Shallow-depth sequencing of cell-free DNA for Hodgkin and diffuse large B-cell lymphoma (differential) diagnosis: a standardized approach with underappreciated potential

Lennart Raman,[1,2,*] Malaïka Van der Linden,[1,2,*] Ciel De Vriendt,[3,*] Bliede Van den Broeck,[4] Kristoff Muylle,[5] Dries Deeren,[6] Franceska Dedeurwaerdere,[7] Sofie Verbeke,[1] Amélie Dendooven,[1,8] Katrien De Grove,[3] Saskia Baert,[3] Kathleen Claes,[2] Björn Menten,[2] Fritz Offner[3#] and Jo Van Dorpe[1#]

[1]Department of Pathology, Ghent University, Ghent University Hospital, Ghent; [2]Center for Medical Genetics, Department of Biomolecular Medicine, Ghent University, Ghent University Hospital, Ghent; [3]Department of Clinical Hematology, Ghent University, Ghent University Hospital, Ghent; [4]Department of Nuclear Medicine, Ghent University, Ghent University Hospital, Ghent; [5]Department of Nuclear Medicine, AZ Delta, Roeselare; [6]Department of Hematology, AZ Delta, Roeselare; [7]Deparment of Pathology, AZ Delta, Roeselare and [8]Faculty of Medicine and Health Sciences, University of Antwerp, Wilrijk, Belgium

**LR, MVDL and CDV contributed equally as co-first authors.*

#JVD and FO contributed equally as co-senior authors.

 doi:10.3324/haematol.2020.268813

Received: August 3, 2020.

Accepted: November 25, 2020.

Pre-published: December 10, 2020.

Correspondence: *JO VAN DORPE* - Jo.VanDorpe@uzgent.be

Supplemental file 1

Shallow-depth sequencing of cell-free DNA for Hodgkin and diffuse large B-cell lymphoma (differential) diagnosis: a standardized approach with underappreciated potential

Supplemental table legends

All tables are included in Supplemental file 2.

Table S1

Summary of cohort characteristics such as demographics, histology, International Prognostic Index/Score (IPI/IPS) at diagnosis, and stage at diagnosis. Counts per category are given with corresponding fractions in parentheses.

Table S2

Patient-wise cohort characteristics such as demographics, histology, IPI/IPS at diagnosis, and stage at diagnosis.

Table S3

Overview of every sequenced sample, including all solid and (serial) liquid biopsies, chronologically ordered per patient. Day 0 corresponds to baseline (~first liquid biopsy). Column 'at' indicates the state of the disease at time of biopsy: 'staging' (possible at baseline or progression), 'interim' (evaluation during therapy) or 'surveillance' (remained complete remission (CR) evaluation). All sequencing-derived information can be found to the right of the second vertical line, including sequencing depth, ichorCNA tumor fraction predictions (these numbers should be interpreted with caution, as this method lacks sensitivity), the copy number profile abnormality (CPA) score and read fractions of multiple viruses (see Supplemental methods below on how these numbers were calculated). For the latter two, significantly elevated numbers in comparison to the controls are marked in green. For the read fractions specifically, the absolute number of non-duplicate viral reads are included in parentheses.

Table S4

Similar to Table S3 but depicts clinical variables including metabolic tumor volume (MTV), Ann Arbor Stage, IPI/IPS, beta-2 microglobulin (B2M) concentration, lactate dehydrogenase (LDH) concentration, chromogenic in situ hybridization (CISH) for Epstein-Barr virus (EBV) testing, fluorescence in situ hybridization (FISH) testing and immunohistochemistry (IHC) assays.

Table S5

Contains significantly aberrant focal gains and losses within Hodgkin lymphoma (HL) and diffuse large B-cell lymphoma (DLBCL) separately according to the GISTIC software. The genes located in these regions are shown in addition. Significance is expressed as a Q-value (~adjusted P-value).

Table S6

Segmental copy number data of all sequenced samples. The ratio (log_2 transformed) presents the observed over the expected number of reads, whereas Z-scores indicate the amount of deviation from healthy background variability at the corresponding locus. Both are inferred by WisecondorX.

