Materials and Methods

In vitro size selection. Between 8-20 ng of DNA were loaded into a 3% agarose cassette (HTC3010, Sage Bioscience), and size selection was performed on a PippinHT (Sage Bioscience) according to the manufacturer's protocol. Quality controls for in vitro size selection were performed on 20 healthy control samples and detailed in **fig. S6**. We observed an increase in duplicate reads with in vitro size selection, and therefore duplicate reads were removed for any downstream size selection analysis in the manuscript. To determine the sequencing noise, we used a QC metric called the median absolute pairwise difference (MAPD) algorithm. MAPD measures the absolute difference between the log2 CN ratios of every pair of neighboring bins and then takes the median across all bins. Higher MAPD scores reflect greater noise, typically associated with poor-quality samples. All samples exhibited a MAPD score of 0.01 (+-0.01), irrespective of the size selection condition.

TAm-Seq. Tagged-Amplicon Deep Sequencing libraries were prepared as previously described (*34*), using primers designed to assess single nucleotide variants (SNVs) and small indels across selected hotspots and the entire coding regions of *TP53*. Libraries were sequenced using MiSeq or HiSeq 4000 (Illumina).

sWGS. Indexed sequencing libraries were prepared using commercially available kits (ThruPLEX-Plasma Seq and/or Tag-Seq, Rubicon Genomics). Libraries were pooled in equimolar amounts and sequenced to <0.4x depth of coverage on a HiSeq 4000 (Illumina), generating 150-bp paired-end reads. Sequence data were analyzed using an in-house pipeline that consists of the following: Paired end sequence reads were aligned to the human reference genome (GRCh37) using BWA-mem after the removal of contaminating adapter sequences *(48)*. PCR and optical duplicates were marked using MarkDuplicates (Picard Tools) feature, and these were excluded from downstream analysis along with reads of low mapping quality and supplementary alignments. When necessary, reads were down-sampled to 10 million in all samples for comparison purposes.

Somatic copy number aberration analysis: The analysis was performed in R using a software suite for shallow Whole Genome Sequencing copy number analysis named CNAclinic (https://github.com/sdchandra/CNAclinic) as well as the QDNAseq pipeline *(49)*. Sequencing reads were randomly sampled to 10 million reads per dataset and allocated into equally sized (30 Kbp) non-overlapping bins throughout the length of the genome. Read counts in each bin were corrected to account for sequence GC content and mappability, and bins overlapping 'blacklisted' regions (derived from the ENCODE project + 1000 Genomes database) prone to artefacts were excluded from downstream analysis. Read counts in test samples were normalized by the counts from an identically processed healthy individual and $\log_2$ transformed to obtained copy number ratio values per genomic bin. Read counts in healthy controls were normalized by their median genome-wide count. Next, bins were segmented using both Circular Binary Segmentation and Hidden Markov Model algorithms, and an averaged $\log_2$R value per bin was calculated.

An in-house empirical blacklist of aberrant read count regions was constructed as follows: 65 sWGS datasets from healthy plasma were used to calculate median read counts per 30 Kbp genomic bin as a function of GC content and mappability. A 2D LOESS surface was applied, and the difference between the actual count and the LOESS fitted values was calculated. The median of these residual values across the 65 controls was calculated for each genomic bin. Regions with median residuals greater than 4 standard deviations were blacklisted. The averaged segmental $\log_2$R values in each test sample that overlap this cfDNA blacklist were trimmed, and the median absolute value was calculated. This score was defined as t-MAD or the trimmed median absolute deviation from $\log_2 R = 0$. The R code to reproduce this analysis is provided in https://github.com/sdchandra/tMAD.

WES. Indexed sequencing libraries were prepared as described above (see sWGS). Plasma DNA libraries from each sample were made and pooled together for exome capture (TruSeq Exome Enrichment Kit, Illumina). Pools were concentrated using a SpeedVac vacuum concentrator (Eppendorf). Exome enrichment was performed following the manufacturer's protocol. Enriched libraries were quantified using quantitative PCR (KAPA library quantification, KAPA Biosystems), and DNA fragment sizes observed by Bioanalyzer (2100 Bioanalyzer, Agilent Genomics) and pooled in equimolar ratio for paired-end next generation sequencing on a HiSeq4000 (Illumina). Sequencing reads were de-multiplexed, allowing zero mismatches in barcodes. Paired-end alignment to the GRCh37 reference genome was performed using BWA-mem for all exome sequencing data, including germline, plasma, and tumor tissue DNA where generated. PCR duplicates were marked using Picard. Base quality score recalibration and local realignment were performed using Genome Analysis Tool Kit (GATK).

Mutation calling. MAFs for each single-base locus were calculated with MuTect2 for all bases with PHRED quality ≥30. After MuTect2, we applied filtering parameters so that a mutation was called if no mutant reads for an allele were observed in germline DNA at a locus that was covered at least 10x, and if at least 4 reads supporting the mutant were found in the plasma data with at least 1 read on each strand (forward and reverse). At loci with <10x coverage in normal DNA and no mutant reads, mutations were called in plasma if a prior plasma sample showed no evidence of a mutation and was covered adequately (10x or more). We aggregated mutations called before and after size selection with a method called Integrated Signal Amplification for Non-invasive Interrogation of Tumors. This method combines different subsets of mutations called from the same plasma DNA sample using different processing approaches. The mutation aggregation as used in this study is formalized as follows: aggregated mutations = mutations detected without size selection **U** (mutations detected with in vitro size selection **U** mutations detected with in silico size selection).

In silico size selection: Paired-end reads were generated by sequencing DNA from both ends of the fragments present in the library. The original length of the DNA could be inferred using the mapping locations of the read ends in the genome. Once alignment was complete, Samtools software was used to select paired reads that

correspond to fragment lengths in a specific range. Mutect2 was used to call mutations from these in silico size selected data as described in the previous section.

Tumor-guided capture sequencing. Matched tumor tissue DNA and plasma DNA samples of 19 patients with advanced cancer from the RigsHospitalet (Copenhagen, Denmark) were sequenced by WES. Variants were called from these samples as previously described (see Mutation calling). Hybrid-based capture for longitudinal plasma sample analysis was designed to cover these variants for each patient using SureDesign (Agilent). A median of 160 variants were included per patient, and in addition, 41 common genes of interest for pan-cancer analysis were included in the tumor-guided sequencing panel. Indexed sequencing libraries were prepared as described above (see sWGS). Plasma DNA libraries from each sample were made and pooled together for tumor-guided capture sequencing (SureSelect, Agilent). Pools were concentrated using a SpeedVac vacuum concentrator (Eppendorf). Capture enrichment was performed following the manufacturer's protocol. Enriched libraries were quantified using quantitative PCR (KAPA library quantification, KAPA Biosystems), and DNA fragment sizes controlled by Bioanalyzer (2100 Bioanalyzer, Agilent Genomics) and pooled in equimolar ratio for paired-end next generation sequencing on a HiSeq4000 (Illumina). Sequencing reads were de-multiplexed, allowing zero mismatches in barcodes. Paired-end alignment to the GRCh37 reference genome was performed using BWA-mem for all exome sequencing data including germline, plasma, and tumor tissue DNA where generated. PCR duplicates were marked using Picard. Base quality score recalibration and local realignment were performed using Genome Analysis Tool Kit (GATK).

Classification analysis. The preliminary analysis was carried out on 304 samples (182 high ctDNA cancer samples, 57 low ctDNA cancer samples, and 65 healthy controls). For each sample, the following features were calculated from sWGS data: t-MAD, amplitude_10bp, P(20-150), P(160-180), P(20-150)/P(160-180), P(100-150), P(100-150)/P(163-169), P(180-220), P(250-320), P(20-150)/P(180-220). The data were arranged in a matrix where the rows represented each sample and the columns held the aforementioned features with an extra "class" column with the binary labels of "cancer"/"healthy". The following analysis was carried out in R using *RandomForest, caret*, and *pROC* packages. The pairwise correlations between the features were calculated to assess multi-collinearity in the dataset (**fig. S19**). A single variable was selected for removal from pairs with Pearson correlation > 0.75. Highly correlated fragmentation features that were composite of individual variables already in the dataset, such as P(20-150)/P(180-220), were prioritized for removal. The features were also assessed for zero variance and linear dependencies, but none were flagged. After this pre-processing, the following 5 variables were selected for further analysis: t-MAD, amplitude_10 bp, P(160-180), P(180-220), and P(250-320). All 57 low ctDNA samples were set aside for validation of the models. The data matrix for the remaining high ctDNA cancer samples and healthy controls (n = 247) was randomly partitioned in a 60:40 split into 1 training and 1 validation dataset with the different cancer types and healthy samples represented in similar proportions. Hence, the training data contained 153 samples (cancer=114, healthy=39), and the first validation set of high ctDNA cancers contained 94 samples (cancer=68,

healthy=26). This validation dataset was only used for final assessment of the classifiers.

Classification of samples as healthy or cancer was performed using one linear and one non-linear machine learning algorithm, namely logistic regression (LR) and random forest (RF). Each algorithm was paired with recursive feature selection to identify the best predictor variables. This analysis was carried out with *caret* within the framework of 5 repeats of 10-fold cross-validation on the training set. The algorithm was configured to explore all possible subsets of the features. The optimal model for each classifier was selected using ROC metric. Separately, a logistic regression model was trained only using t-MAD as a predictor to assess the difference in performance without the addition of fragmentation features. Finally, the 68 high ctDNA cancer samples, 57 low ctDNA cancer samples, and 26 healthy controls set aside for validation were used to test the classifiers, using the area under the curve in a ROC analysis to quantify their performance.

A secondary analysis was carried out on the same training and validation cohorts, with the only difference being the features used in the model. Here, we tested the predictive ability of fragmentation features without the addition of information from SCNAs such as t-MAD. Hence the features used were: amplitude_10 bp, P(160-180), P(180-220), and P(250-320).

Quantification of the 10 bp periodic oscillation. The amplitude of the 10 bp periodic oscillation observed in the size distribution of cfDNA samples was determined from the sWGS data as follows. Local maxima and minima in the range 75 bp to 150 bp were identified. The average of their positions across the samples was calculated (for minima: 84, 96, 106, 116, 126, 137, 148, and maxima: 81, 92, 102, 112, 122, 134, 144). To compute the amplitude of the oscillations with 10 bp periodicity observed below 150 bp, we calculated the sum of the heights of the maxima and subtracted the sum of the minima. The larger this difference, the more distinct were the peaks. The height of the x bp peak was defined as the number of fragments with length x divided by the total number of fragments. To define local maxima, we selected the positions y such that y was the largest value in the interval [y-2, y+2]. The same rationale was used to pick minima.

Supplementary figures:

Supplementary figure 1:

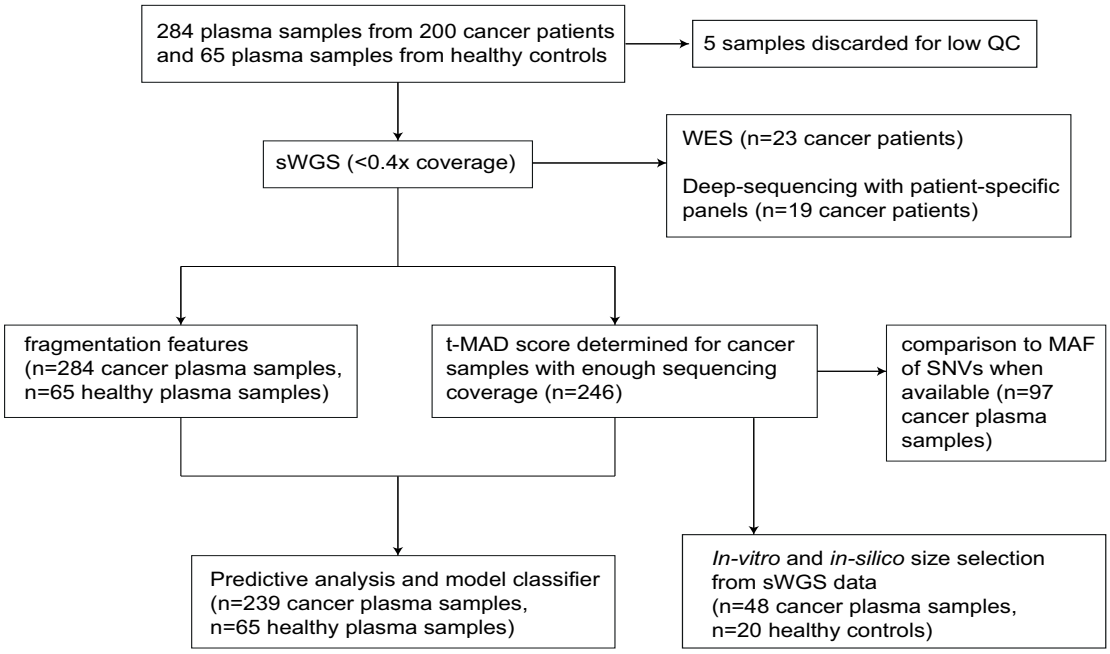


Fig. S1. Flowchart summarizing the experiments performed in this study and the sample numbers used at each step.

Supplementary figure 2:

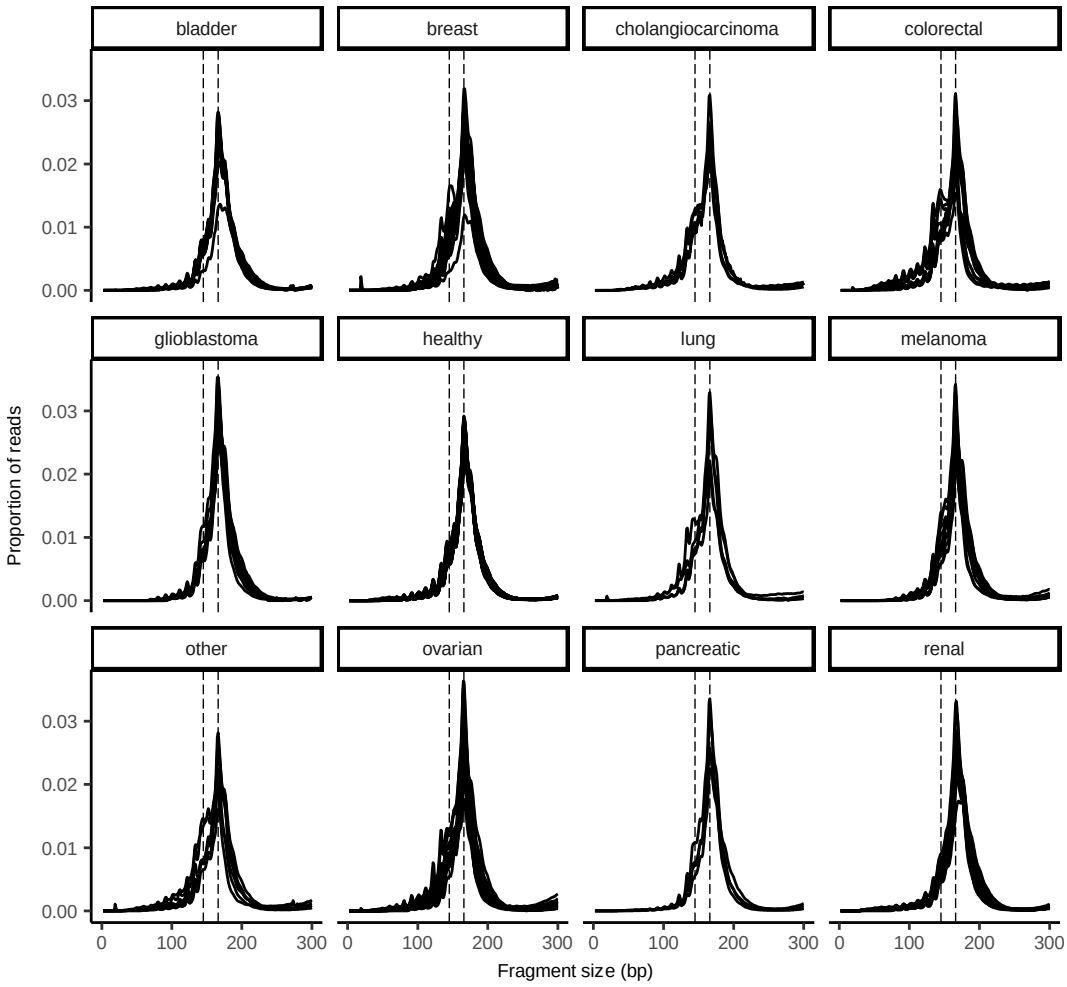


Fig. S2. Size distribution of cfDNA determined by sWGS for different cancer types. Size distribution of cfDNA determined by sWGS for all plasma samples of healthy individuals and cancer patients included in this study grouped by cancer type. The size profile of cfDNA from healthy controls (n=65) is also shown.

Supplementary figure 3:

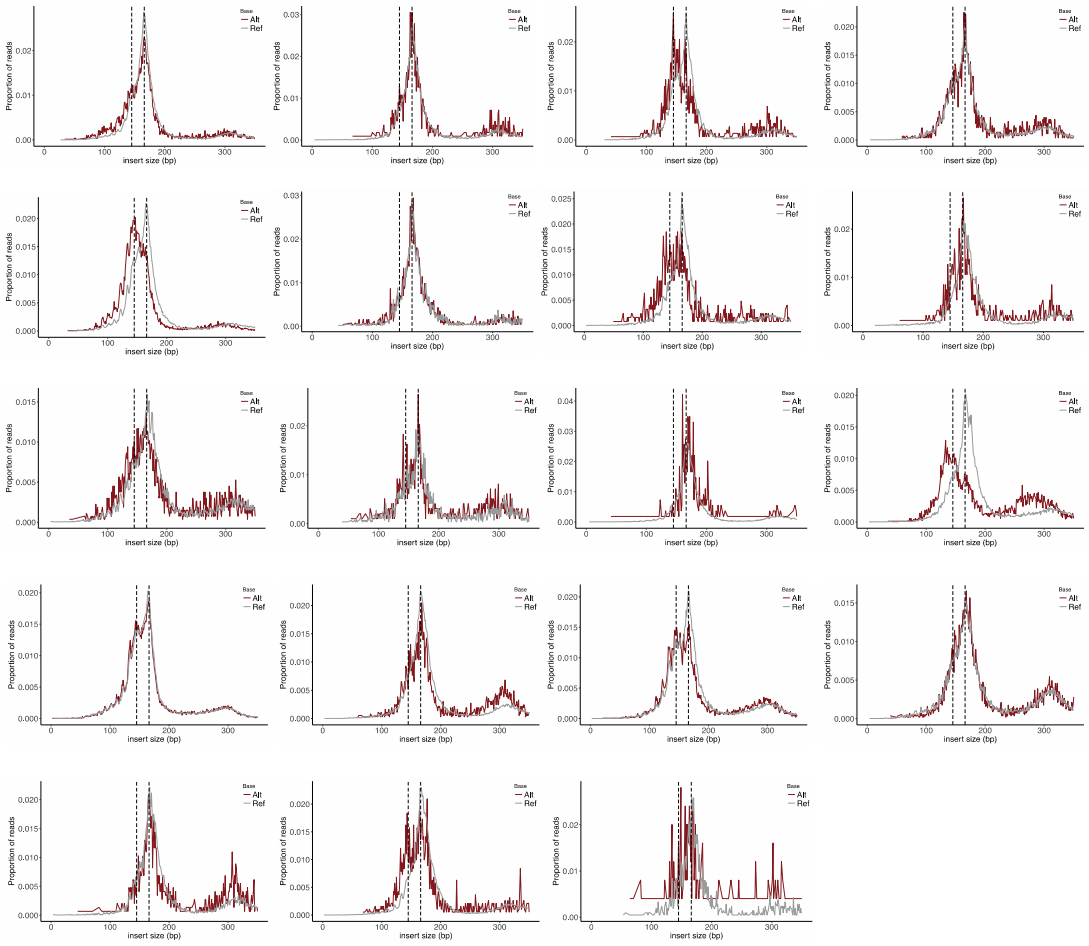


Fig. S3. Insert size distribution of mutant cfDNA determined with hybrid-capture sequencing for 19 patients. DNA fragment size distribution determined by hybrid-capture sequencing for 19 samples included in the mutant DNA size distribution analysis. The size distribution of DNA fragments carrying mutations identified in the corresponding patient tissue samples is shown in red, and the distribution of reference cfDNA from the same sample is shown in gray. The vertical dashed lines indicate 145 bp and 167 bp.

Supplementary figure 4:

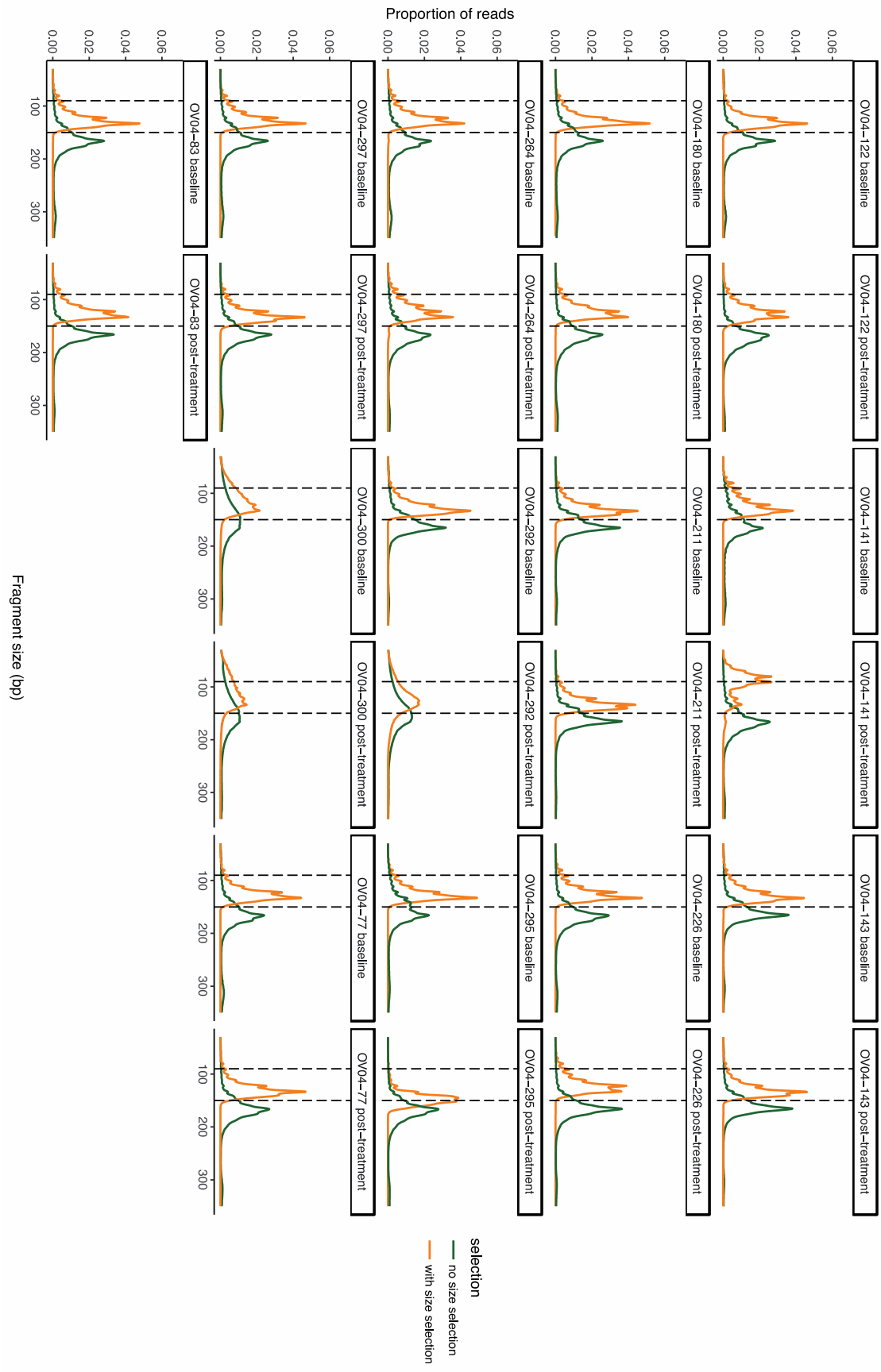


Fig. S4. DNA fragment size distribution for plasma samples from patients with ovarian cancer. DNA fragment size distribution determined by sWGS for 25 plasma samples from the 13 patients with ovarian cancer, collected before and after treatment. The distribution of cell-free DNA (cfDNA) without size selection is shown is green, and the distribution of the same cfDNA samples after in vitro size selection is shown in orange. The vertical lines represent the range of fragments selected with the PippinHT cassettes, between 90 bp and 150 bp. Samples from patient OV04-292 and patient OV04-300 exhibited an altered fragmentation profile, indicating a possible issue with the preparation or pre-analytical preservation of the samples.

