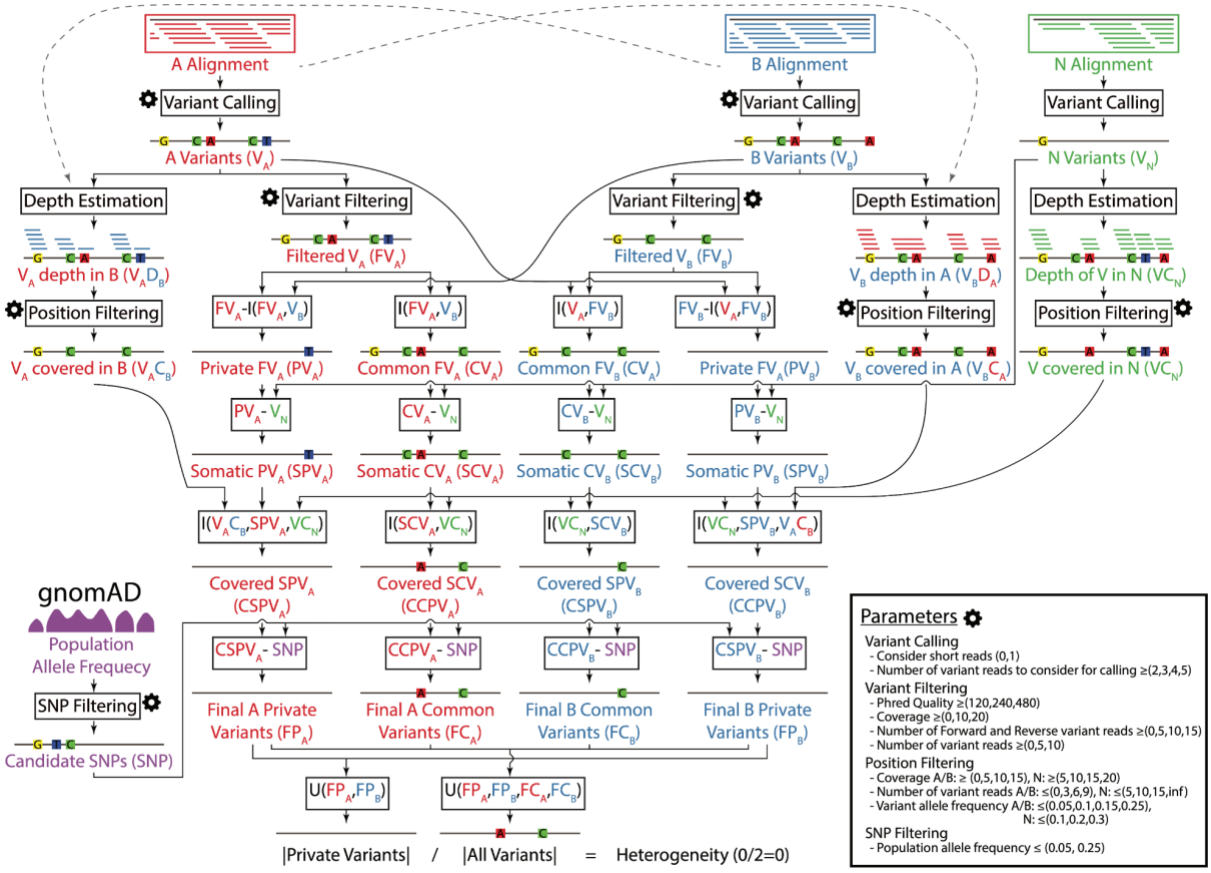
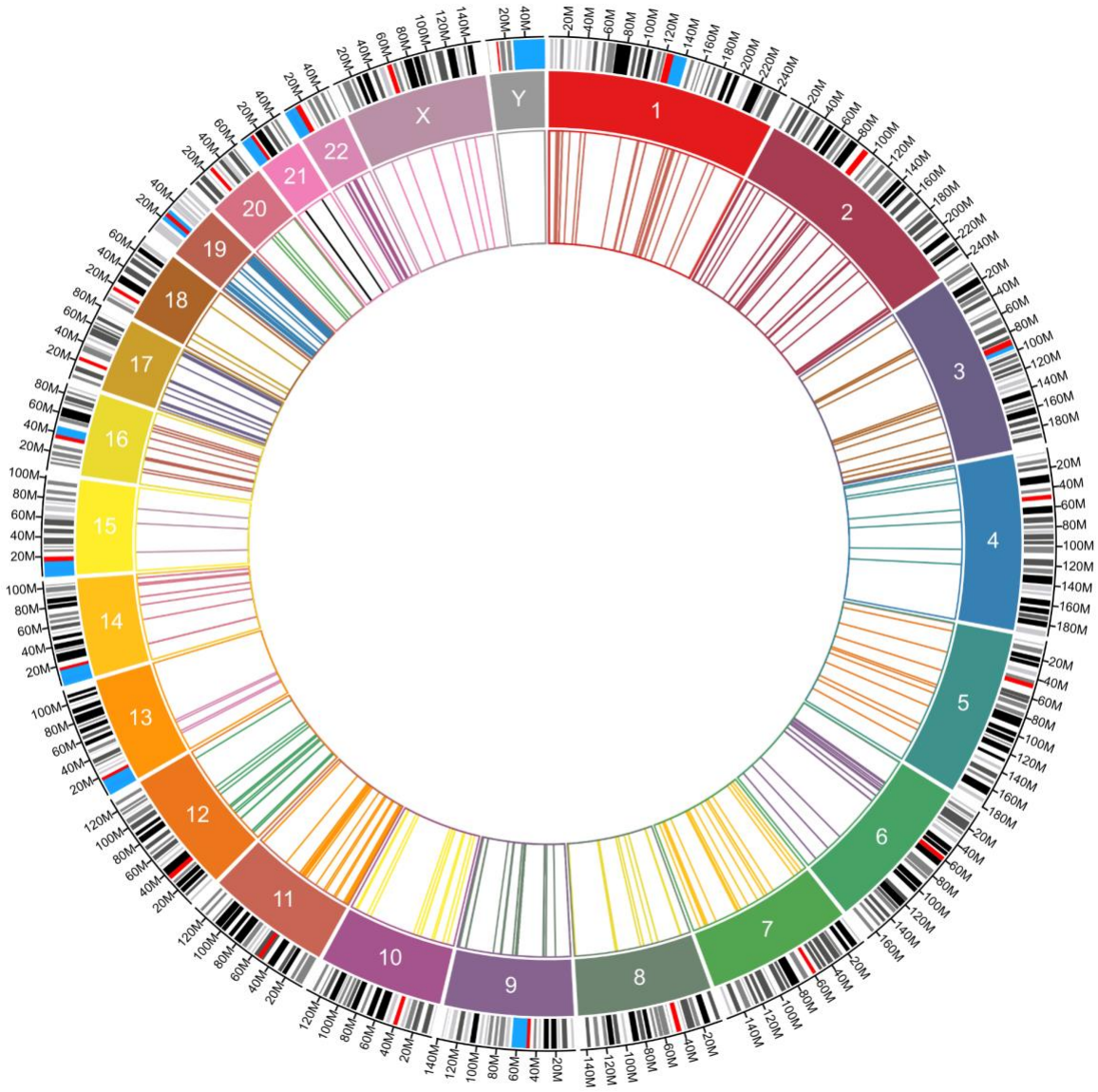
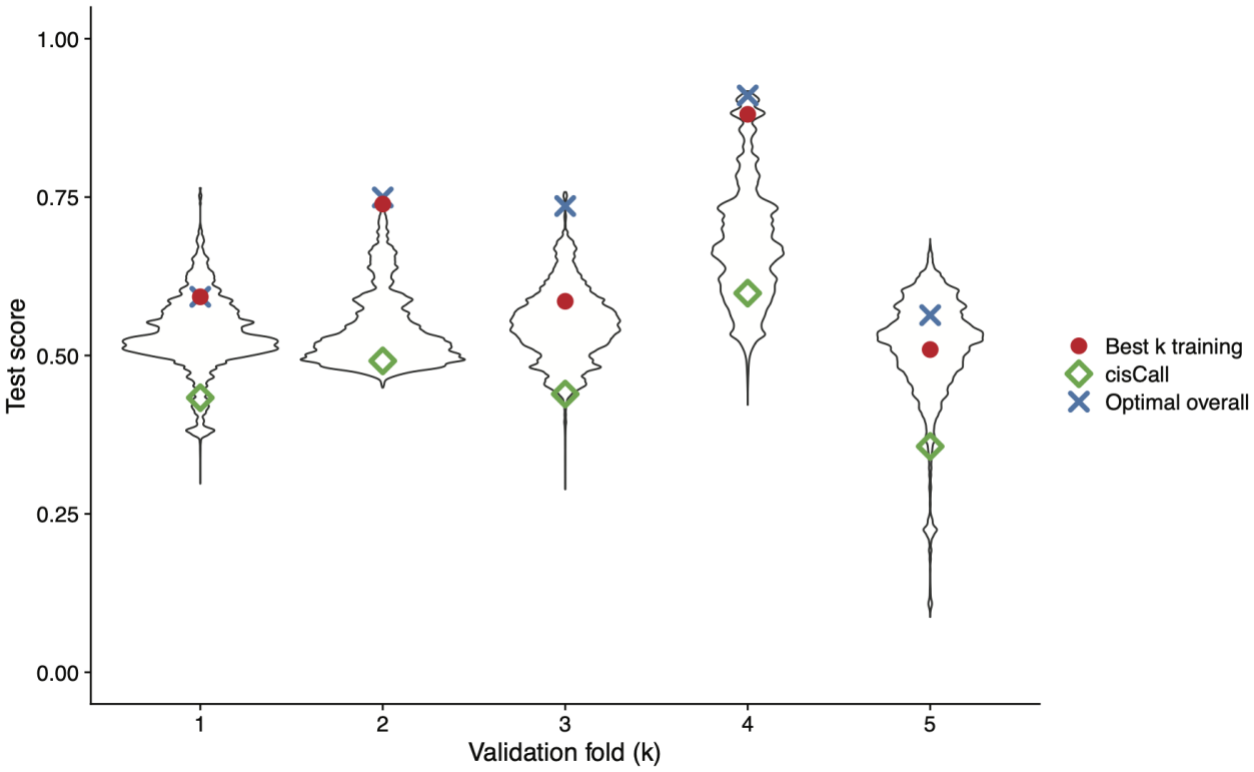
Supplementary figures



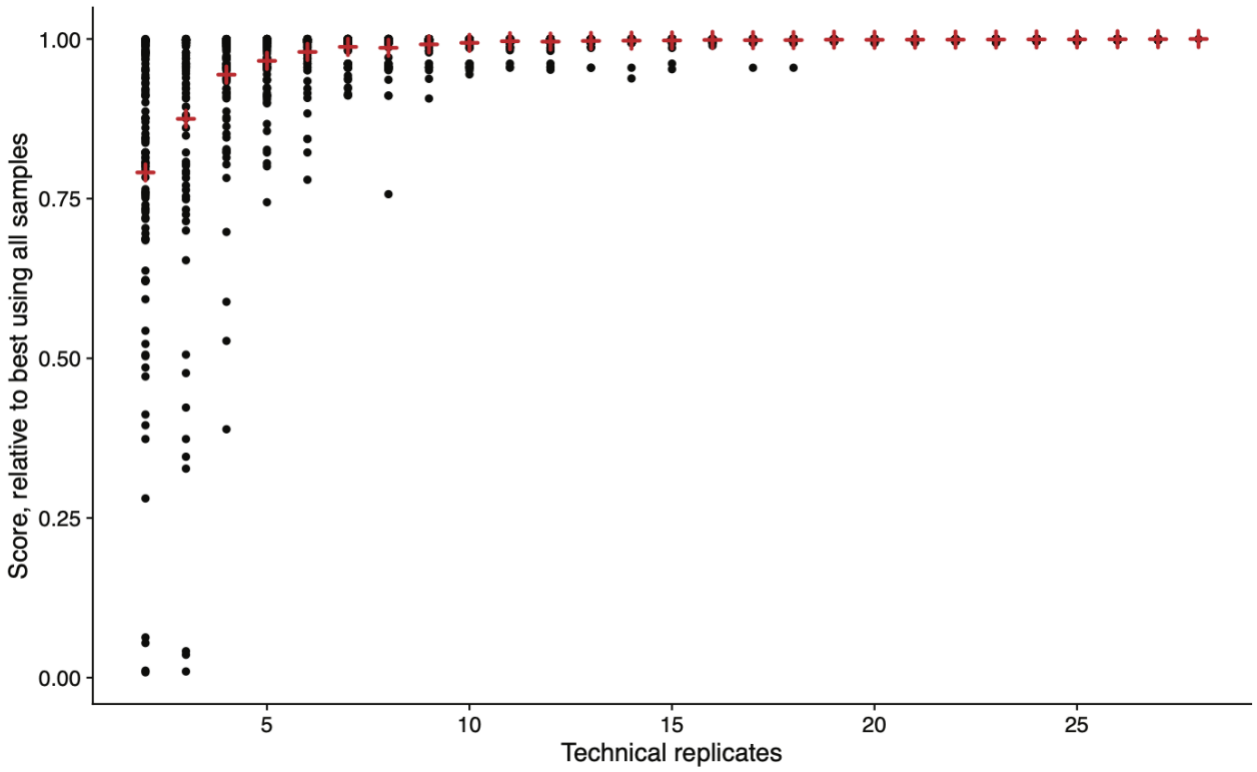
Supplementary figure 1S: Detailed flowchart of the algorithm used to estimate the genetic heterogeneity between two samples and details of its optimization. Inputs: aligned sequences (BAM files) of the two samples (A, in red; and B, in blue) and their healthy tissue control (N, in green), population allele frequency data from the gnomAD database (single nucleotide polymorphisms, SNPs, in purple), and user-specified configuration parameters (gear icon). Outputs: estimate of the genetic heterogeneity between samples A and B, and sets of variants (level of detail user-specified). All parameters that control this pipeline are detailed in the Parameters box, accompanied by the range of values assayed during optimization between parentheses. Boxes in the flowchart indicate the operations used on sets of variants in the different steps of the algorithm using set arithmetics. U = Union, I = Intersection, |x| = cardinality (number of elements).