Table S7

Mapped read counts (non-duplicate raw and locus-filtered; see Supplemental methods below) and subsequently derived normalized viral read fractions across all sequenced samples.

Supplemental methods

Formalin-fixed paraffin-embedded (FFPE) DNA sequencing

DNA extraction was performed with the QIAamp DNA FFPE Tissue Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions. DNA shearing to 200 bp fragments was executed by Covaris' Adaptive Focused Acoustics technology (Covaris, Woburn, Massachusetts). Using 200 ng of starting material, library construction was completed by use of the NEXTflex Rapid DNA-Seq Kit and NEXTflex DNA Barcodes (Bioo Scientific, Austin, Texas). Cluster generation and sequencing were executed by respectively a cBot 2 and HiSeq 3000 system (Illumina, Essex, UK). The minimal number of reads (single-read; 50-cycle mode) per sample was aimed to be 15 million (mean coverage of 0.25x).

Cell-free DNA (cfDNA) sequencing

Blood samples were collected in cfDNA BCT tubes (Streck, La Vista, Nebraska) or PAXgene Blood ccfDNA tubes (PreAnalytiX, Hombrechtikon, Switzerland). Within 72 hours of collection, plasma isolation was executed by one (PAXgene) or two consecutive (BCT) centrifugation steps, according to the manufacturer's protocol. cfDNA extraction from 4 mL of plasma was performed using the Maxwell RSC ccfDNA Plasma Kit (Promega, Madison, Wisconsin), following the manufacturer's instructions. Using 25µL of cfDNA, library preparation was executed by use of the NEXTflex cfDNA-Seq Library Prep Kit (Bioo Scientific) and NEXTflex DNA Barcodes (Bioo Scientific). Cluster generation and sequencing were performed in correspondence to the solid biopsies.

Copy number profiling

Raw reads were mapped by Bowtie 2[1] (v2.3.2) onto human reference genome GRCh38, using the *fast-local* flag. Biobambam's bamsormadup[2] (v2.0.87) was used to mark duplicate reads and to sort the resulting bam files. No additional quality filtering was applied. The latter files were indexed by SAMtools[3] (v1.4.1). The novel WisecondorX[4] (v1.1.2) was deployed to reliably deduce normalized genome-wide $\log_2$ ratios across 100 kb bins, representing copy number. Technical normalization was performed using two healthy reference sets: one for cfDNA (n=333), derived from non-invasive prenatal assays; and one for FFPE DNA (n=181), represented by genomic DNA samples routinely collected in the context of congenital diseases. Stretches of expected equal copy number were defined by circular binary segmentation segments (contain at least two bins; therefore, the limit of detection in terms of alteration size is 200kb).[5] These were seen as aberrant once they had an absolute Z-score[4] of 3 or more. Bins with missing data were interpreted as loci of undeterminable copy number (e.g., at centromeres).

Next to bin-wise and segmental $\log_2$ ratios;[4] and segmental Z-scores,[4] copy number profiles can be represented by smoothened $\log_2$ ratios: using a one-dimensional sliding window, sized 100 bins (10 Mb), the average was interpolated to the window's middle bin at each movement. If the copy number of more than half of the covered bins was undeterminable, the middle bin was set to undetermined as well.

Copy number tumor burden

The CPA score aims at quantifying tumor burden according to copy number profiles, as shown by equation 1. This formula is discussed in greater depth by Raman et al.[6]

$$CPA = \sum_{i=1}^{n}\left(\left|Z_{segment_i}\right| * l_{segment_i}\right)/n \ * 1e^{-8} \qquad \text{(equation 1)}$$

Above, $Z_{segment_i}$ represents the Z-score[4] of $segment_i$. The length of this segment is given by $l_{segment_i}$. Copy number profiles are defined by n segments. The CPA score is expressed per 100 Mb.

A less sensitive but more interpretable way to define overall copy number deviance is given by the estimated tumor fraction, which was derived for each sample using ichorCNA[7] (v0.2.0). This tool generates elementary copy number profiles and returns the most likely tumor fraction according to the segments' distribution, using maximum likelihood. Three subclonal states were allowed. The maximum copy number was set to 5. The optimization algorithm was reinitiated using five different starting values, representing the non-tumoral contamination: 20%, 35%, 50%, 65% and 80%. The obtained tumor fraction corresponding to the highest likelihood was interpreted as the final estimated tumor fraction.