Supplementary figure 5:

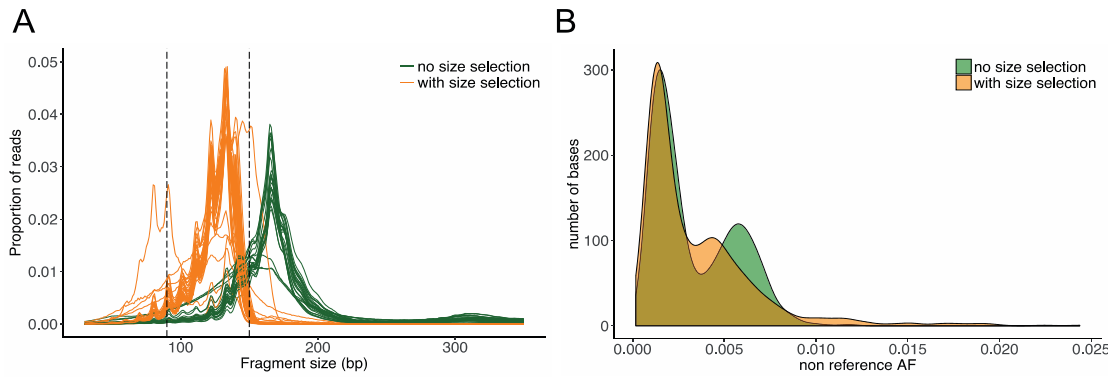


Fig. S5. Quality control assessed for in vitro size selection. A) Size distribution of DNA fragments from the plasma samples included in the size selection study, assessed by sWGS, before size-selection (green) and after in vitro size-selection (orange). The two dotted vertical lines indicate the size selection range between 90 bp and 150 bp. B) Proportion of non-reference allele fractions corresponding to the sequencing background noise as determined during targeted sequencing (TAm-Seq) of plasma DNA samples from ovarian cancer patients, with and without in vitro size selection.

Supplementary figure 6:

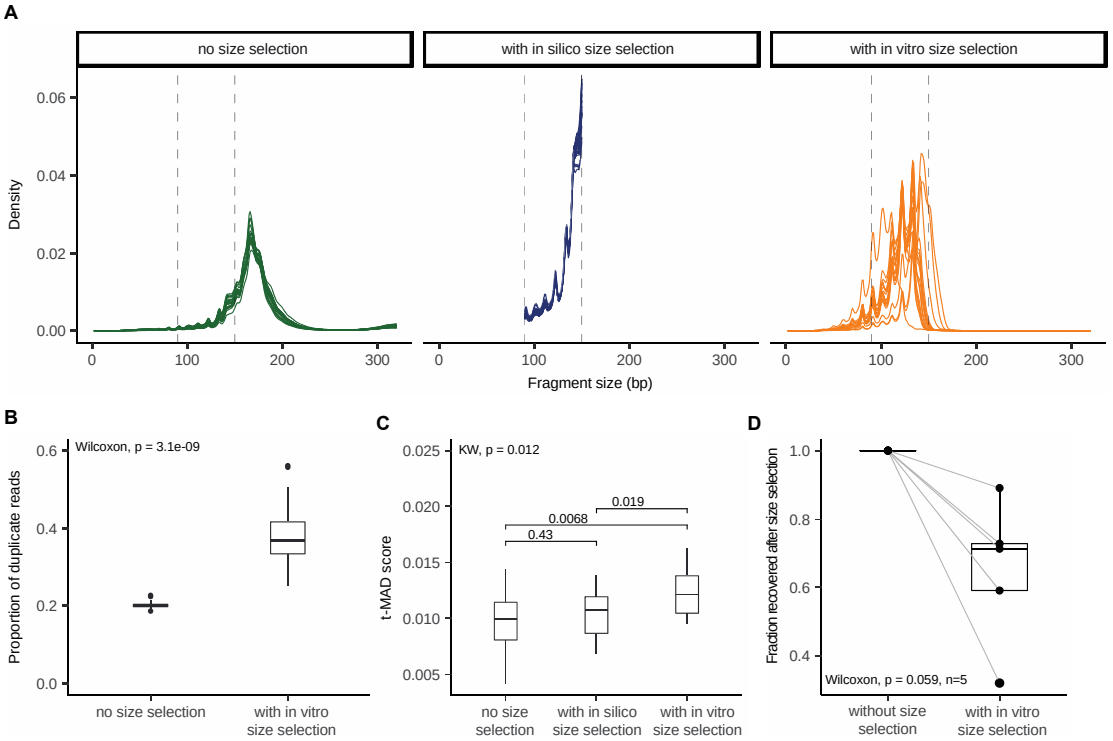


Fig. S6. Quality control assessed for in vitro and in silico size selection on healthy control samples. We selected 20 plasma samples from healthy controls, extracted DNA, and performed sWGS without size selection or with in vitro or in silico size selection. A) The size profile for each of these methods, showing one line for each sample. B) The fraction of duplicated reads without and with in vitro size selection. We observed an increase in duplicate reads with in vitro size selection, and therefore duplicate reads have been removed for any downstream size selection analysis in the manuscript (see Materials and Methods). C) t-MAD scores in 20 control samples without size selection and with in vitro or in silico size selection. The t-MAD score from the samples without size selection was not significantly different from the t-MAD determined after in silico size selection (t-test, p=0.43), but a significant difference was observed after in vitro size selection (t-test, p=0.0068). The mean (0.011) and the maximum (0.016) for t-MAD were within the threshold limit determined empirically from the cohort of controls (n=65). D) The yield of DNA recovered after in vitro size selection. This yield is estimated from bioanalyzer data in the size range of interest.

Supplementary figure 7:

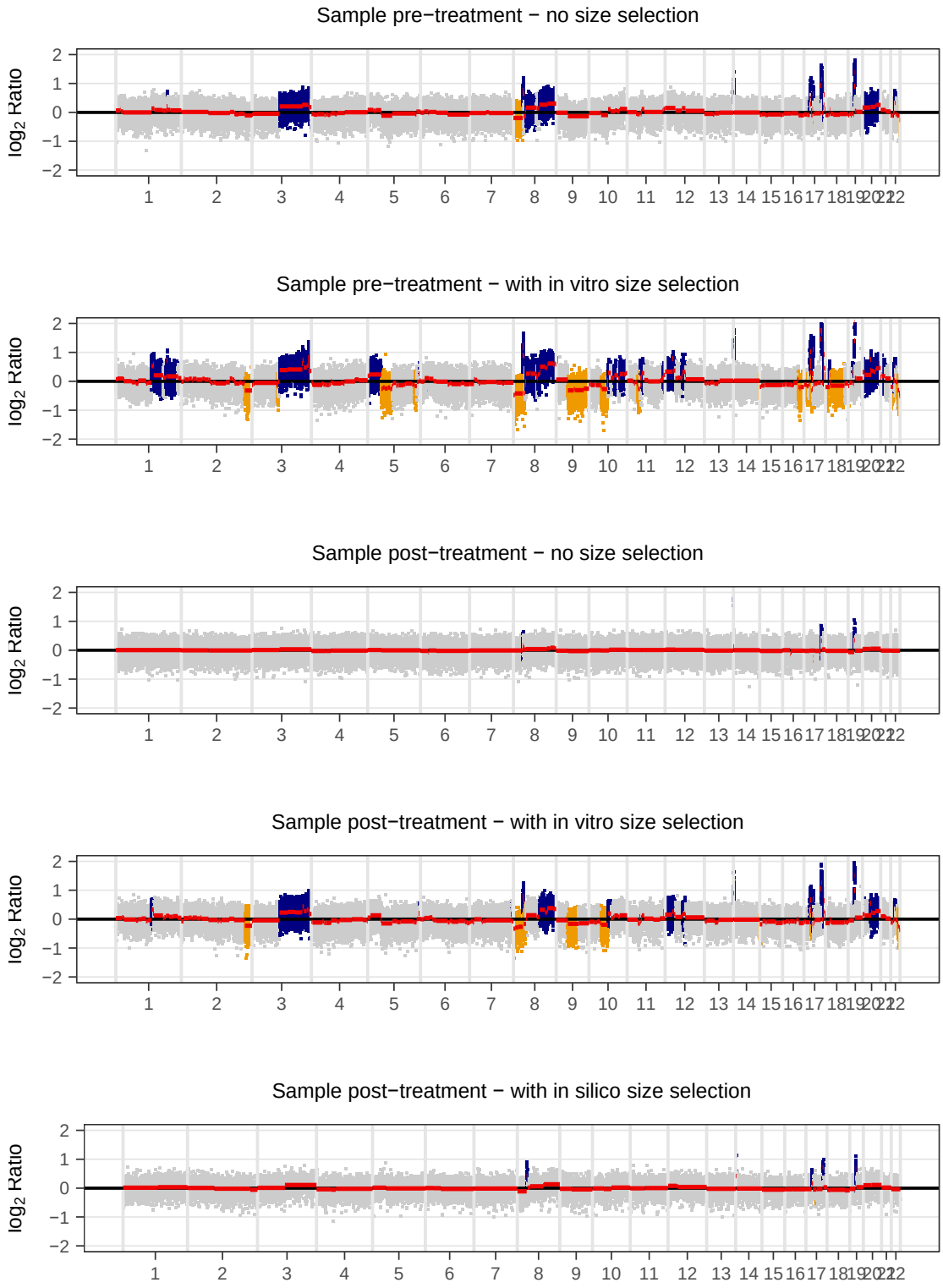
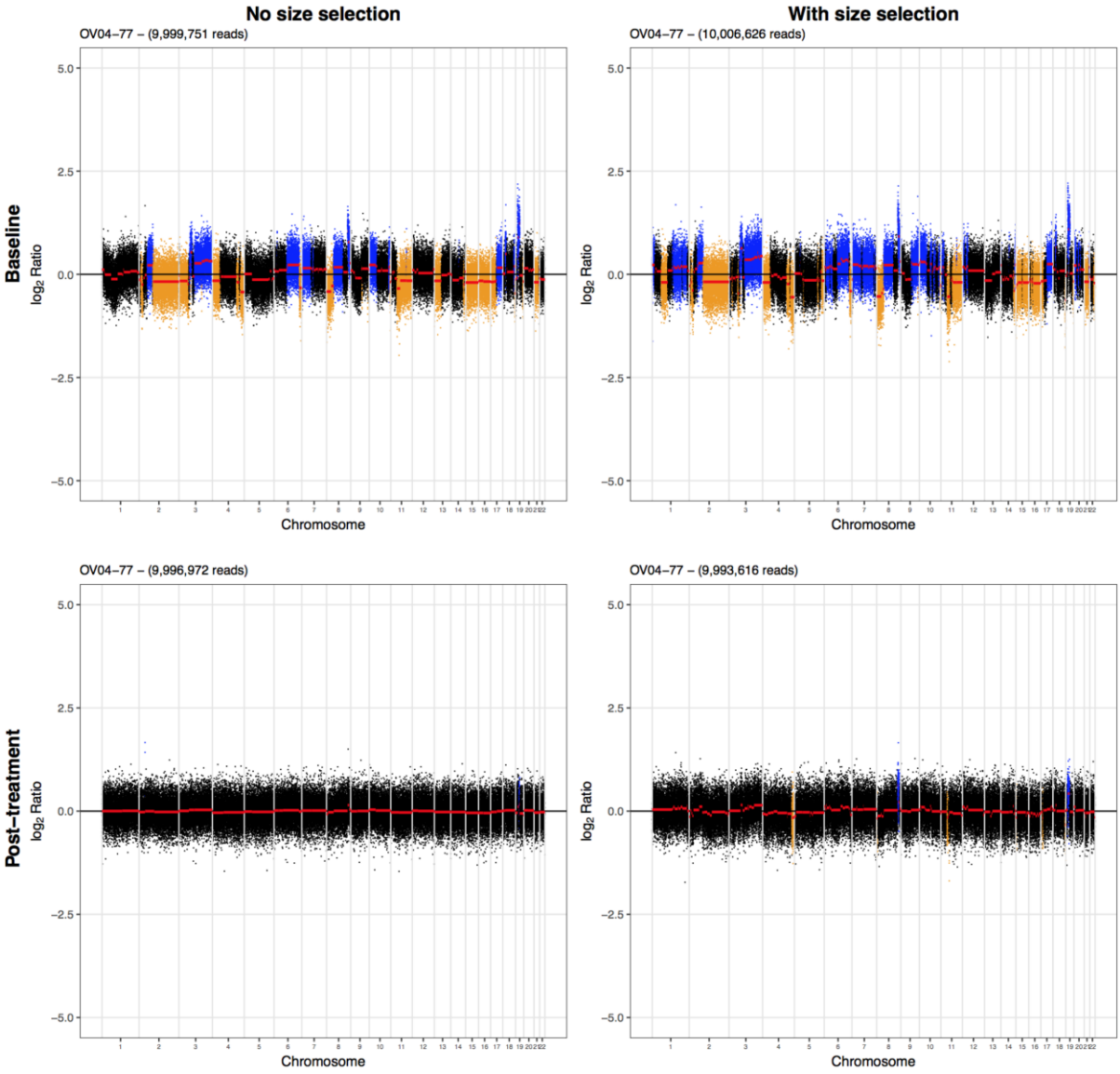


Fig. S7. SCNA analysis of the segmental $\log_2$ ratio determined after sWGS (<0.4× coverage) for the patient OV04-83. This shows the data presented in Fig. 3 (for reference and comparison), together with additional

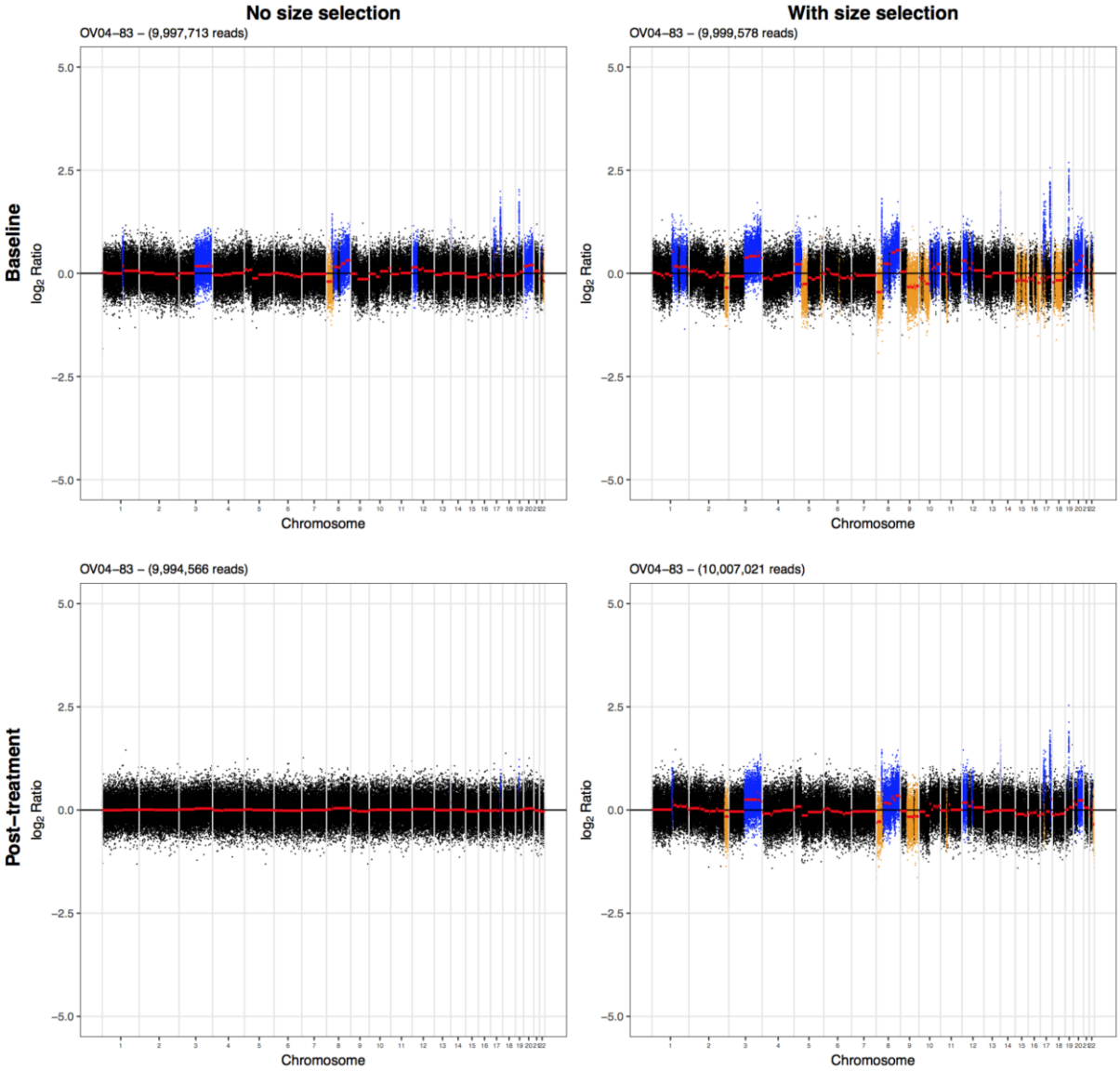
data for the sample collected pre-treatment analyzed after in vitro size selection, and the sample that was collected post-treatment analyzed after in silico size selection for DNA fragments between 90-150 bp.

Supplementary figure 8:

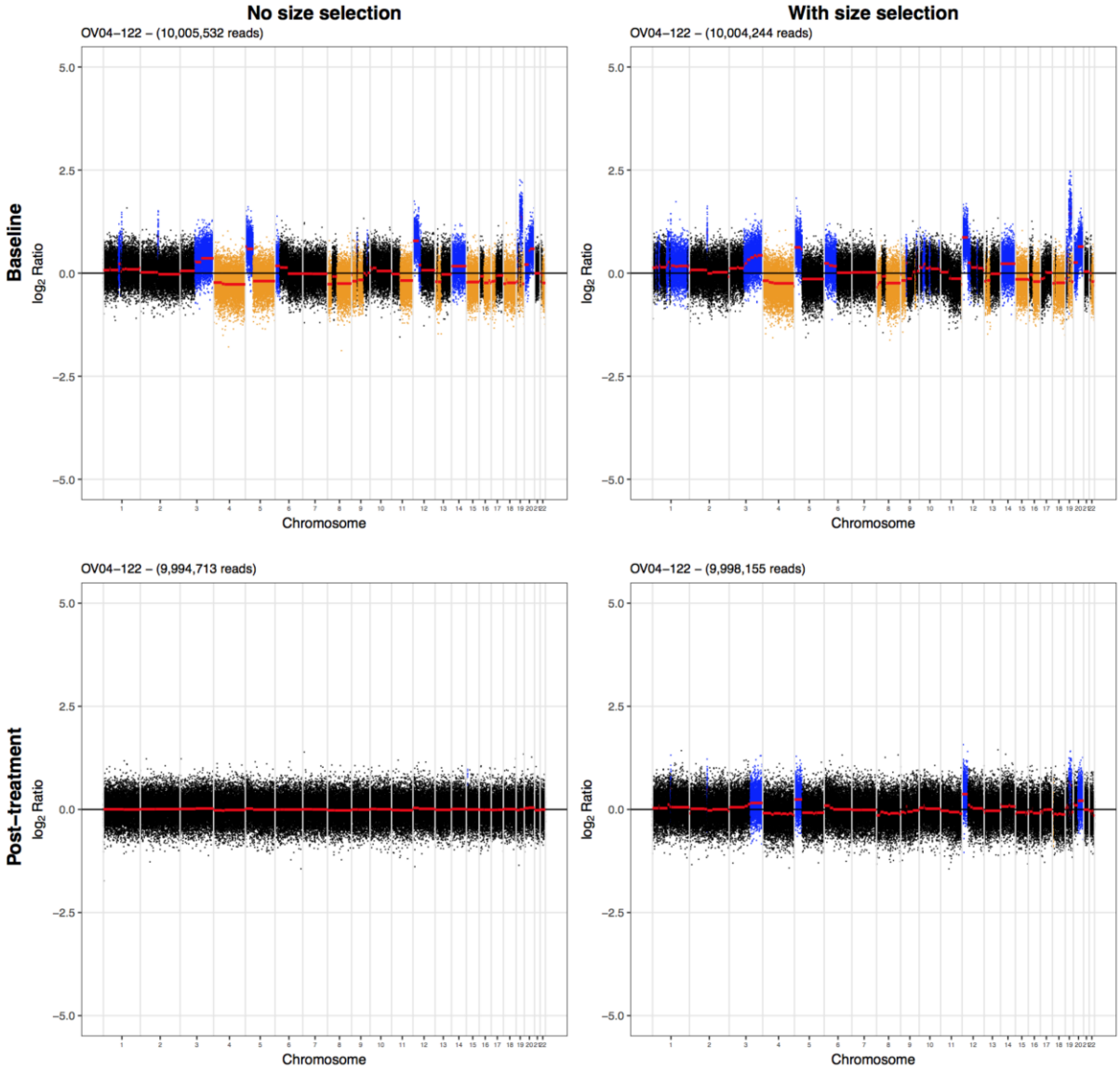
Patient OV04-77: the TP53 MAF determined by TAm-Seq was 0.346 at baseline and 0.068 post-treatment. The enrichment after in vitro size selection was 1.33 times at baseline and 1.89 times post-treatment.