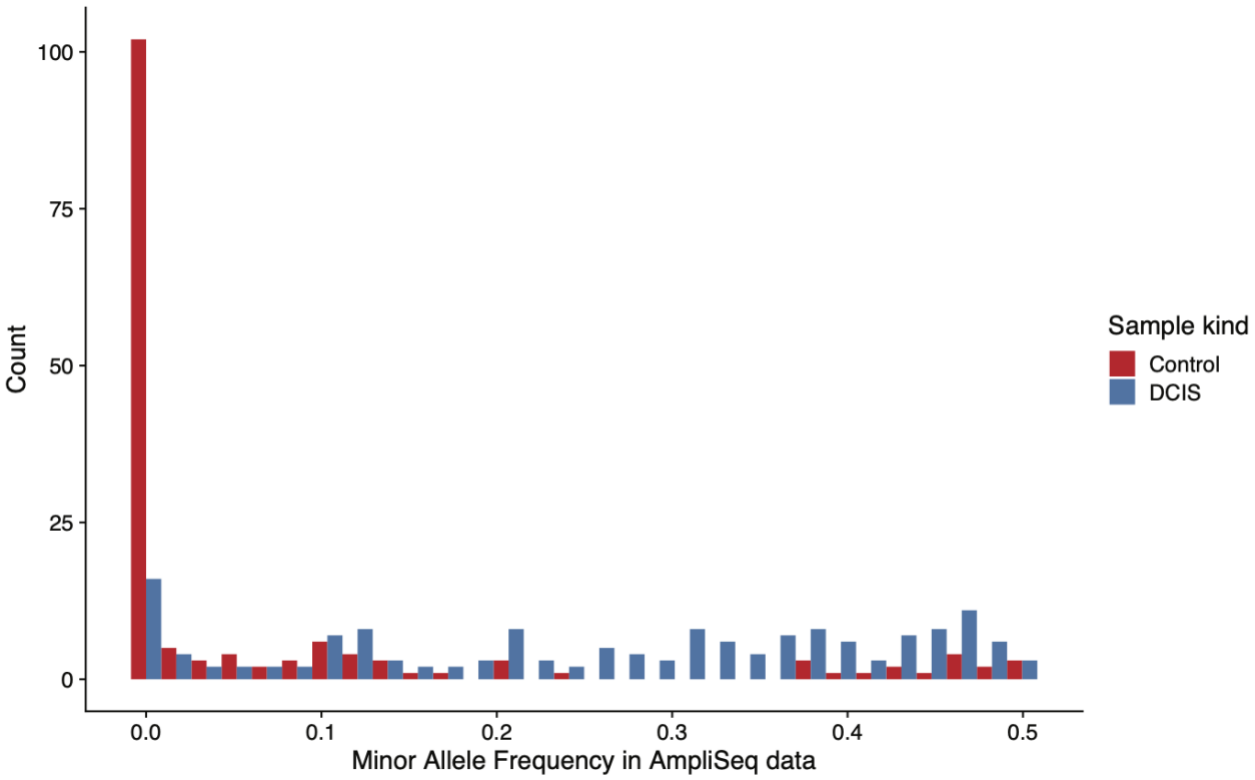
Supplementary figure 2S: Genomic distribution of mutations. The cancer variants (O) were distributed across the entire exome. The colored lines in the inner circle represent mutations.



Supplementary figure 3S: Cross-validation. 5-fold cross-validation using ITHE. For each fold (x axis) we report the distribution of training scores of all assayed parameter value combinations (violin), the test score of the optimized parameter values for that fold (red dot), the test score of the optimized parameter using all samples (blue cross), and the test score obtained by cisCall using only targeted regions (green diamond).



Supplementary figure 4S: Sensitivity analysis on the number of technical replicates. The x-axis shows the size of the datasets subsampled from our 28 technical replicates randomly without replacement. Each subsampled dataset was used to optimize ITHE independently, and its score on the full dataset (relative to the best) is shown as a black dot (y-axis). The mean score per dataset size is shown as a red cross.



Supplementary figure 5S: Distribution of minor allele frequencies in the validation data. Histograms of the minor allele frequencies estimated using the validation (AmpliSeq) sequencing data in the control (red) and DCIS samples (blue).