Viral segment detection

Raw reads mapped to GRCh38, using Bowtie 2, with mapping quality 40 or higher, were removed using SAMtools, preserving bam files containing exclusively unmapped and low-quality mapped reads. Using consecutively SAMtools and Bowtie 2, the latter sequences were extracted and remapped onto seven different viral reference genomes, retrieved from the National Center for Biotechnology Information, selected for their proven relation to lymphoma: NC_001802.1, Human immunodeficiency virus 1; NC_001722.1, Human immunodeficiency virus 2; NC_007605.1, Human gammaherpesvirus 4 or EBV; NC_001699.1, JC polyomavirus; NC_001436.1, Human T-lymphotropic virus 1; NC_009333.1, Human herpesvirus 8; NC_004102.1, Hepatitis C virus genotype 1. Duplicate reads were removed. Four narrow but highly non-specific regions (NC_007605.1:37230-37260; NC_004102.1:9350-9550; NC_009333.1:23910-25780; NC_009333.1:124860-131250), characterized by an enrichment of mapped reads in both cases and controls (EBV example in Figure S3), were omitted and thereby not considered for the remainder of the study. Finally, the normalized viral read fraction, given in parts per million, was calculated for samples with more than two mapped reads using equation 2.

$$viral\ read\ fraction = \frac{n_{virus}}{n_{total}} * \frac{\mu_{CN}}{2} * 1e^{6} \qquad \text{(equation 2)}$$

Above, n_{virus} and n_{total} equal the number of mapped reads on the corresponding viral reference genome and the number of raw reads, respectively. $\frac{\mu_{CN}}{2}$ is a continuous correction factor for hidden between-sample bias introduced to n_{total} by factors such as sex (males are gonosomal monoploid), or the presence of nuclear germline and somatic deletions, gains or amplifications. μ_{CN} represents the overall mean copy number, which can be approximated using equation 3.

$$\mu_{CN} = \sum_{i=1}^{n}\left(2^{R_{segment_i}} * \ l_{segment_i} * \mathbf{2}\right) / \sum_{i=1}^{n} l_{segment_i} \qquad \text{(equation 3)}$$

$R_{segment_i}$ represents the $\log_2$ ratio[4] of $segment_i$, with length $l_{segment_i}$. Copy number profiles are defined by n segments. Note that the '$* 2$', shown in bold, is not applied when $segment_i$ covers a (part of a) male sex chromosome, since these are naturally present in a single copy.

Random forest modeling

Bin-wise 100 kb copy number Z-scores, interpolated from segmental Z-scores,[4] were used as modeling features. Bins for which one or more samples had missing data were removed from the feature matrix. Predictive modeling was executed by the R package RandomForest[8] (v4.6-14), under default parameters, with the exception of the *sampsize* option: to class-balance training, $.632 * n$ samples[9] were drawn from each of the strata, where n matched the number of samples in the smallest class. The performance was evaluated using out of bag voting[10] (i.e., cross-validation for bootstrapping methods, omitting train/test bias), in combination with receiver operating characteristic analysis, which determined the area under the curve as an overall performance metric.

Detecting driver peaks

Driver genes were predicted by GISTIC[11] (v2.0.23), which detects significantly enriched copy number profile peaks across a predefined set of samples. All parameters remained default with the exception of *scent* and *peak_type*, set to 'none' and 'loo', respectively. A Q-value significance level of 0.1 was adopted.

General statistical testing

Differences between means of normally distributed variables were evaluated using the two-way paired (for solid-liquid pairs) t-test and unpaired Welch's t-test; for the viral fractions, a non-parametric variant was required, thereby preferring the Wilcoxon signed-rank test; binary variables where subjected to Fisher's exact tests. Concordance was assessed by Pearson's correlation coefficient (r) and corresponding P-value. A significance level of 0.05 was adopted.