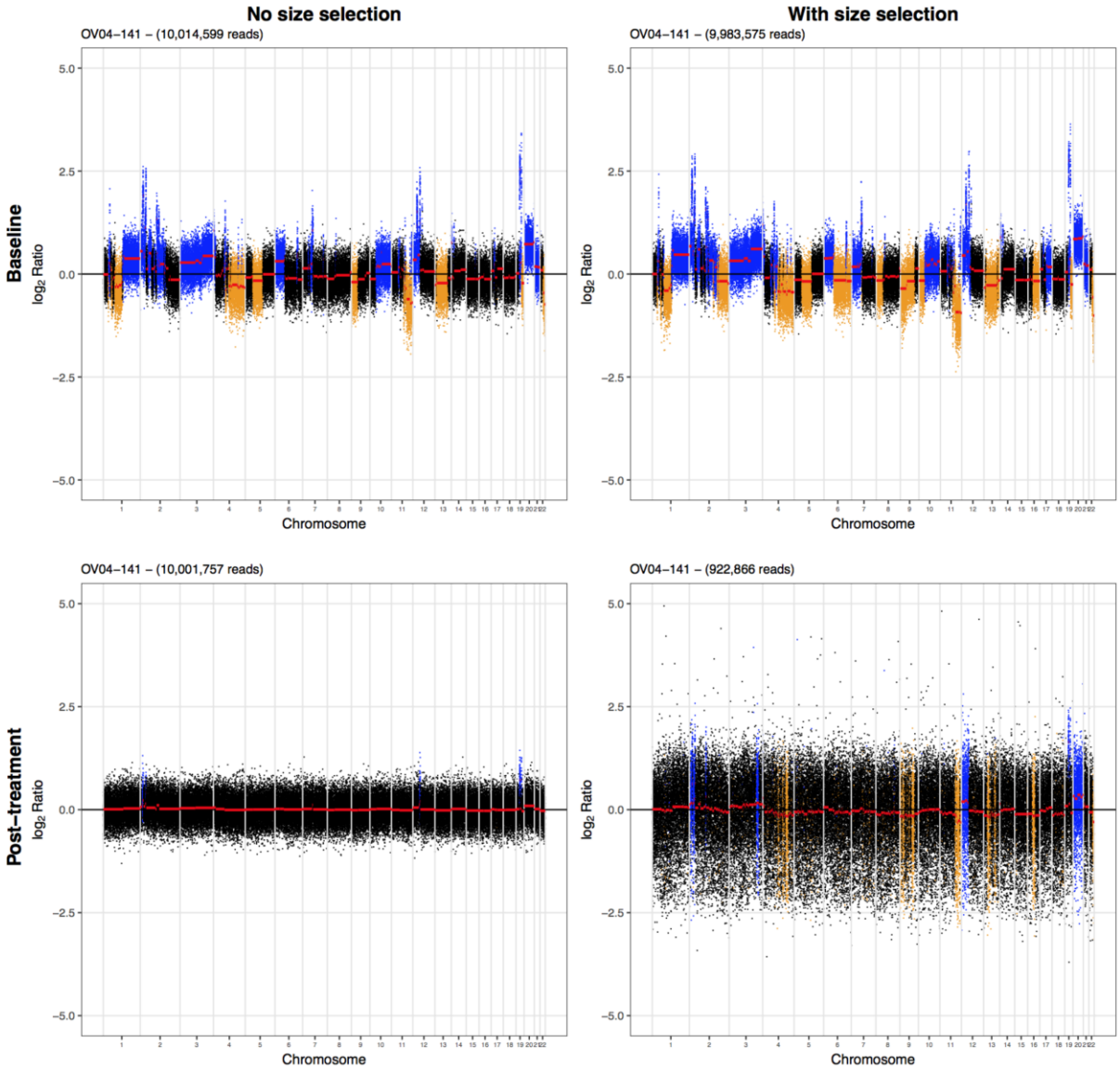
Patient OV04-83: the TP53 MAF determined by TAm-Seq was 0.271 at baseline and 0.068 post-treatment. The enrichment after in vitro size selection was 2.69 times at baseline and 6.87 times post-treatment.



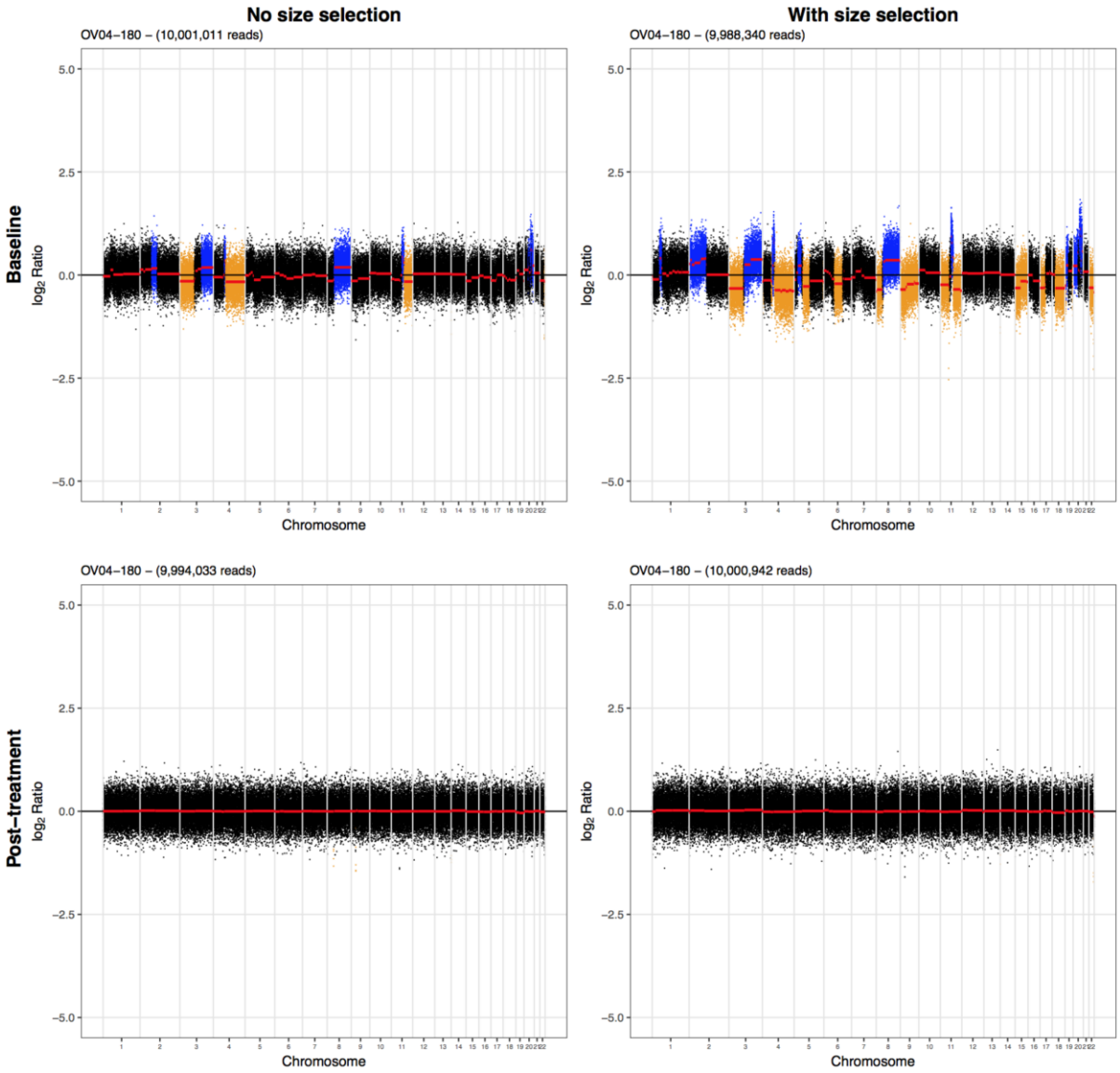
Patient OV04-122: the TP53 MAF determined by TAm-Seq was 0.483 at baseline and 0.036 post-treatment. The enrichment after in vitro size selection was 1.03 times at baseline and 6.28 times post-treatment.



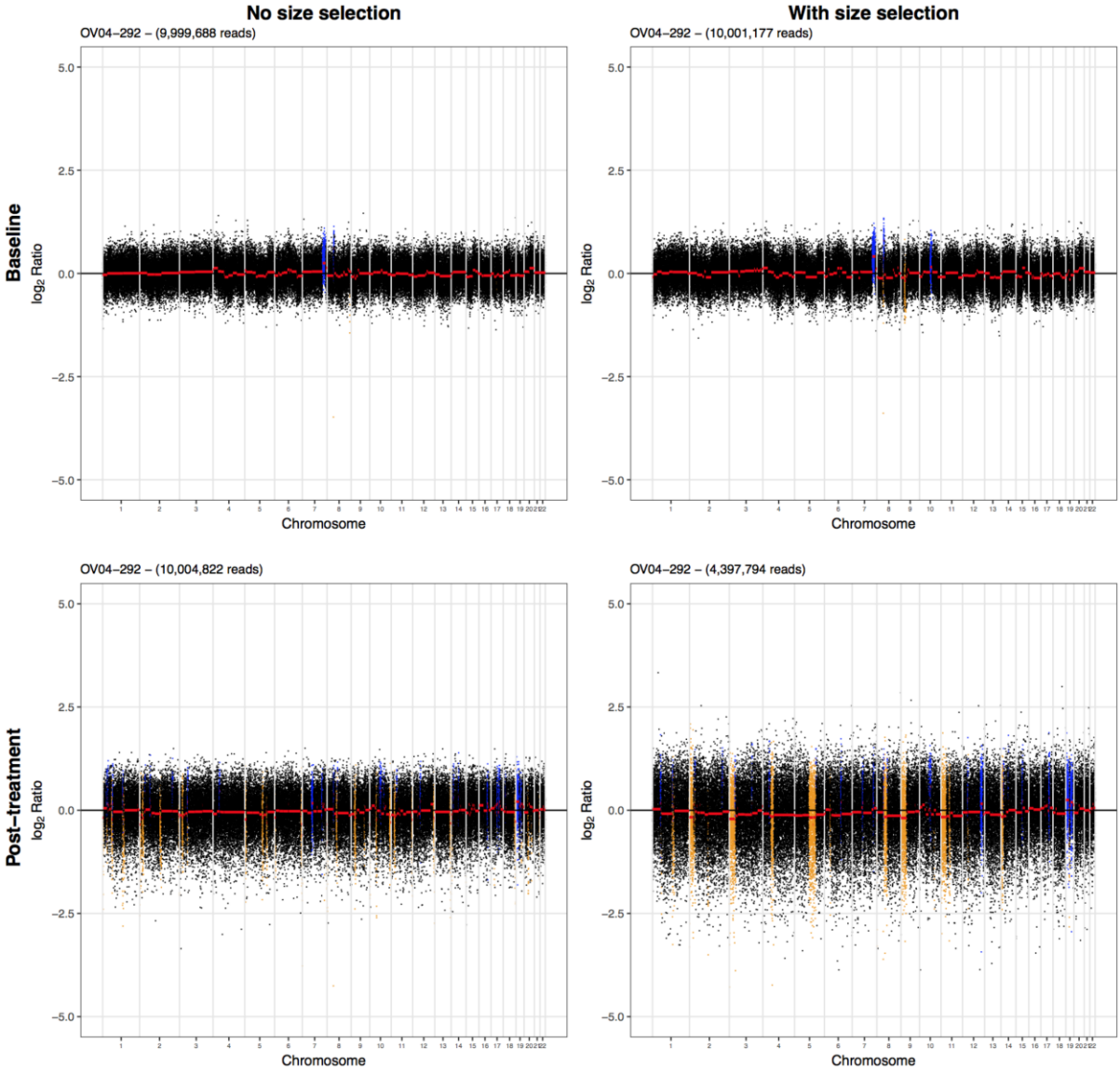
Patient OV04-141: the TP53 MAF determined by TAm-Seq was 0.610 at baseline and 0.064 post-treatment. The enrichment after in vitro size selection was 1.27 times at baseline and 4.91 times post-treatment.



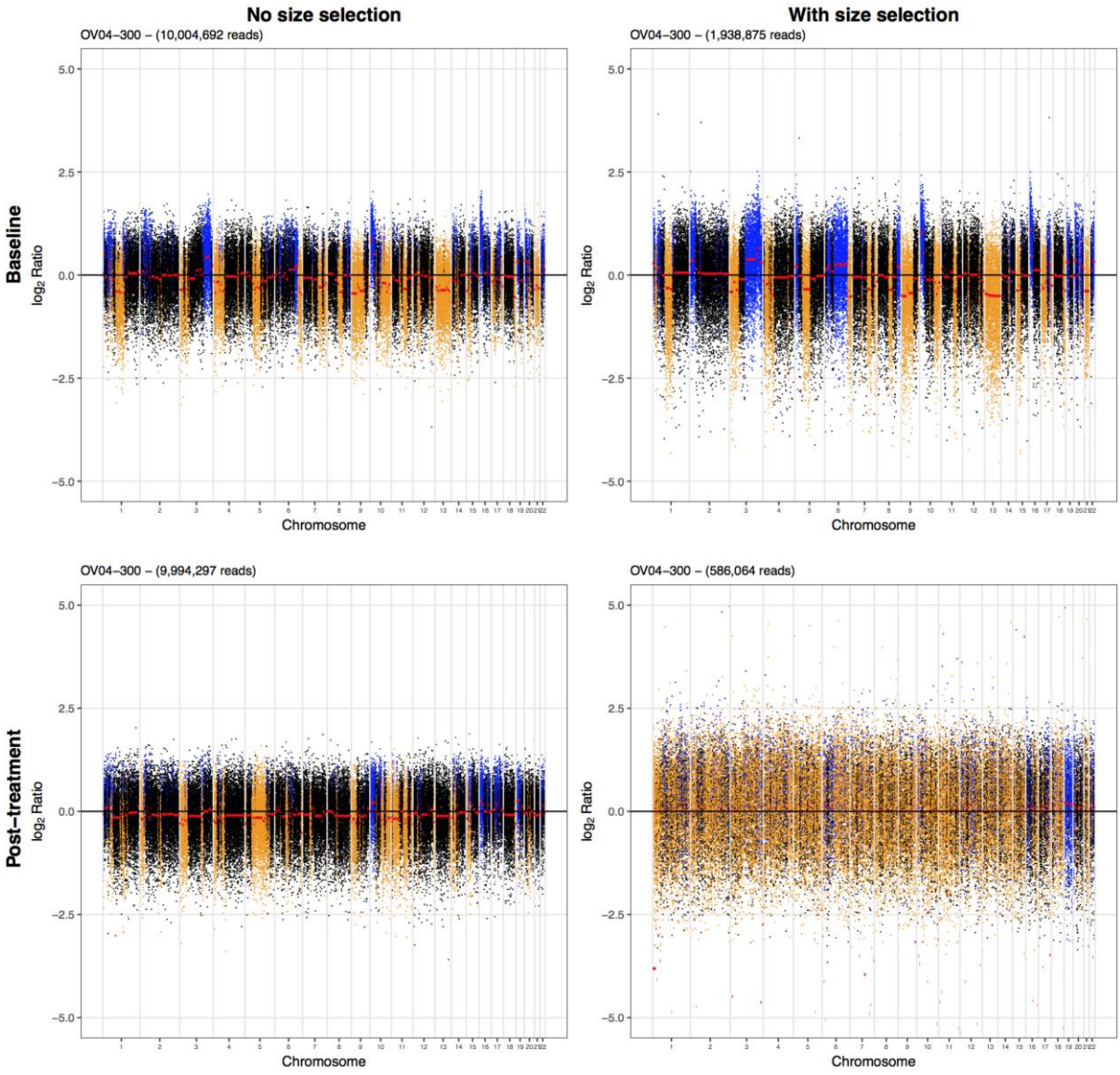
Patient OV04-180: the TP53 MAF determined by TAm-Seq was 0.212 at baseline and 0.001 post-treatment. The enrichment after in vitro size selection was 2.83 times at baseline and 1.81 times post-treatment.



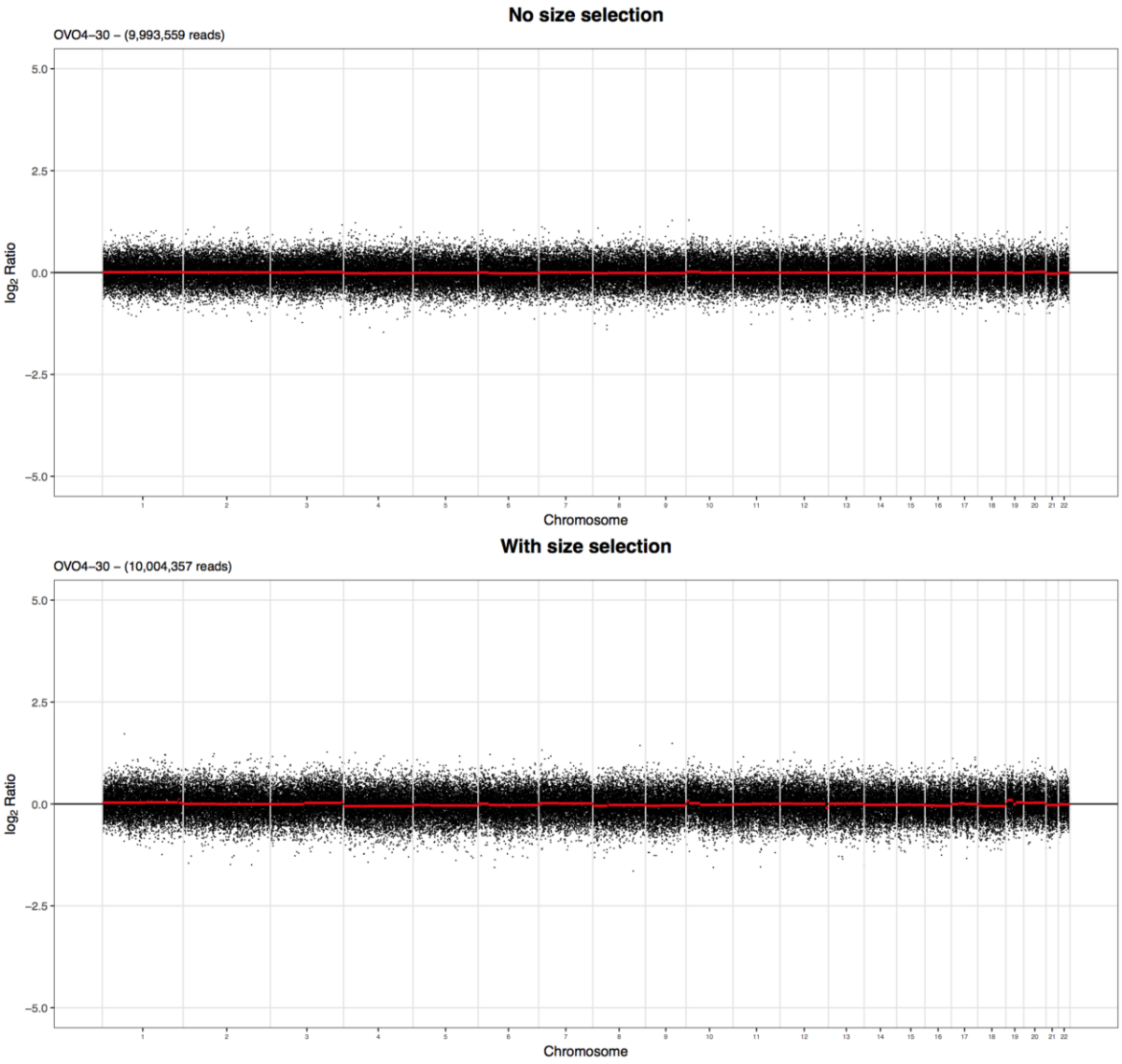
Patient OV04-292: the TP53 MAF determined by TAm-Seq was 0.147 at baseline and 0.069 post-treatment. The enrichment after in vitro size selection was 1.28 times at baseline and 1.89 times post-treatment. The samples collected post-treatment for this patient exhibited a degraded pattern of fragmentation (see fig. S4). The size selection was affected by this pattern, and the number of reads after sequencing was <10M.



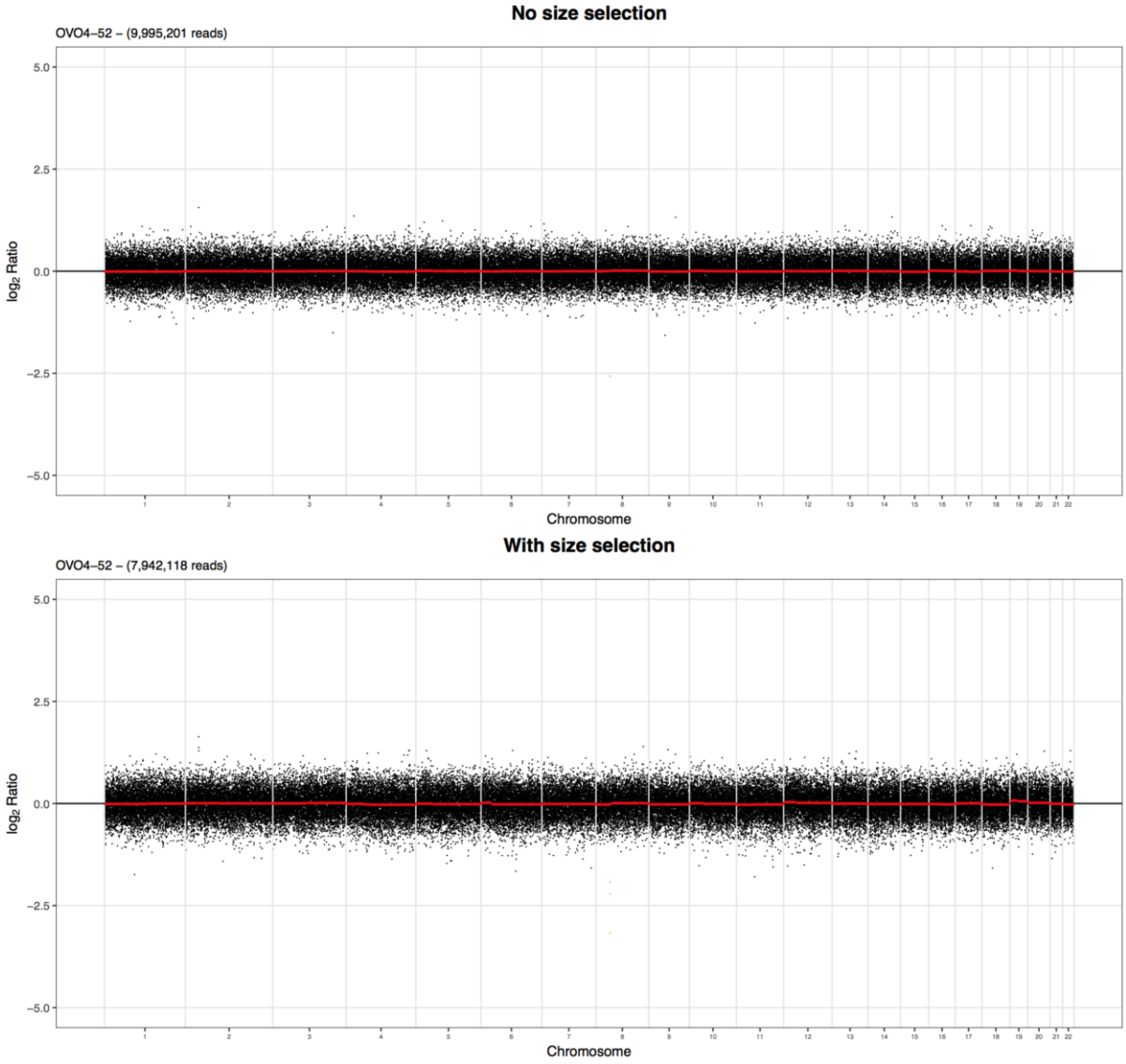
Patient OV04-300: the TP53 MAF determined by TAm-Seq was 0.266 at baseline and 0.039 post-treatment. The enrichment after in vitro size selection was 0.89 times at baseline and 1.08 times post-treatment. The samples collected post-treatment for this patient exhibited a degraded pattern of fragmentation (see fig. S4). The size selection was affected by this pattern, and the number of reads after sequencing was <10M.



