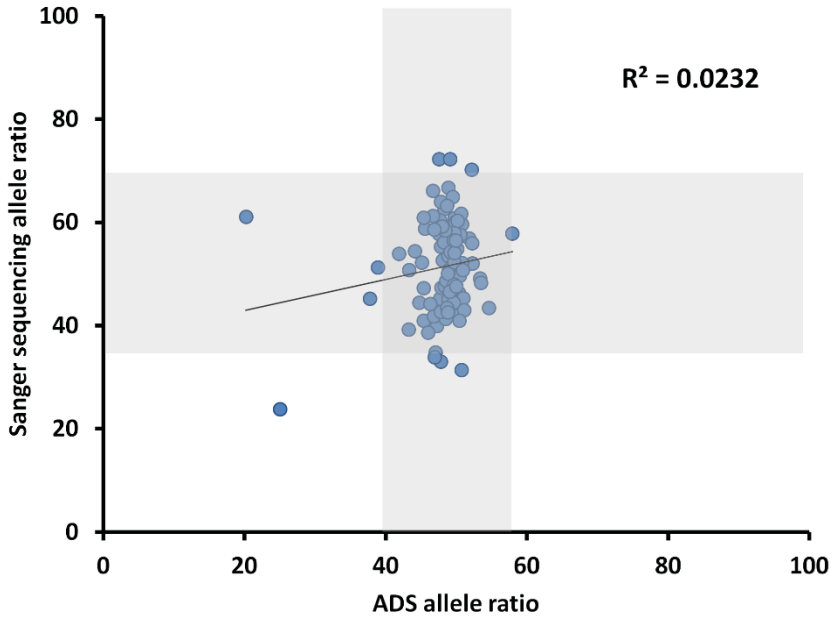

Figure S12. Comparison of the allele ratio obtained for different sequencing techniques in truly heterozygous variants. A group of 109 inherited variants was amplified using the same primer pair and sequenced both by ADS and Sanger sequencing. Each circle represents one variant, while the gray rectangles 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.