Supplementary Methods

Whole exome sequencing and data processing were performed by the Genomics Platform at the Broad Institute of MIT and Harvard. Libraries from DNA samples (>250 ng of DNA, at >2 ng/ul) were created with an Illumina Nextera or Twist exome capture (~38 Mb target) and sequenced (150 bp paired reads) to cover >85% of targets at 20x and a mean target coverage of >55x. Sample identity quality assurance checks were performed on each sample. The exome sequencing data were de-multiplexed and each sample’s sequence data were aggregated into a single Picard BAM file. Exome sequencing data were processed through a pipeline based on Picard, using base quality score recalibration and local realignment at known indels. The BWA aligner was used for mapping reads to the human genome build 38. Single nucleotide variants (SNVs) and insertions/deletions (indels) were jointly called across all samples using Genome Analysis Toolkit (GATK) HaplotypeCaller package version 3.5. Default filters were applied to SNV and indel calls using the GATK Variant Quality Score Recalibration (VQSR) approach. Annotation was performed using Variant Effect Predictor (VEP). Lastly, the variant call set was uploaded to seqr for collaborative analysis between the CMG and investigator.

Whole genome sequencing and data processing were performed by the Genomics Platform at the Broad Institute. PCR-free preparation of sample DNA (350 ng input at >2 ng/ul) is accomplished using Illumina HiSeq X Ten v2 chemistry. Libraries are sequenced to a mean target coverage of 30x. Genome sequencing data were processed through a pipeline based on Picard, using base quality score recalibration and local realignment at known indels. The BWA aligner was used for mapping reads to the human genome build 38. Single Nucleotide Variants (SNVs) and insertions/deletions (indels) were jointly called across all samples using the Genome Analysis Toolkit (GATK) HaplotypeCaller package version 4.0. Default filters
were applied to SNV and indel calls using the GATK Variant Quality Score Recalibration (VQSR) approach. Annotation was performed using Variant Effect Predictor (VEP). Lastly, the variant call set was uploaded to seqr for collaborative analysis between the CMG and investigator.

Human whole transcriptome sequencing was performed by the Genomics Platform at the Broad Institute of MIT and Harvard. The transcriptome product combines poly(A)-selection of mRNA transcripts with a strand-specific cDNA library preparation, with a mean insert size of 550bp. Libraries were sequenced on the HiSeq 2500 platform to a minimum depth of 50-75 million STAR-aligned reads. ERCC RNA controls are included for all samples, allowing additional control of variability between samples.

Copy-number variants (CNVs) were identified from whole-exome sequencing following GATK-gCNV best practices, as follows: read coverage was first calculated for each exome using GATK CollectReadCounts. After coverage collection, all samples were subdivided into batches for gCNV model training and execution; these batches were determined based on a principal components analysis (PCA) of sequencing read counts. After batching, one gCNV model was trained per batch using GATK GermlineCNVCaller on a subset of training samples, and the trained model was then applied to call CNVs for each sample per batch. Finally, all raw CNVs were aggregated across all batches and post-processed using quality- and frequency-based filtering to produce a final CNV callset.

The Infinium Omni5-4-kit microarray was performed using the Infinium LCG Quad assay protocol supplied by the manufacturer (Illumina, Inc., San Diego, CA). The array was scanned for fluorescence signal intensity and measured using the iScan System (Illumina,
Inc., San Diego, CA). Raw data were processed using Illumina GenomeStudio v2.0 software, analysed using Illumina KaryoStudio v1.4.3 software and curated for copy number and loss of heterozygosity (LOH). Interpretation was based on the UCSC GRCh37/hg19 human reference sequence.