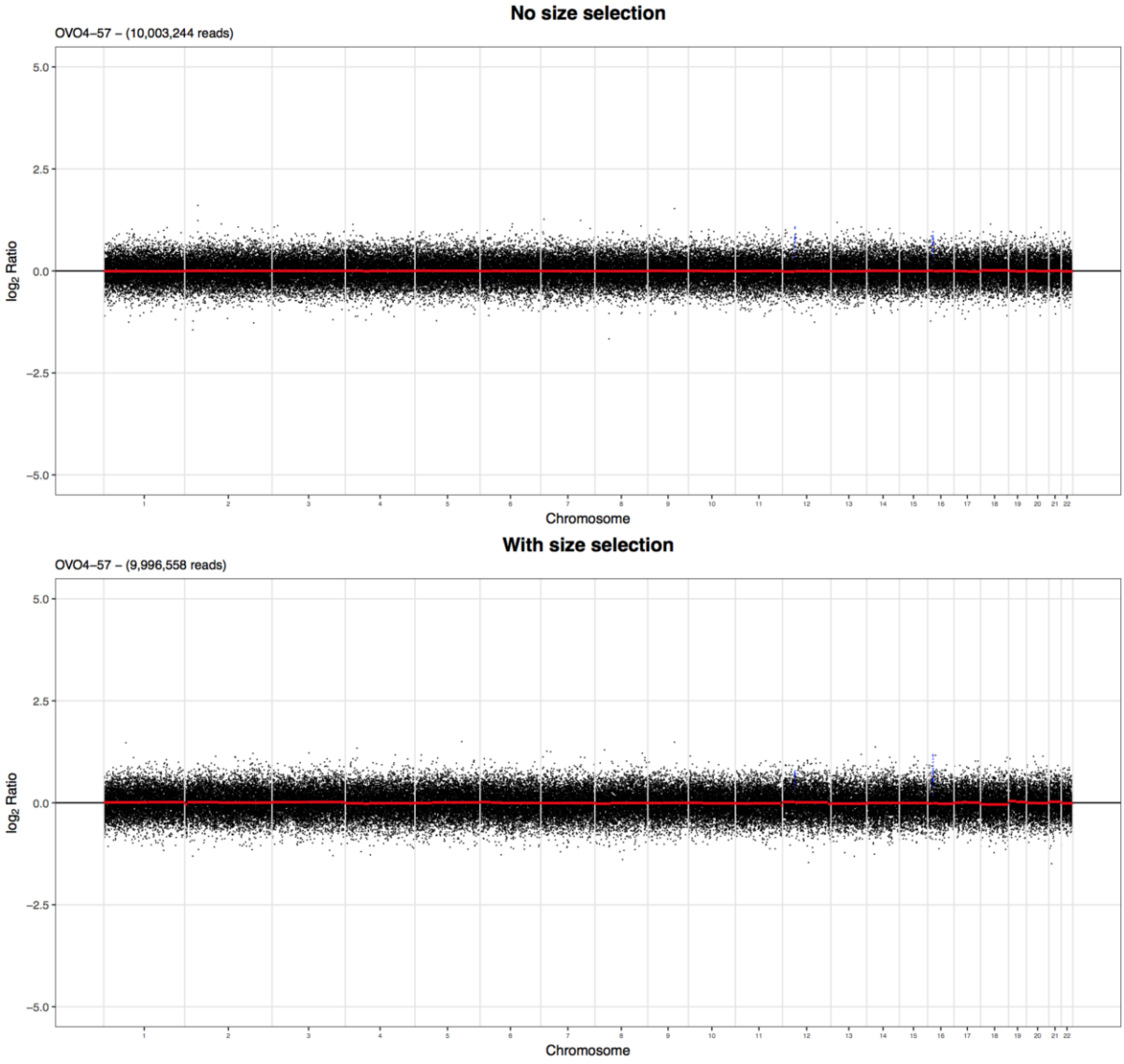
Patient OV04-30: the TP53 MAF determined by TAm-Seq was 0.032. The enrichment after in vitro size selection was 2.73 times.



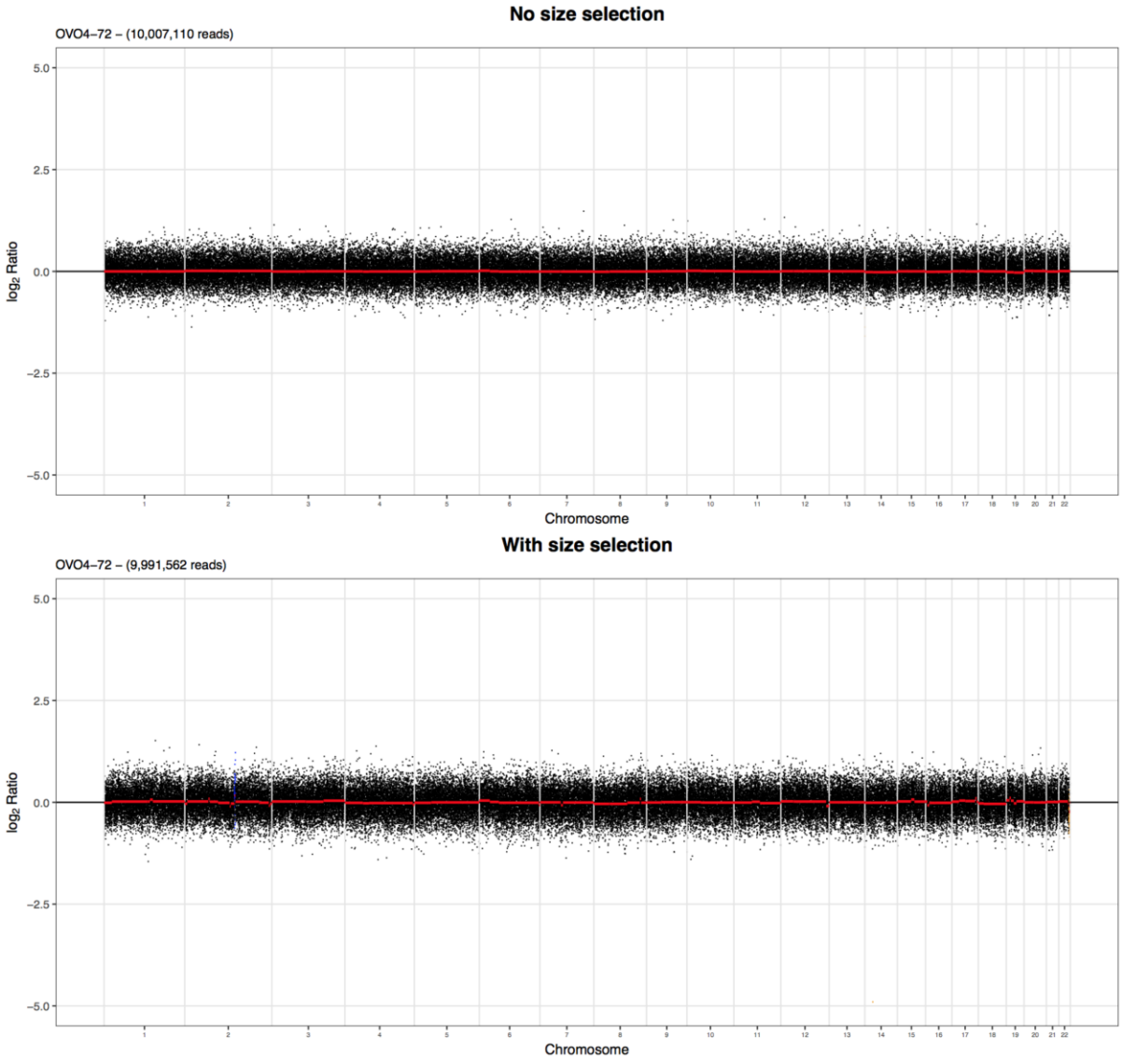
Patient OV04-52: the TP53 MAF determined by TAm-Seq was 0.002. The enrichment after in vitro size selection was 3.48 times.



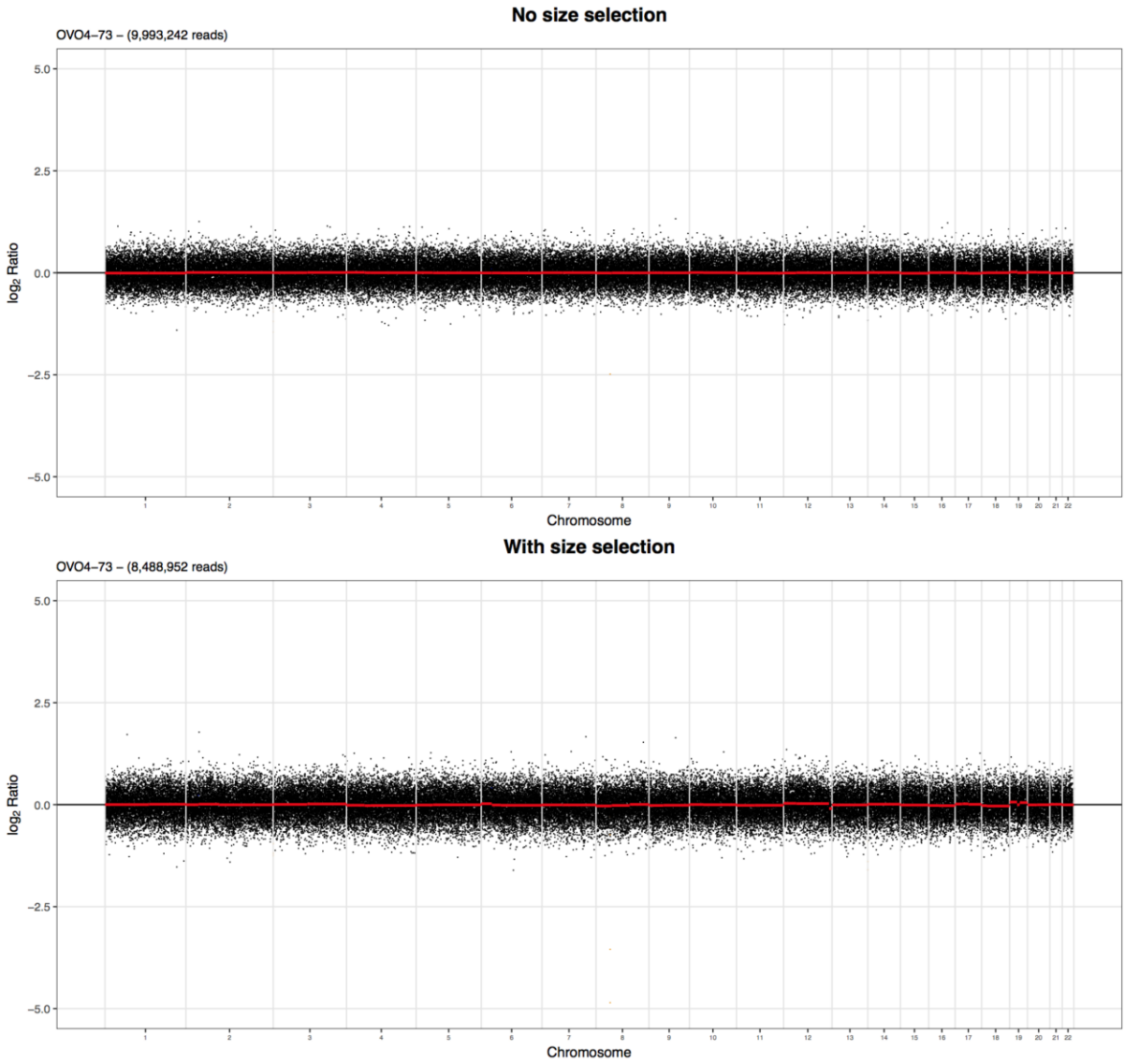
Patient OV04-57: the TP53 MAF determined by TAm-Seq was 0.001. The enrichment after in vitro size selection was 2.67 times.



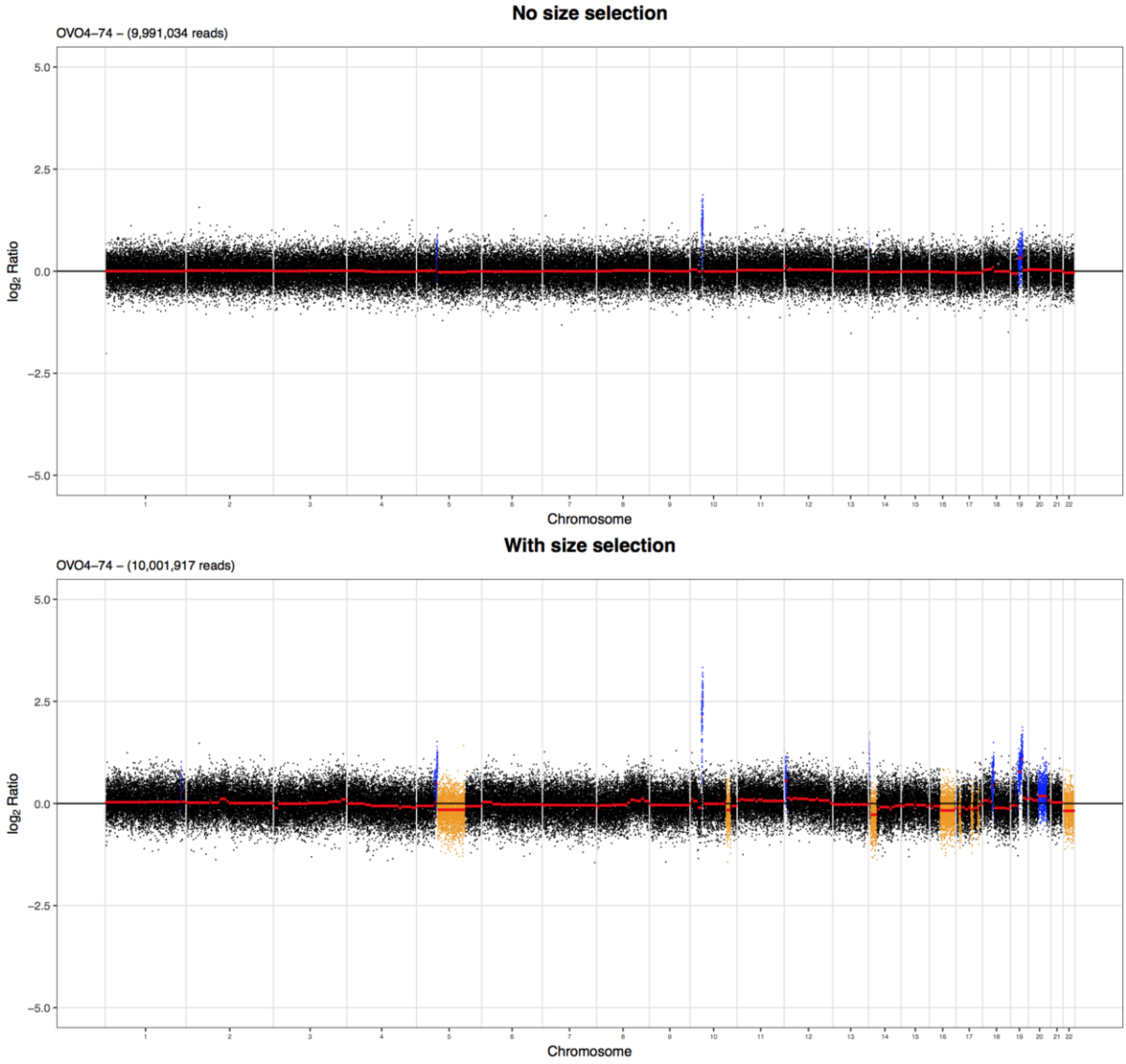
Patient OV04-72: the TP53 MAF determined by TAm-Seq was not detected.



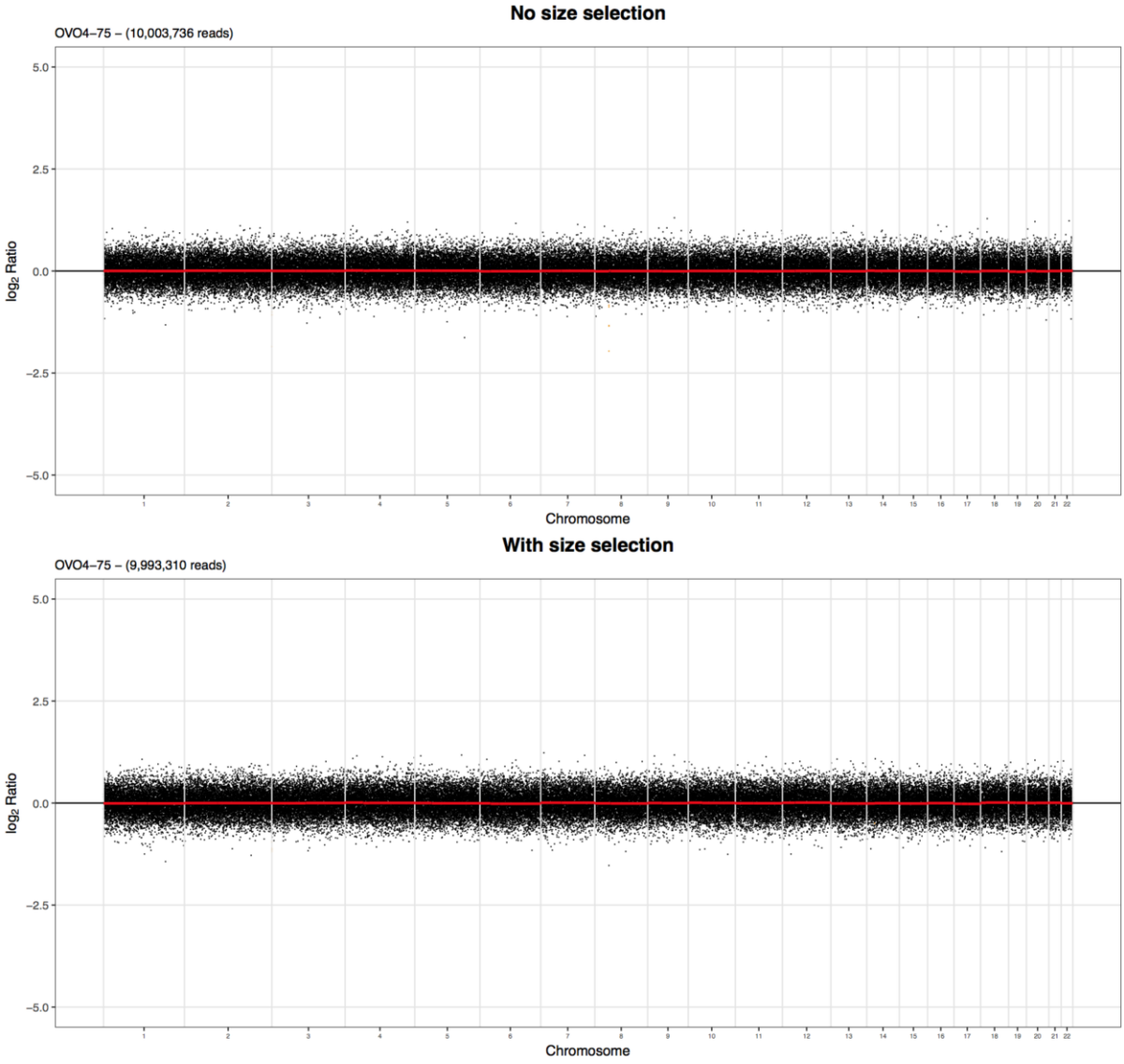
Patient OV04-73: the TP53 MAF determined by TAm-Seq was 0.002. The enrichment after in vitro size selection was 2.52 times.



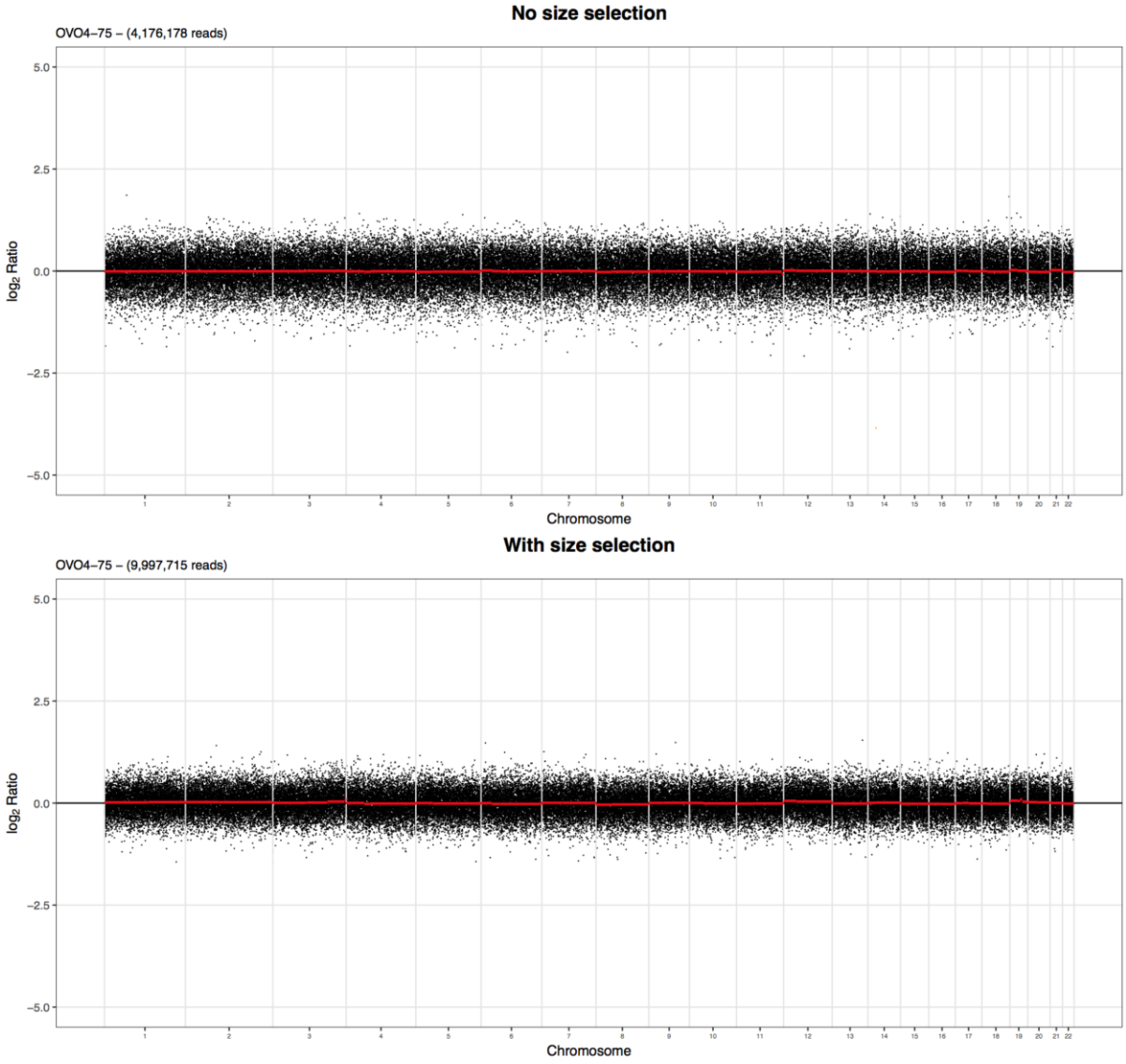
Patient OV04-74: the TP53 MAF determined by TAm-Seq was 0.001. The enrichment after in vitro size selection was 3.86 times.



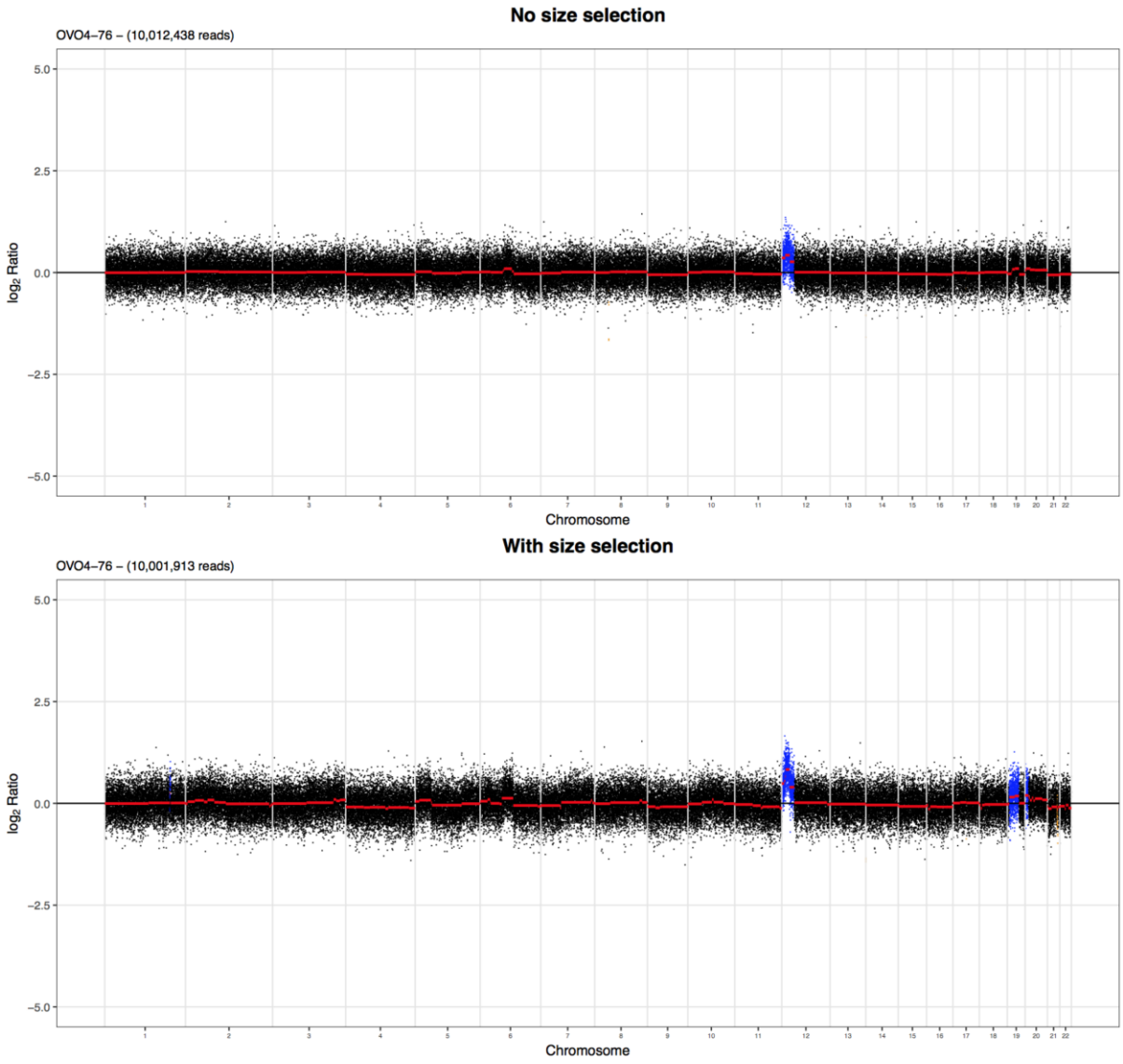
Patient OV04-75: the TP53 MAF determined by TAm-Seq was 0.004. The enrichment after in vitro size selection was 2.46 times.



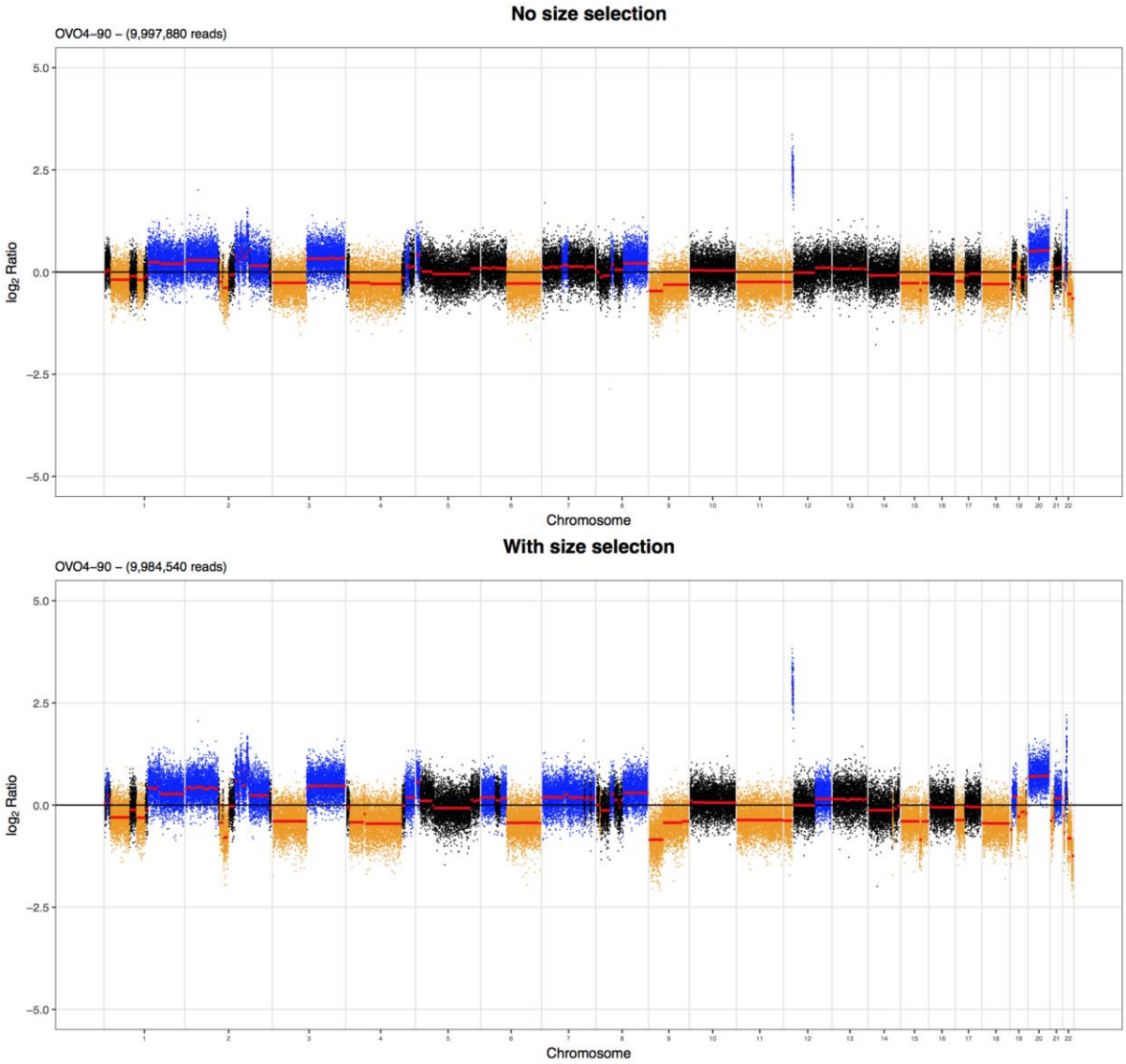
Patient OV04-75-2: the TP53 MAF determined by TAm-Seq was 0.001. The enrichment after in vitro size selection was 2.09 times.



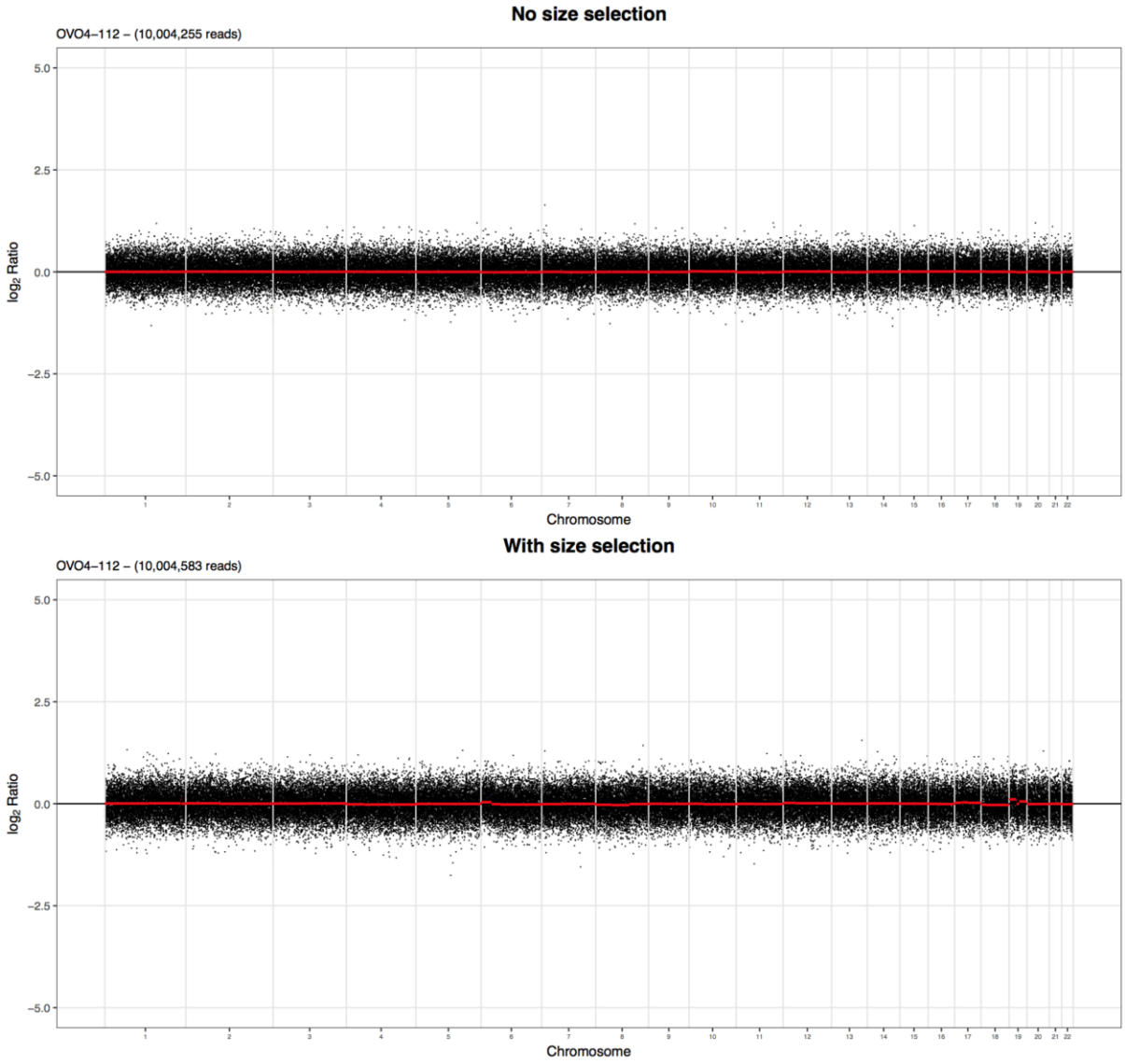
Patient OV04-76: the TP53 MAF determined by TAm-Seq was undetected.



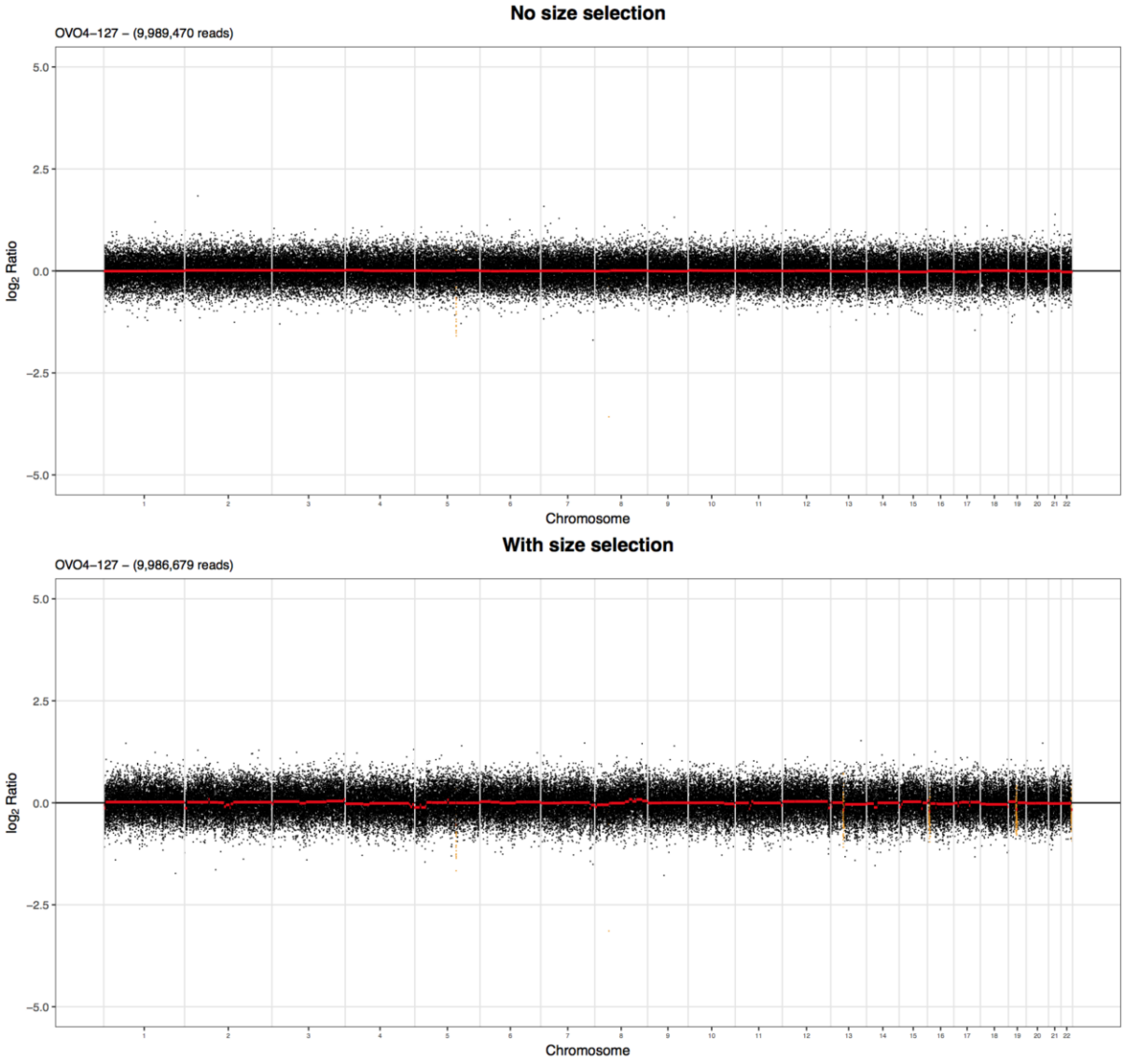
Patient OV04-90: the TP53 MAF determined by TAm-Seq was 0.003. The enrichment after in vitro size selection was 1.39 times.



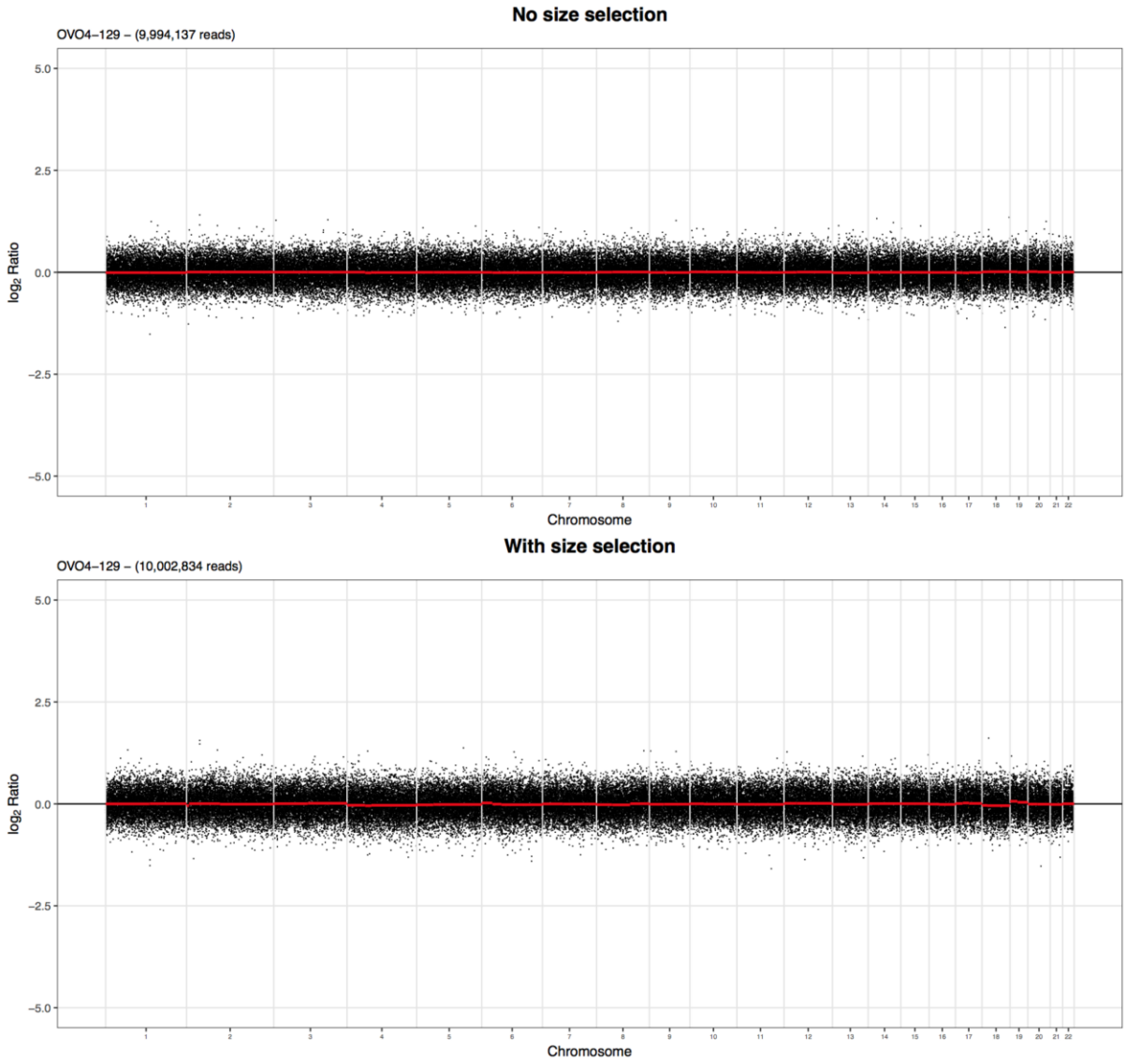
Patient OV04-90: the TP53 MAF determined by TAm-Seq was undetected.



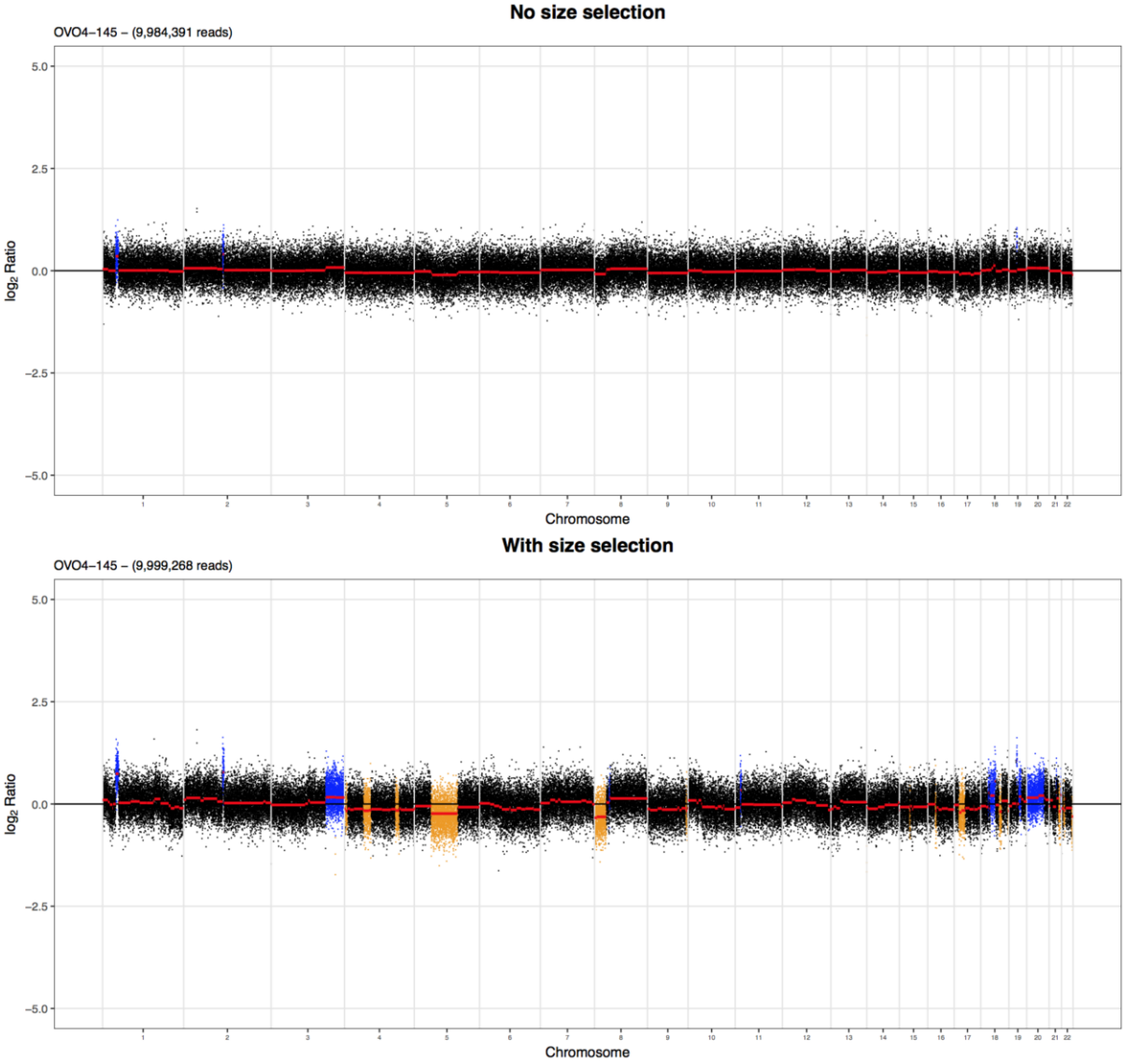
Patient OV04-127: the TP53 MAF determined by TAm-Seq was 0.002. The enrichment after in vitro size selection was 3.01 times.



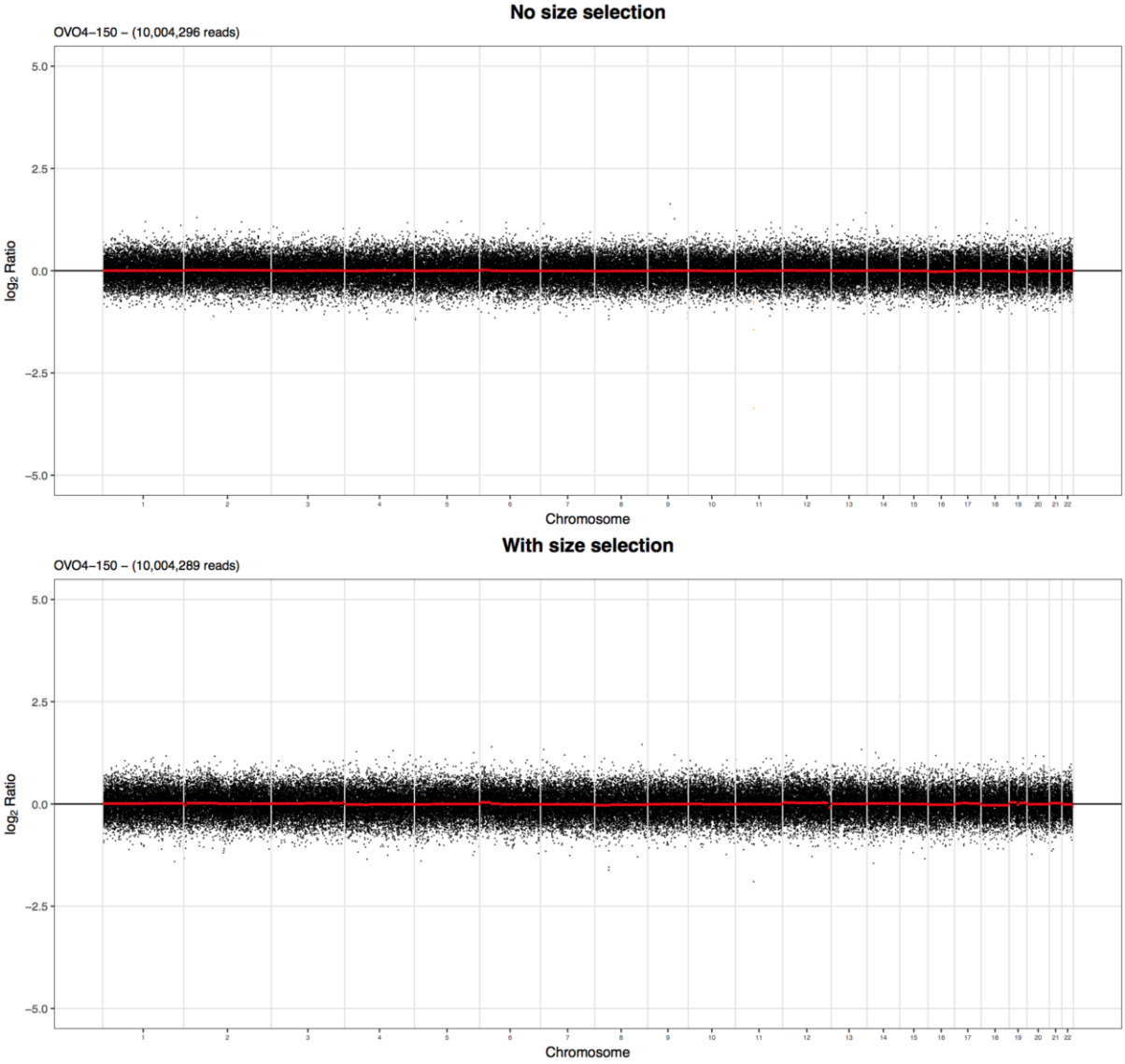
Patient OV04-129: the TP53 MAF determined by TAm-Seq was 0.001. The enrichment after in vitro size selection was 1.97 times.



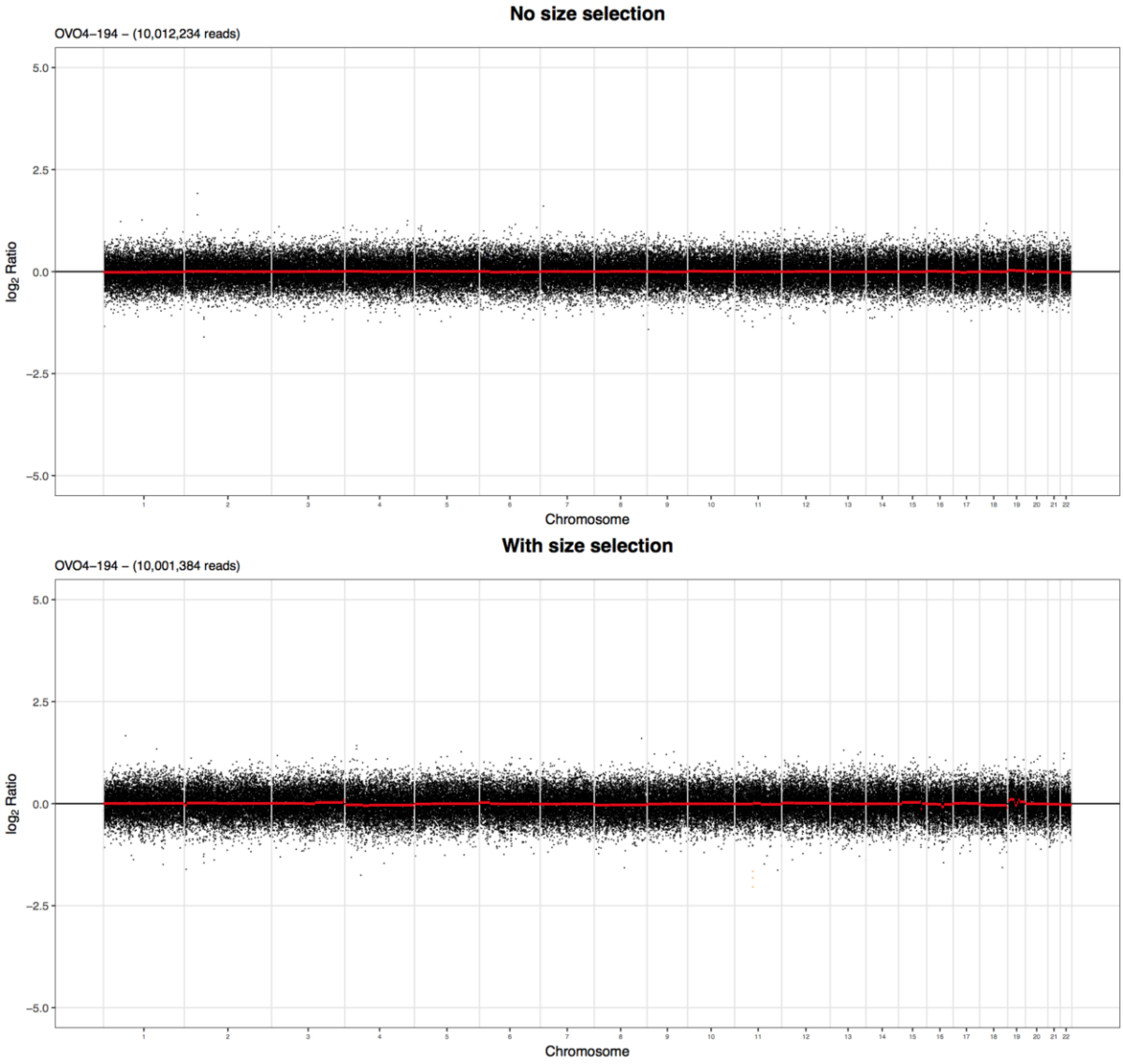
Patient OV04-145: the TP53 MAF determined by TAm-Seq was undetected.



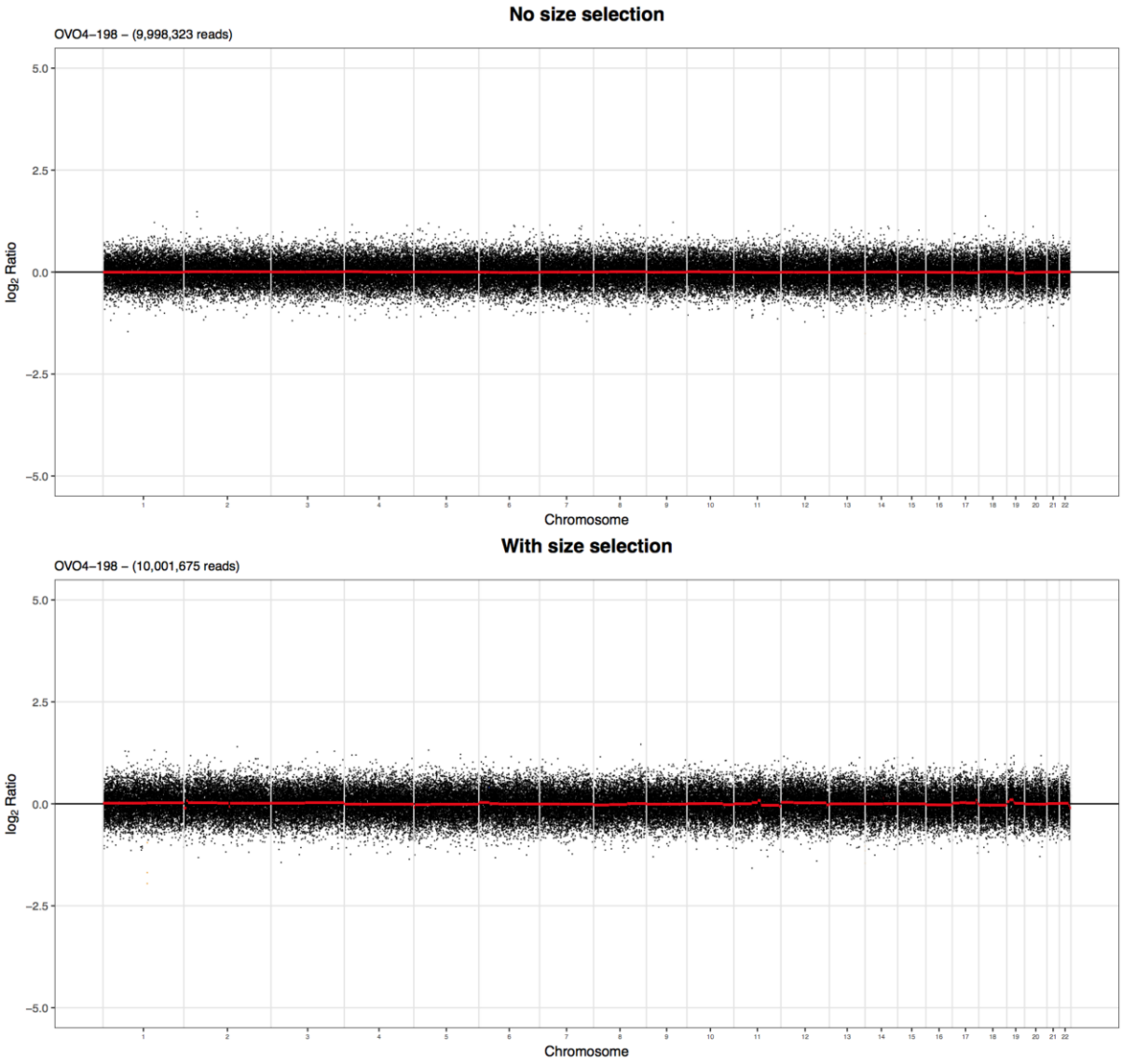
Patient OV04-150: the TP53 MAF determined by TAm-Seq was undetected.



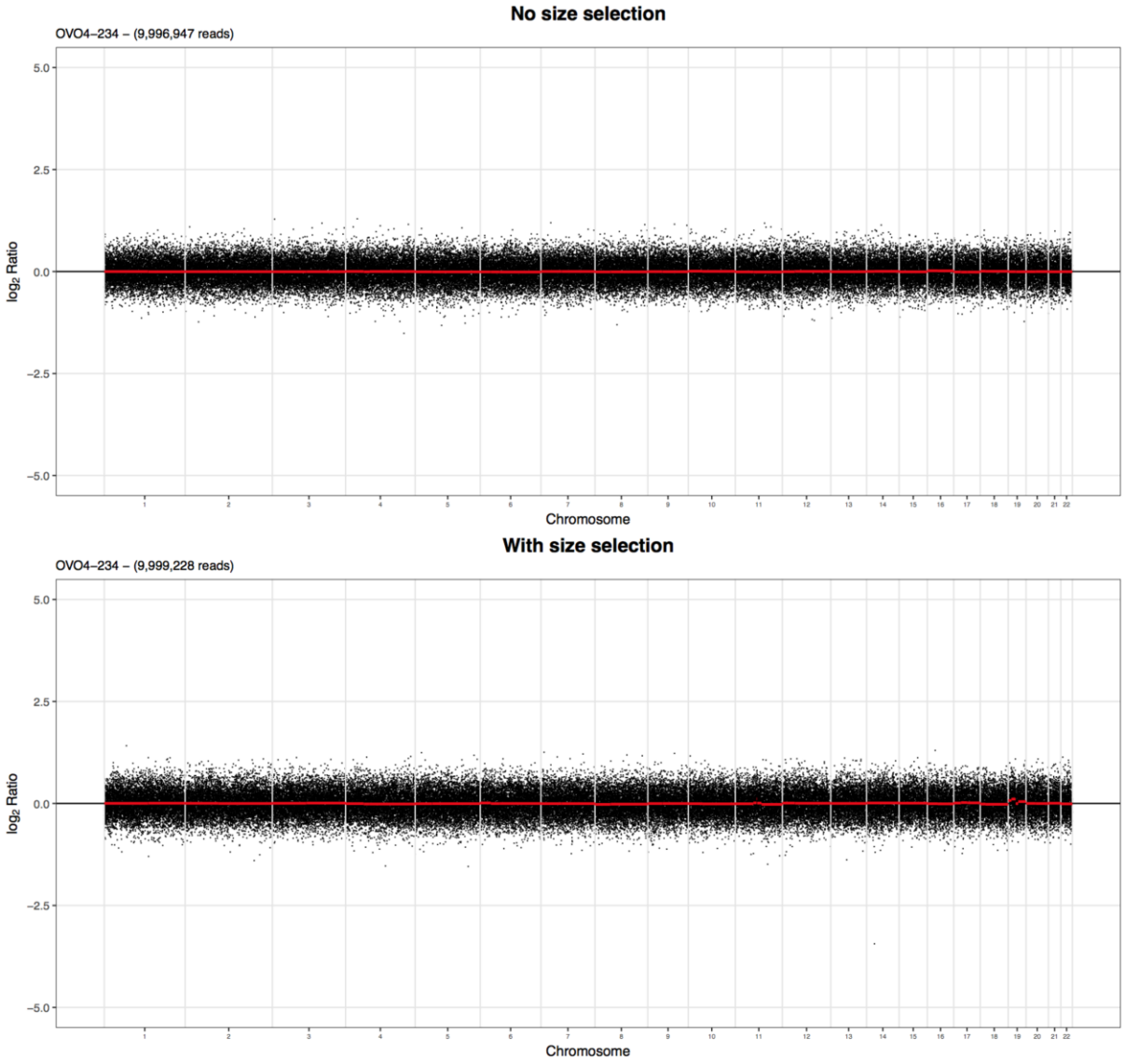
Patient OV04-194: the TP53 MAF determined by TAm-Seq was undetected.



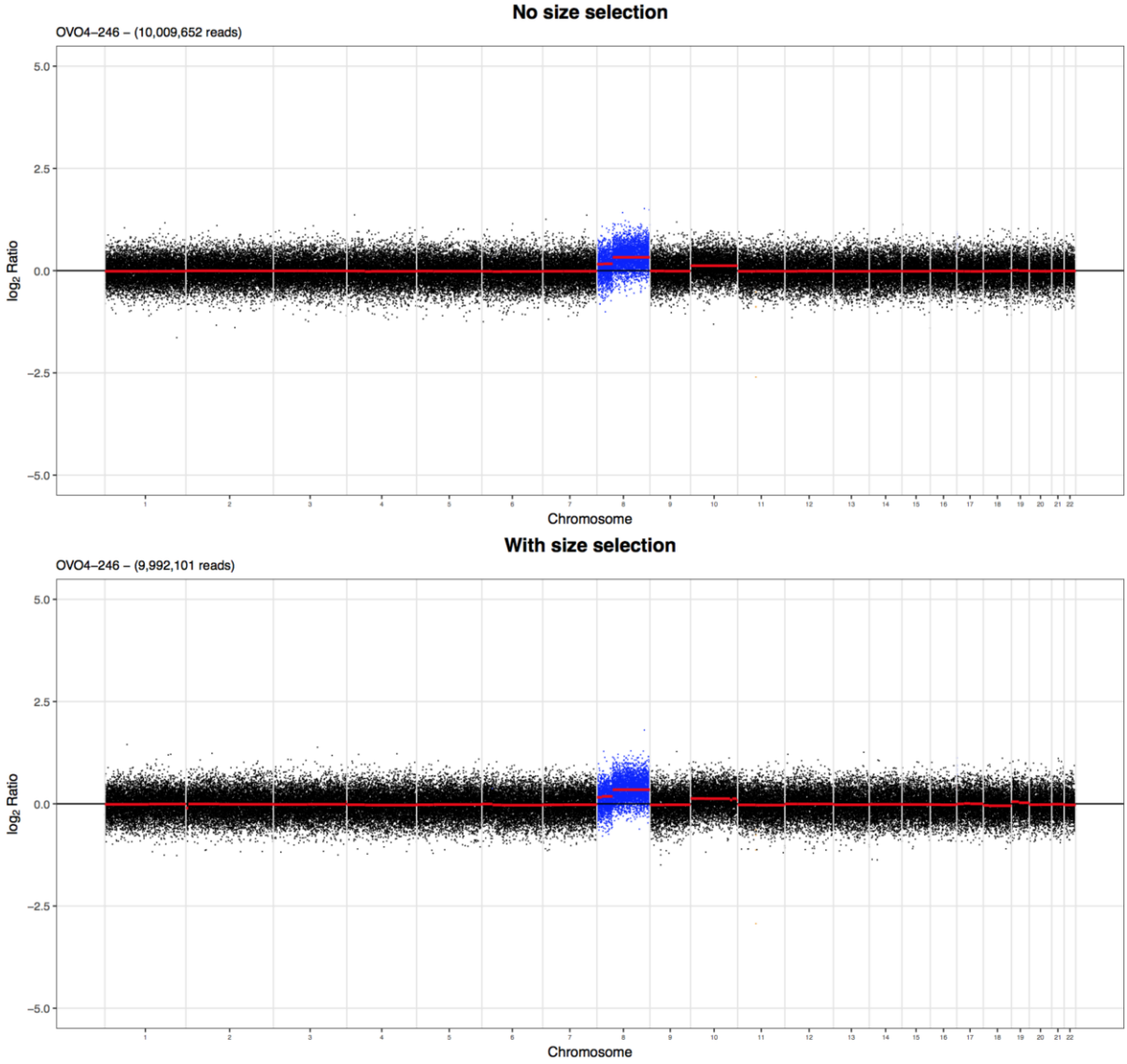
Patient OV04-198: the TP53 MAF determined by TAm-Seq was undetected.



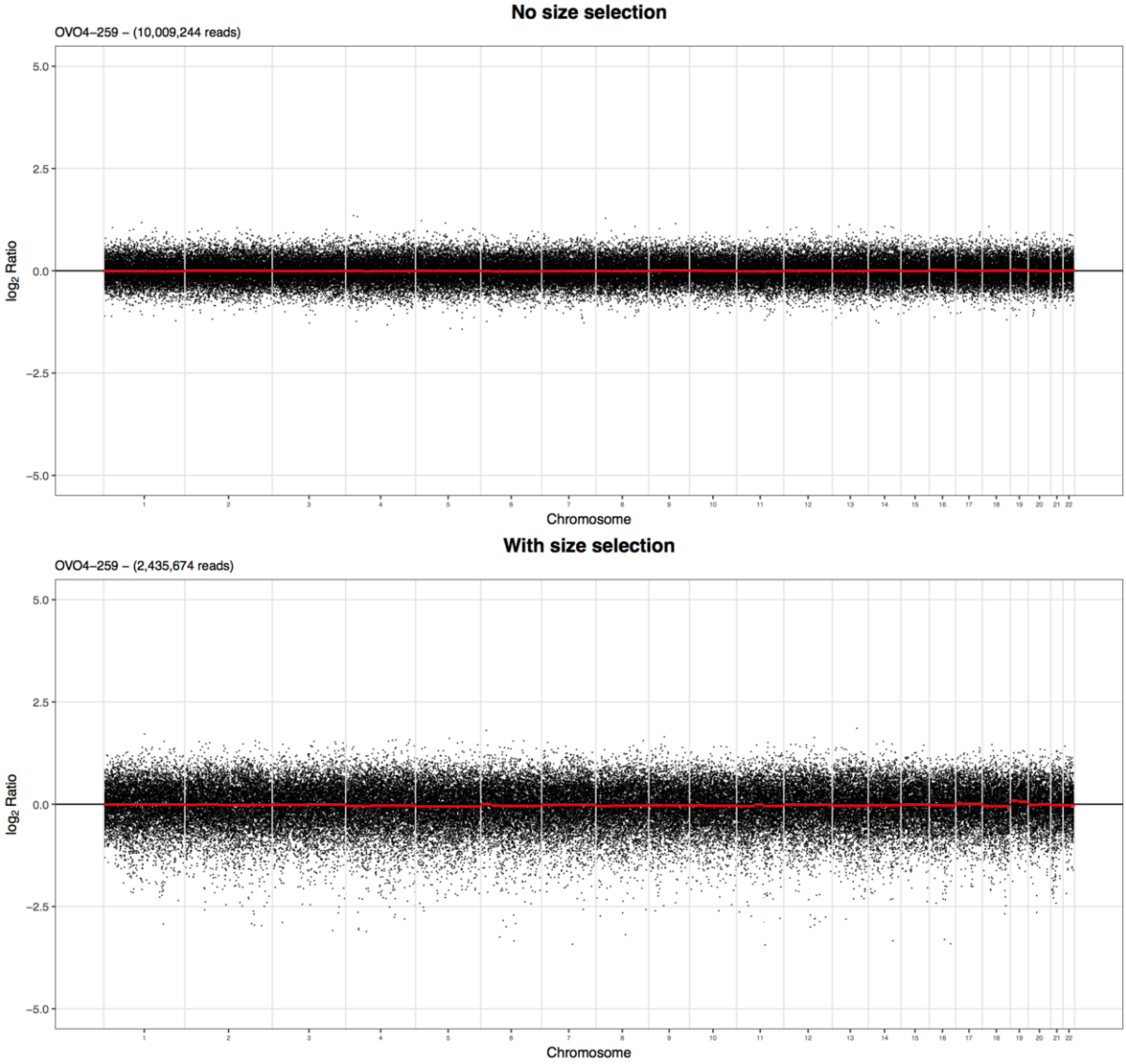
Patient OV04-234: the TP53 MAF determined by TAm-Seq was undetected.



Patient OV04-246: the TP53 MAF determined by TAm-Seq was undetected.



Patient OV04-259: the TP53 MAF determined by TAm-Seq was undetected.



Patient OV04-270: the TP53 MAF determined by TAm-Seq was undetected.

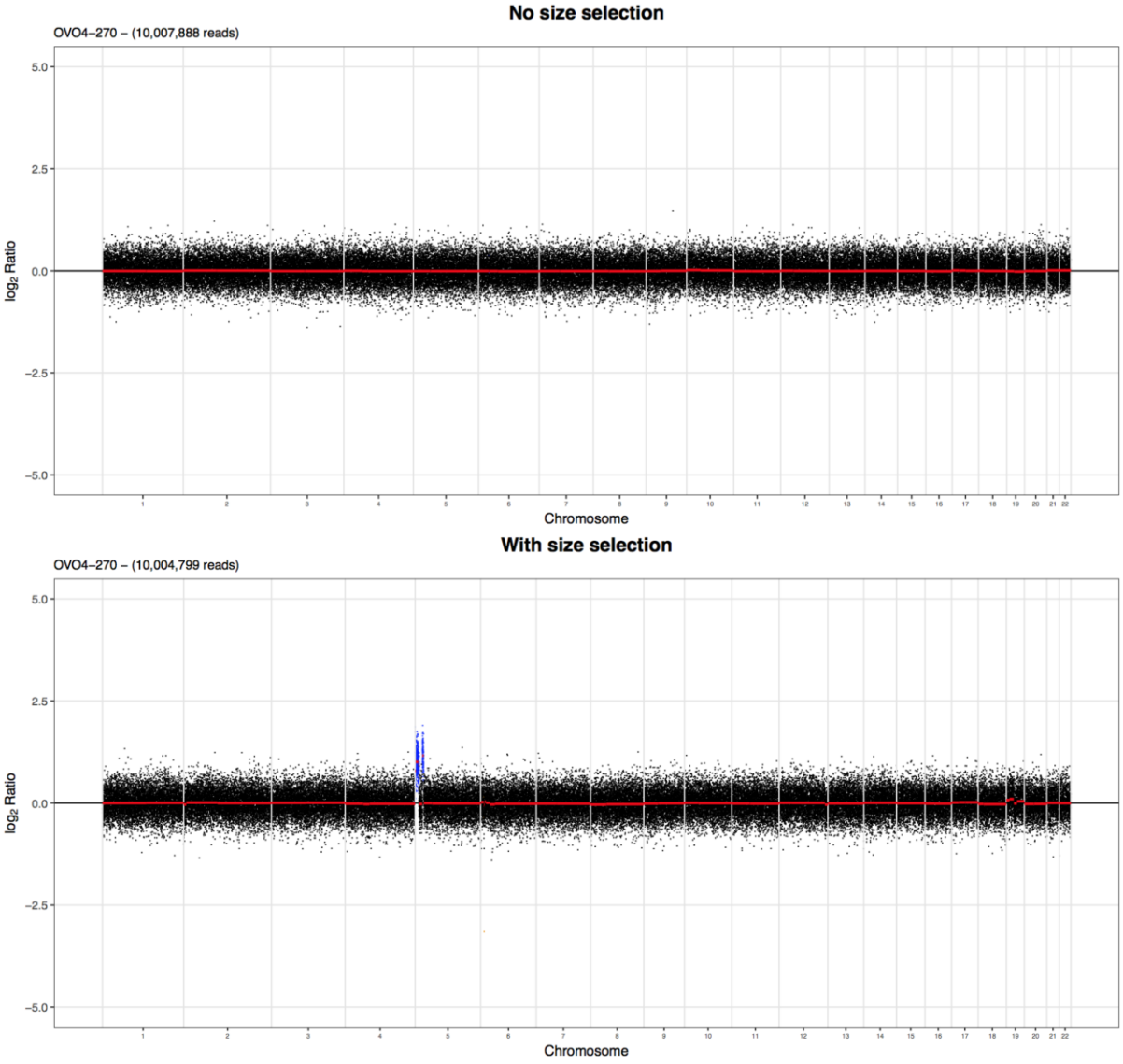


Fig. S8. SCNA analysis of the segmental log_2 ratio determined after sWGS (<0.4× coverage) for plasma samples from patients with ovarian cancer (from the OV04 study). The copy number plots are shown without size selection and with in vitro size selection of the shorter DNA (90-150 bp). Amplifications are shown in blue, deletions in orange, and copy number neutral regions in black. The TP53 MAF value was determined by TAm-Seq before size selection. The enrichment ratio was calculated from sWGS data.

Supplementary figure 9:

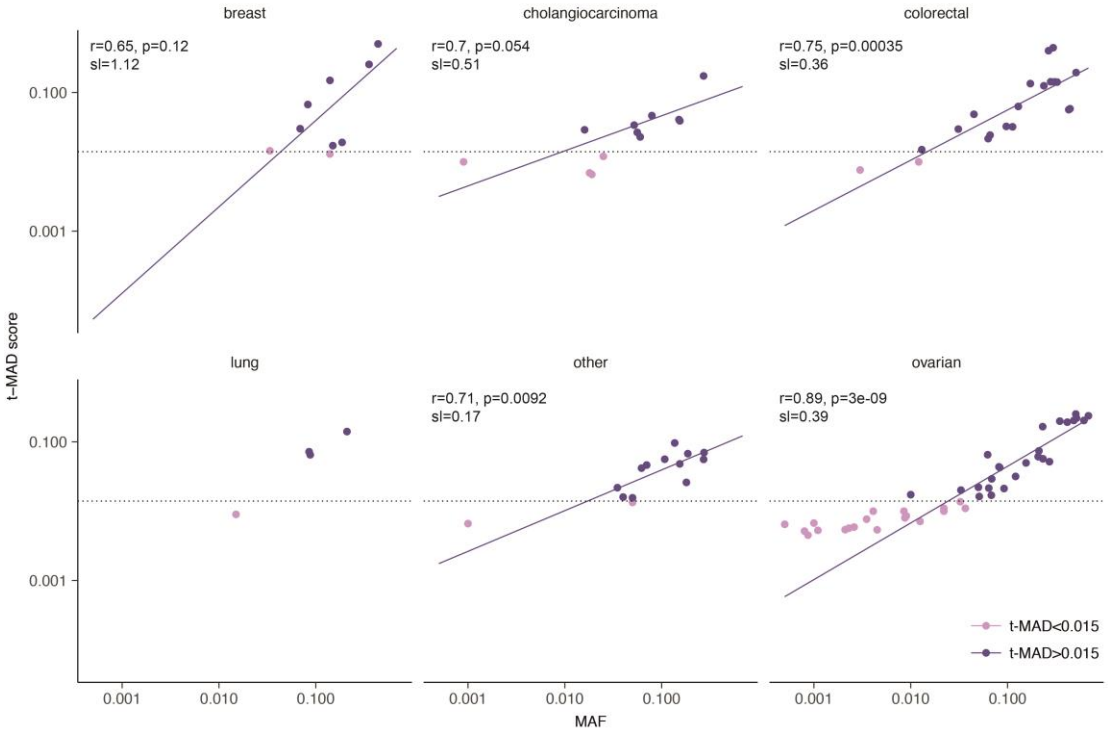


Fig. S9. MAF and t-MAD score compared for different cancer types. Data from ovarian, breast, cholangiocarcinoma, colorectal, and lung cancers are indicated for cases where matched MAF and t-MAD data were available. Other cancer types are grouped in the category “other”. Samples are labeled depending on their t-MAD score, with t-MAD<0.015 colored in light purple, and t-MAD>0.015 colored in dark purple. Pearson correlations, p values, and slopes are indicated when n>5 and t-MAD>0.015.

Supplementary figure 10:

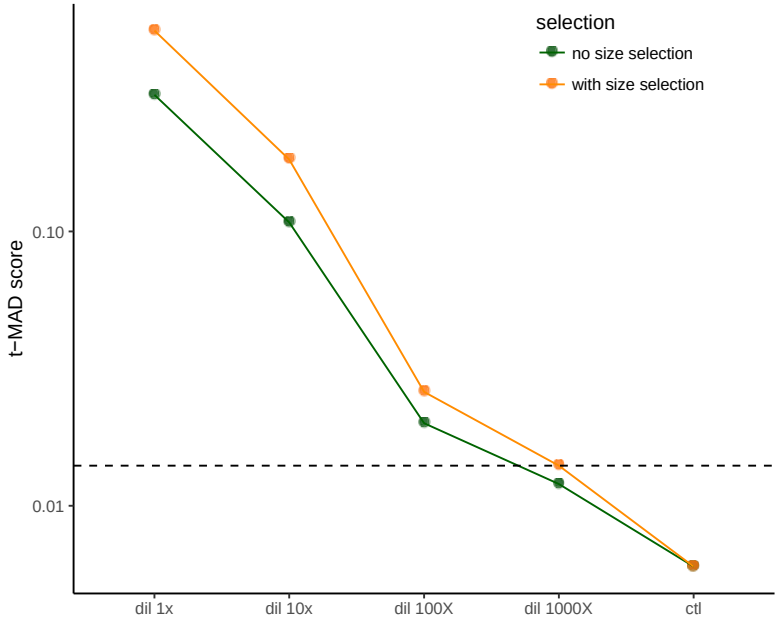


Fig. S10. t-MAD score measured on a plasma DNA dilution series. Plasma DNA from a breast cancer patient was spiked into pooled plasma DNA derived from healthy individuals. This was serially diluted 10-fold to achieve dilutions of 1, 0.1, 0.01, and 0.001 compared to the original. A total of 10 ng of DNA was used for the initial DNA library preparation. The allele fraction for a TP53 mutation of the neat sample was estimated by both WES and TAm-Seq to be ~45.6% and was used as the reference for the dilution. In the dilution series data, the t-MAD score appears to detect SCNA with very low coverage and mutant AF down to ~0.46% AF, or the 100x diluted sample. In addition, the sequencing data have been in silico size selected for the short fragments (90-150 bp), leading to an increase in the t-MAD score (yellow dots and line; green dots and line show t-MAD without size selection).

Supplementary figure 11:

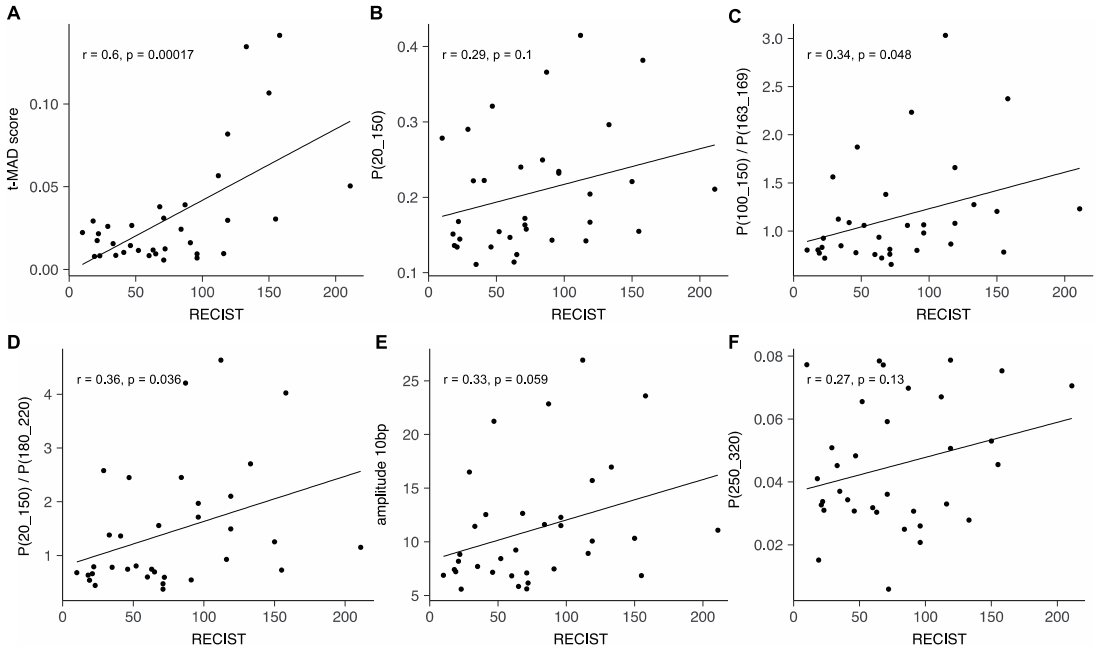


Fig. S11. t-MAD scores and fragmentation features compared to tumor volume. The available RECIST volume (in mm) determined by CT scan was compared to the t-MAD score and fragmentation features. The RECIST volume was compared to the t-MAD score (A), the proportion of fragments between 20 and 150 bp (B), the ratio of the proportion of fragments between 100-150 bp and the proportion of fragments between 163-169 bp (C), the ratio of the proportion of fragments between 20-150 bp and the proportion of fragments between 180-220 bp (D), the amplitude of the 10 bp peaks and valleys (E), and the proportion of fragments between 250-350 bp (F). Correlation and p values are shown for each comparison.

Supplementary figure 12:

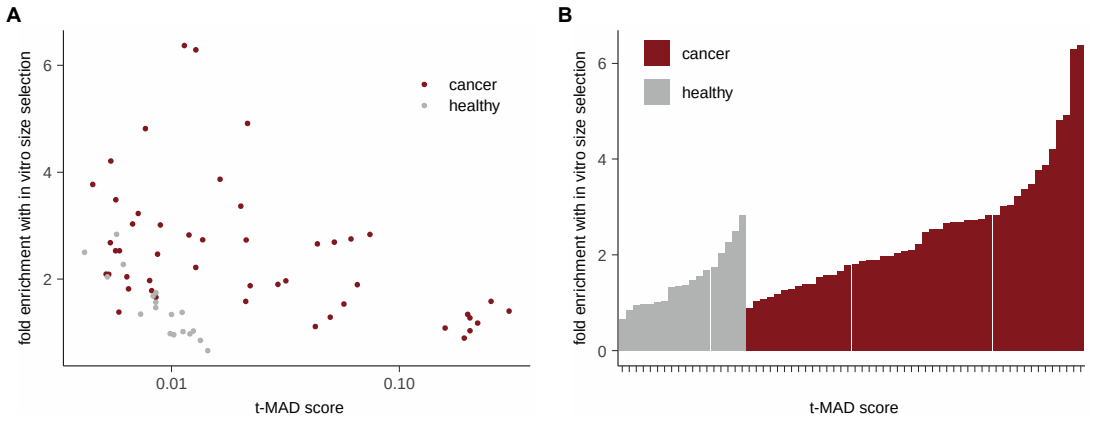


Fig. S12. Changes to t-MAD after in vitro size selection. The t-MAD score was determined after in vitro size selection for 48 plasma samples collected from 35 ovarian patients and from 18 healthy controls. A) Samples with higher t-MAD scores before size selection had less enrichment (Pearson correlation, r=-0.49, p<0.001). B) The t-MAD score determined from the sWGS with in vitro size selection was higher than without size selection for nearly all samples, with a median increase of 2.1-fold.

Supplementary figure 13:

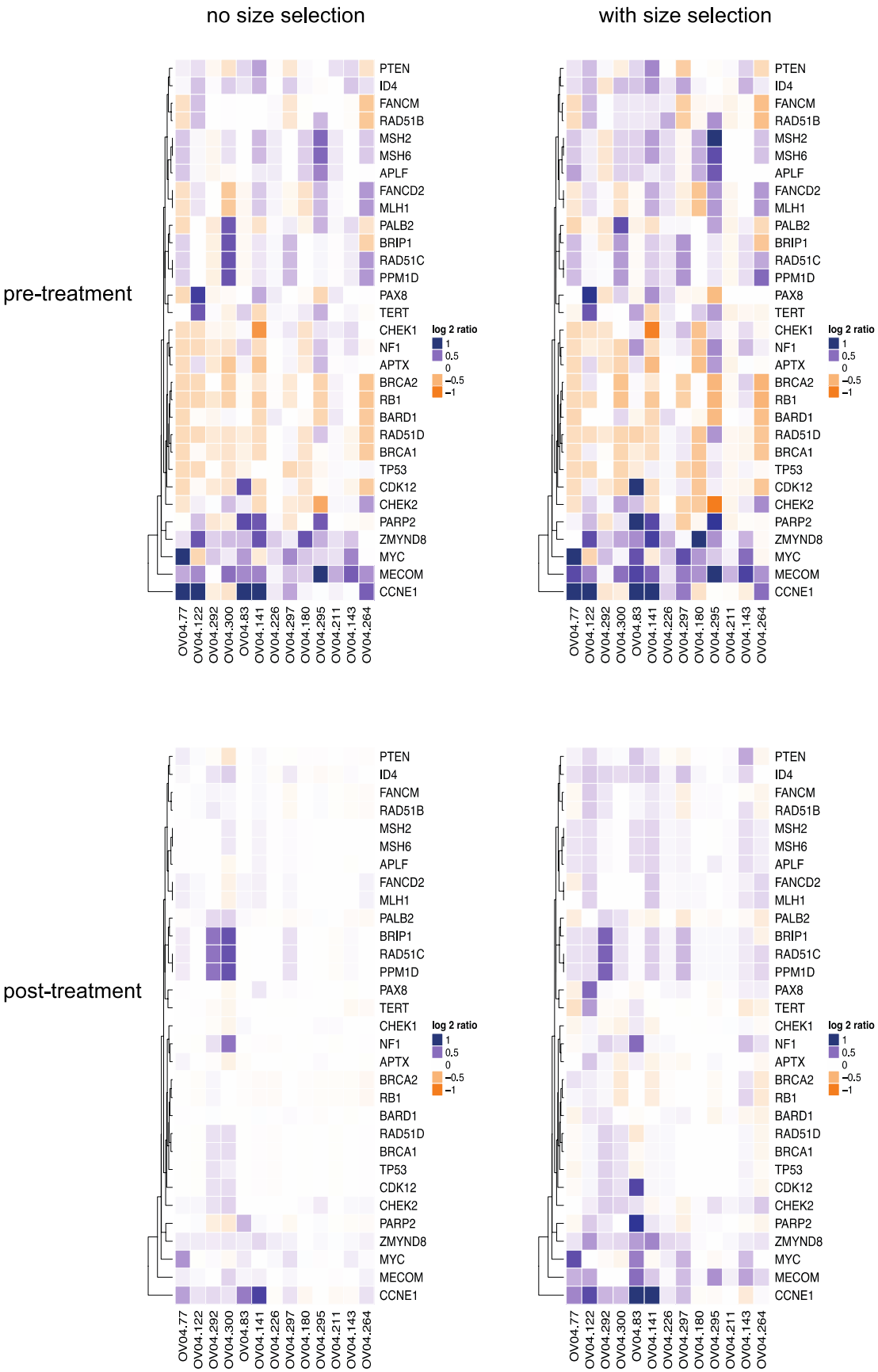


Fig. S13. SCNA analysis in cfDNA from plasma samples collected at baseline and after treatment for 13 patients with HGSOC. Segmental $\log_2$ ratio of the SCNA was determined by sWGS across a list of 29 genes frequently mutated in recurrent ovarian cancer. The $\log_2$ ratios are shown for the samples without size selection and with in vitro size selection of the shorter DNA (90-150 bp).

Supplementary figure 14:

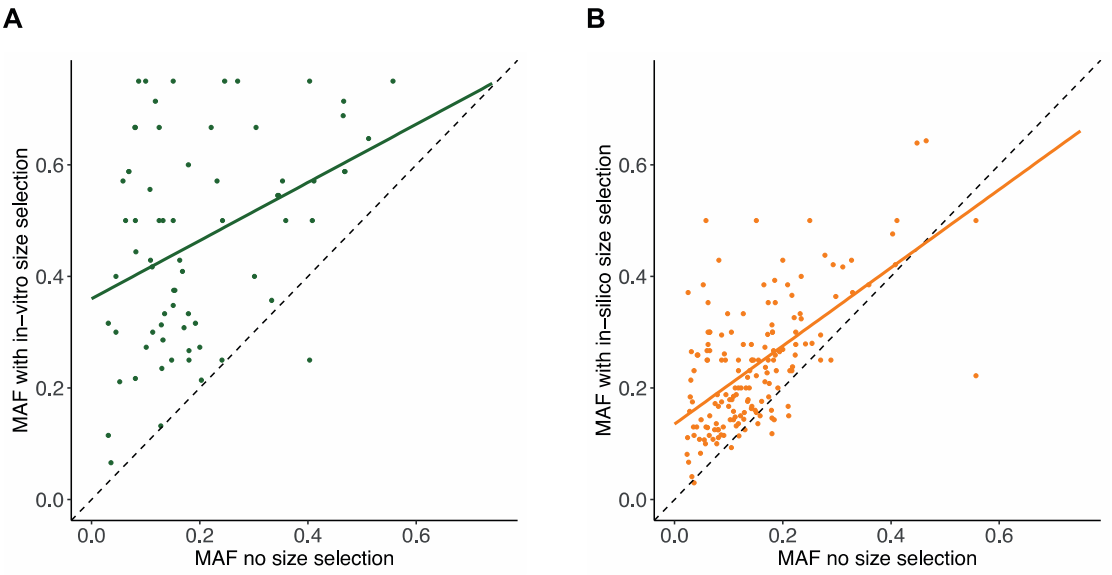


Fig. S14. MAF for SNVs called by WES with and without size selection. Mutant allelic fractions (MAFs) for single nucleotide variants (SNVs) were called by WES analysis of plasma DNA samples from patients with HGSOC, without size selection and with in vitro size selection. A) The MAF determined by WES with in vitro size selection (vertical) was higher than without in vitro size selection (horizontal) for most of the mutations detected from the plasma samples of 6 HGSOC patients. B) A lower rate of enrichment was also observed in the same samples after in silico size selection of the WES data that were generated without in vitro size selection.

Supplementary figure 15:

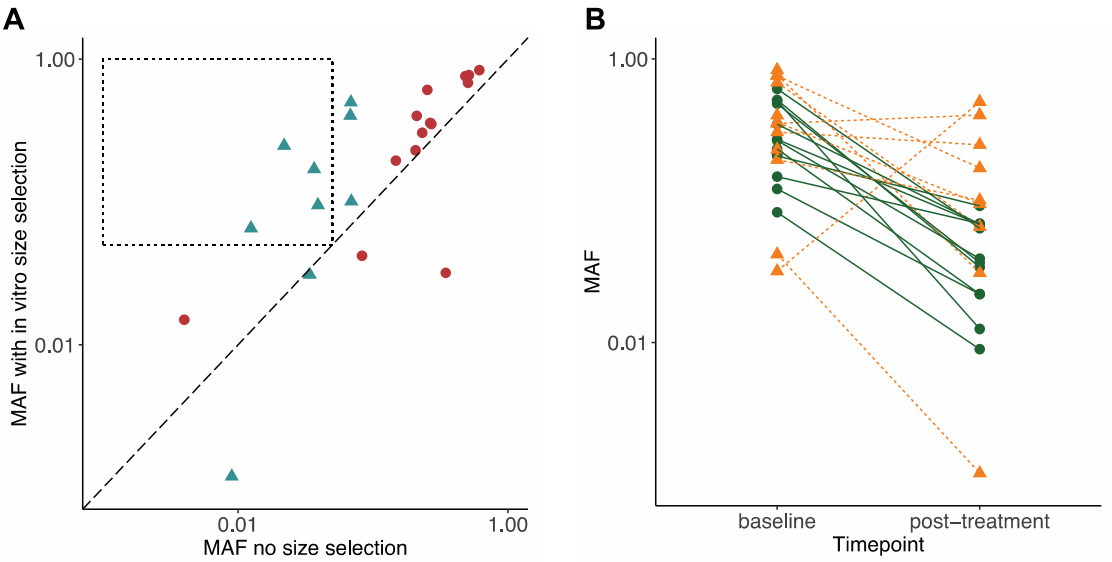


Fig. S15. TAm-Seq before and after in vitro size selection. A, The MAF for TP53 mutations determined by TAm-Seq without and with in vitro size selection, for samples collected at baseline (red circles) and after initiation of treatment (blue triangles) from 13 patients with HGSOC. The MAF was higher after size selection for most samples, especially for samples collected after initiation of treatment (blue triangles). The dotted area highlights samples which initially had low MAF (<5%), where methods such as whole-exome sequencing (at sequencing depth of ~100x) would not be effective, and where in vitro size selection enriched the MAF to >5% and therefore made it accessible for wide-scale analysis. **B,** Comparison of the MAF detected by TAm-Seq before treatment and after initiation of treatment, with in vitro size selection (yellow triangles) and without size selection (green circles).

Supplementary figure 16:

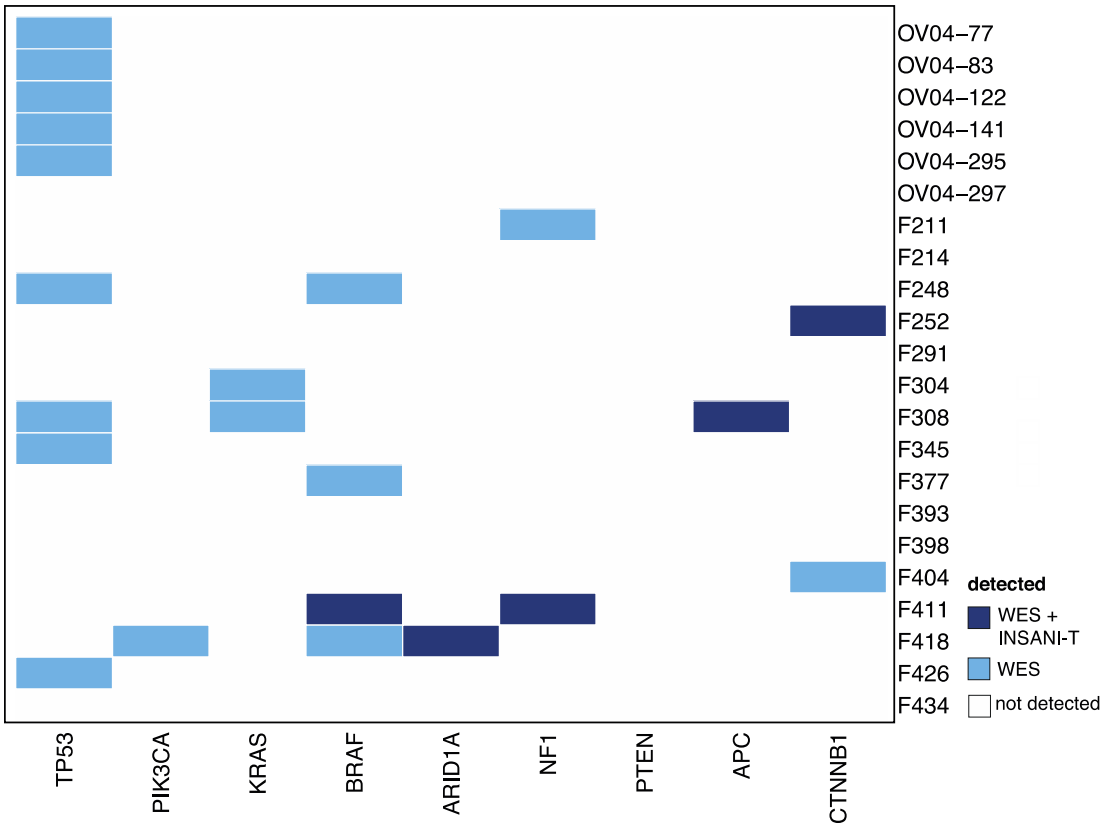


Fig. S16. Mutations in clinically relevant genes detected by WES with and without in silico size selection. Mutations were analyzed for 9 genes of clinical importance. Here, we analyzed all the plasma samples submitted to WES (from 6 patients with HGSOC and from 16 patients with different cancer types treated in early-phase clinical trials). The figure shows mutations called without size selection and mutations called by WES only after in silico size selection.

Supplementary figure 17:

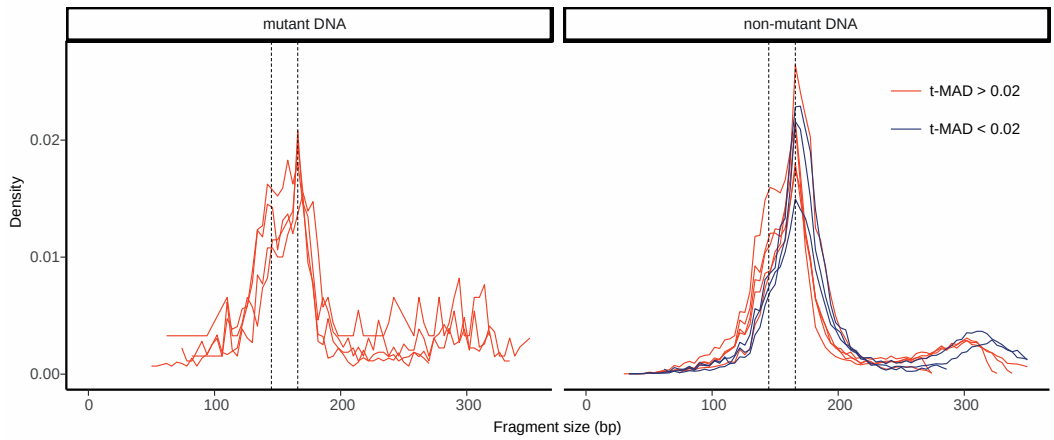


Fig. S17. Size distribution of nonmutant DNA and ctDNA concentration. Size distribution of mutant and non-mutant DNA obtained from the personalized sequencing (as described in **Fig. 2D**). We sub-selected 4 patients from this figure with more than 200 mutant reads. 3 patients with no mutant DNA were added on the right panel for comparison. The loci selected corresponded to variants identified by WES of the tumor tissue DNA. The left panel exhibits the size distribution of mutant DNA, and the right panel the size distribution of the corresponding non-mutant DNA. The color represents the t-MAD value for these patients (red for t-MAD score > 0.02 and blue for t-MAD score <0.02). The size distribution of mutant reads confirms enrichment in the size range 90-150 bp. The non-mutant reads exhibited a lower enrichment in the size range 90-150 bp, but varied between patients.

Supplementary figure 18:

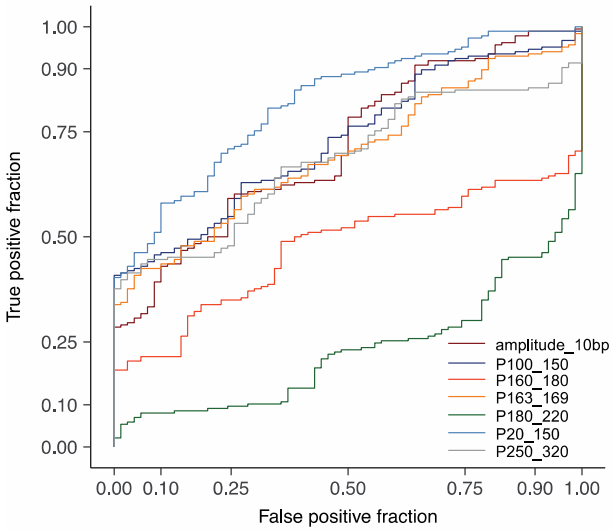


Fig. S18. ROC curve for individual fragmentation features in high ctDNA cancers versus controls. ROC analysis comparing the classification of plasma samples of high ctDNA cancer patients (n=191) and plasma samples from healthy controls (n=65) using 7 cfDNA fragmentation features. The proportion of fragments between 20 and 150 bp exhibited the highest AUC (AUC=0.819).

Supplementary figure 19:

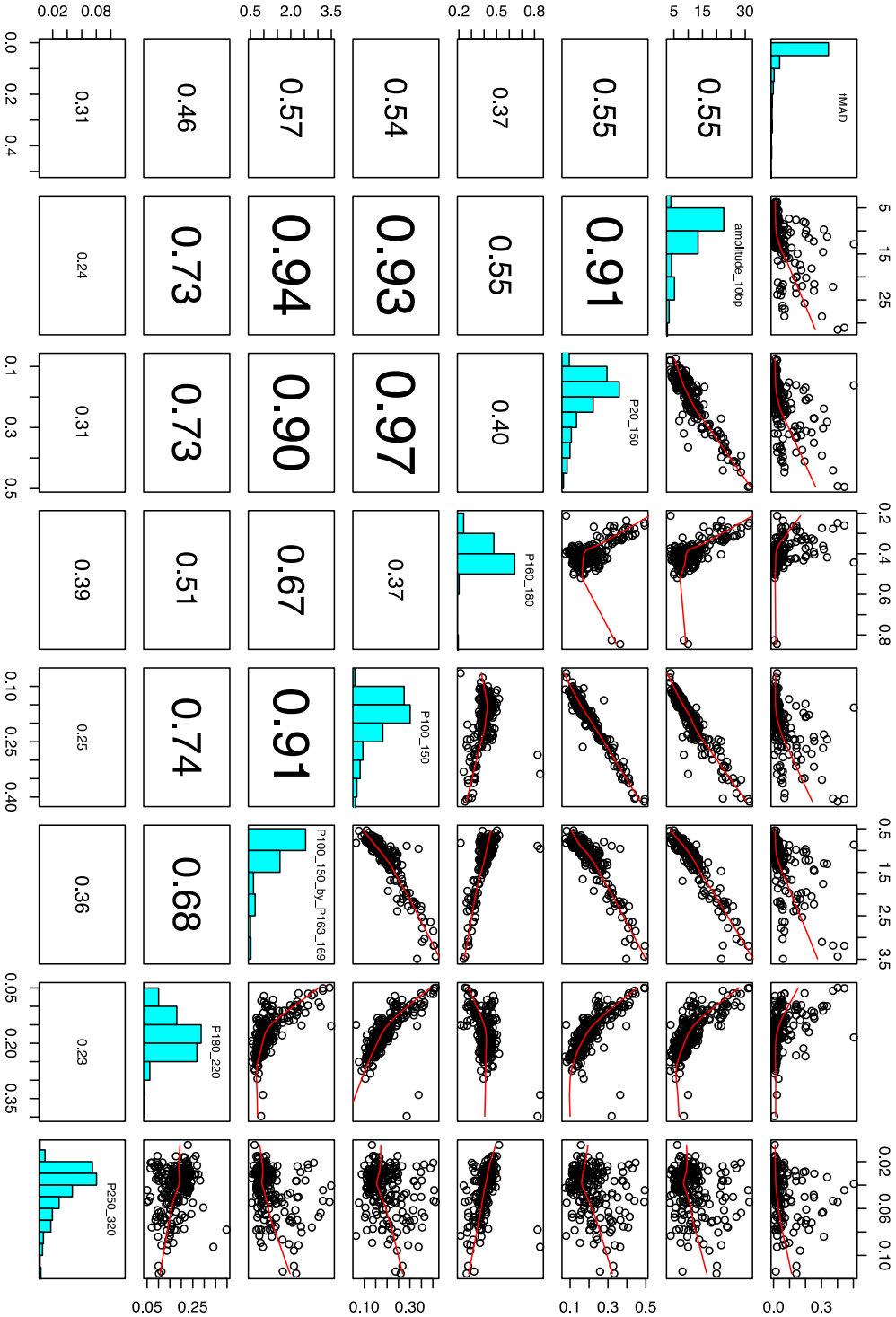


Fig. S19. t-MAD score compared with seven fragmentation features. The fragmentation features and t-MAD were determined by sWGS from the 147 plasma samples from cancer patients included in the training and validation

datasets of the classifier models. Dot-plots are shown in the panels on the top right part of the matrix of panels, and the correlation scores for each comparison are displayed in the corresponding panels on the bottom left part of the matrix. The panels along the diagonal illustrate the distribution of values for each parameter.

Supplementary figure 20:

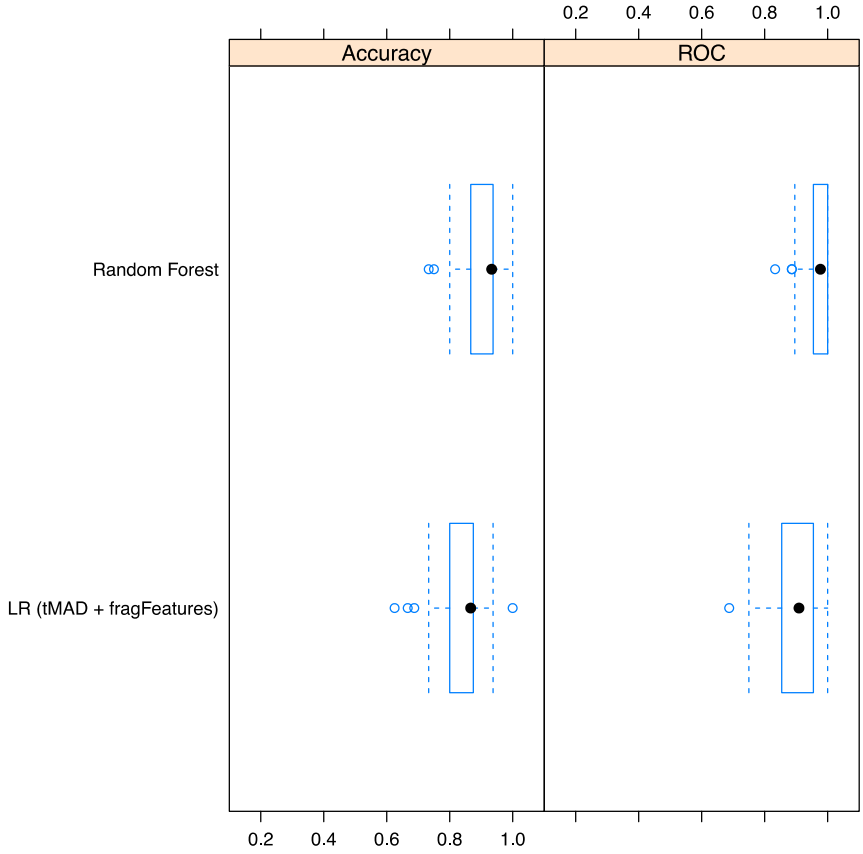


Fig. S20. Performance metrics for the two algorithms, LR and RF. The logistic regression and RF were both run with t-MAD score and the fragmentation features, on training set data from sWGS (n=153; 114 cancer samples and 39 healthy controls). The median ROC score and accuracy values are displayed for each model, as is the 0.95 confidence level.

Supplementary figure 21:

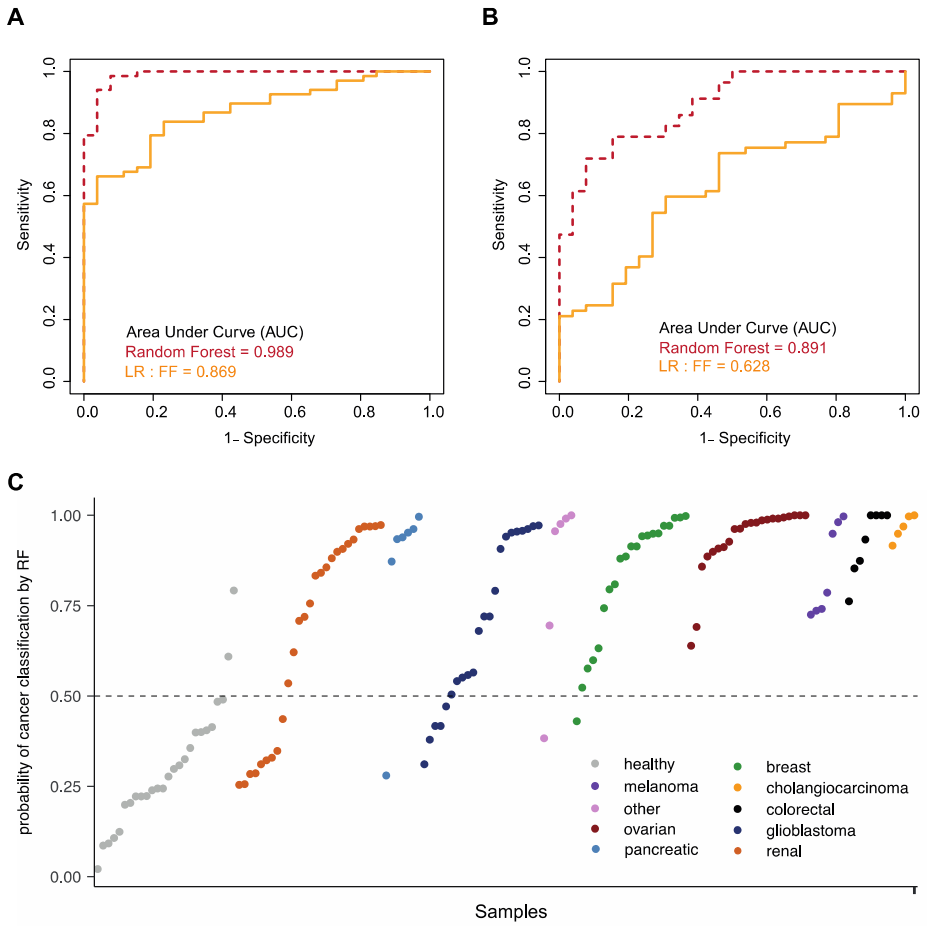


Fig. S21. LR and RF models using the fragmentation features without t-MAD. A) ROC curves from the first validation sample set (cancer=68, healthy=26) for 2 classifiers built on the pan-cancer training cohort (cancer=114, healthy=39). The orange curve represents the ROC for the LR model trained only with the fragmentation features without t-MAD, and the dashed red curve shows the result for a random forest classifier trained on the combination of the best 3 predictive fragmentation features (amplitude_10bp, P(160-180), and P(250-320). B) ROC curves from the second validation sample set (cancer=57, healthy=26) for the same 2 classifiers as in (A). C) The probability of classification as cancer with the RF model for the validation datasets shown in A and B. Samples are ranked by cancer type and by probability of classification as cancer. The dashed horizontal line represents 50% probability.

Supplementary table legends:

Supplementary table 1:

Table S1. Summary table of the patients and samples included in this study. Patients and samples included in this study, listing the DNA extraction type, plasma collection time point, and cancer type for the 344 plasma samples from 200 cancer patients and the 65 healthy control samples included in the study.

Supplementary table 2:

Table S2. Values for nine fragmentation features determined from sWGS data for the samples included in the study. For each sample, the following features were calculated from sWGS data: the proportion (P) of fragments between 20 bp and 150 bp [P(20-150)], P(160-180), P(20-150)/P(160-180), P(100-150), P(100-150)/P(163-169), P(180-220), P(250-320), P(20-150)/P(180-220). The amplitude of the periodic 10 bp oscillations was also quantified as described in the Materials and Methods section.

Supplementary table 3:

Table S3. t-MAD score for the 48 plasma samples of the OV04 cohort before and after in vitro size selection. MAF value for a TP53 mutation (different for each patient) is also indicated, measured by TAm-Seq without size selection.

Supplementary table 4:

Table S4. Log_2 of the signal ratio observed by sWGS of the plasma samples from the OV04 cohort. The ratios are determined from sWGS data of plasma samples from 35 patients, across a list of genomic positions corresponding to the 38 most frequent genes of interest in ovarian cancer as

defined by cBioportal and the catalogue of somatic mutations in cancer (COSMIC database).

Supplementary table 5:

Table S5. Mutations called by WES of six patients selected from the OV04 cohort. Mutations were called by Mutect2 with subsequent stringent filtering (as described in Materials and Methods). The table lists the genomic position and base change, as well as the reference and the detected (alternate) base (ref and alt, respectively).

Supplementary table 6:

Table S6. Mutations called by WES data of the plasma samples from 16 patients from the CoPPO cohort. Mutations were called by Mutect2 and subjected to stringent filtering (as described in Materials and Methods). The table details the genomic position and base change, as well as the reference and the detected (alternate) base (ref and alt, respectively), and the mutation called after in silico size selection.