To define aberrant copy number profiles, we calculated the theoretical cumulative distribution function of the control group's CPA variable, which is assumed to be normally distributed. The anomaly cutoff

was chosen at P(x) = 0.99, delineating a type-1 error limit: samples that cross this number are abnormal at the 1% false discovery rate level.

Statistically calling elevated viral read fractions (for EBV, JC polyomavirus, etc.) is less straight-forward: these data are count-derived, and their distributions are heavily zero-inflated. No reliable model could be implemented, therefore an abnormality cutoff at twice the maximum observation of the control group was arbitrarily installed.

References

1. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–359.

2. Tischler G, Leonard S. biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol Med 2014;9:13.

3. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–2079.

4. Raman L, Dheedene A, De Smet M, Van Dorpe J, Menten B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. Nucleic Acids Res 2019;47(4):1605–1614.

5. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5(4):557–572.

6. Raman L, Van der Linden M, Van der Eecken K, et al. Shallow whole-genome sequencing of plasma cell-free DNA accurately differentiates small from non-small cell lung carcinoma. Genome Med 2020;12(1):35.

7. Adalsteinsson VA, Ha G, Freeman SS, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun 2017;8(1):1324.

8. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2(3):18–22.

9. Efron B, Tibshirani R. Improvements on Cross-Validation: The 632+ Bootstrap Method. J Am Stat Assoc 1997;92(438):548–560.

10. Breiman L. Random Forests. Mach Learn 2001;45(1):5–32.

11. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12(4):R41.

Case reports

Amongst cases with follow-up, four interesting patients further illustrate the potential of shallow-depth sequencing and require special emphasis.

Patient 79

Patient 79 was diagnosed with progressive multifocal leukoencephalopathy following a positive magnetic resonance imaging (MRI) scan and brain biopsy, caused by JC polyomavirus activation, 15 months after reaching CR for primary breast DLBCL. Our data suggests that JC polyomavirus could have been diagnosed rapidly and in a non-invasive manner, as 13 plasmatic JC polyomavirus reads were detected, which disappeared post-treatment (Figure S4A). Despite the limited genome length of this virus (5 kb), shallow-depth sequencing did thus expose plasmatic fragments.

Patient 81

For patient 81, eventually diagnosed with intravascular large B-cell lymphoma through an invasive biopsy, positron emission tomography/computed tomography (PET/CT) images could not distinguish normal brain activity from lymphoma, considering the intravascular location of the neoplasm. After treatment, an MRI scan indicated probable lesion reduction. Shallow-depth sequencing of cfDNA would have detected lymphoma earlier and would have allowed more accurate response monitoring of the administered therapy, which indicates (at least partial) remission (Figure S4B).

Patient 35

Patient 35, a recovered DLBCL patient, was suspected of relapse within the central nervous system, despite equivocal findings (September 2019). One month later, doctors decided to initiate ibrutinib treatment, mostly based on symptoms. A liquid biopsy, taken earlier (September 2019), shows an identical pattern in comparison to diagnosis, univocally exposing relapse (Figure S4C). In this narrow time frame, the CPA score increased from 0.986 to 1.888; and the estimated tumor fraction raised from 17.06% to 55.59%.

Patient X

The final patient was excluded from the main cohort, as his condition did not involve lymphoma, despite the presence of lymphoma-related symptoms. A laparoscopic biopsy and two consecutive trephine biopsies were taken, yet all turned out negative. Finally, a biopsy taken during laparotomy, revealed the presence of a seminoma. Unfortunately, this procedure was complicated by transection of the left ureter. A non-invasive liquid biopsy, collected for this study—given the lymphoma suspicion—between both trephine biopsies, showed an amplification of 12p12.1 (Figure S4D), which is highly suggestive of a germ cell tumor, illustrating the advantage of whole-genome sequencing compared to a targeted variant.

Supplemental figures

Figure S1. Sensitivity analyses in relation to Ann Arbor stage

CPA score comparison between Ann Arbor stages of HL and DLBCL. Dots represent blood samples at baseline; default box plots indicate underlying distributions. The bottom grey box delineates the abnormality cutoff defined by the controls. According to this limit, the numbers on top define which samples have abnormal copy number profiles. The overlapping histograms at the left additionally display the overall CPA distribution.

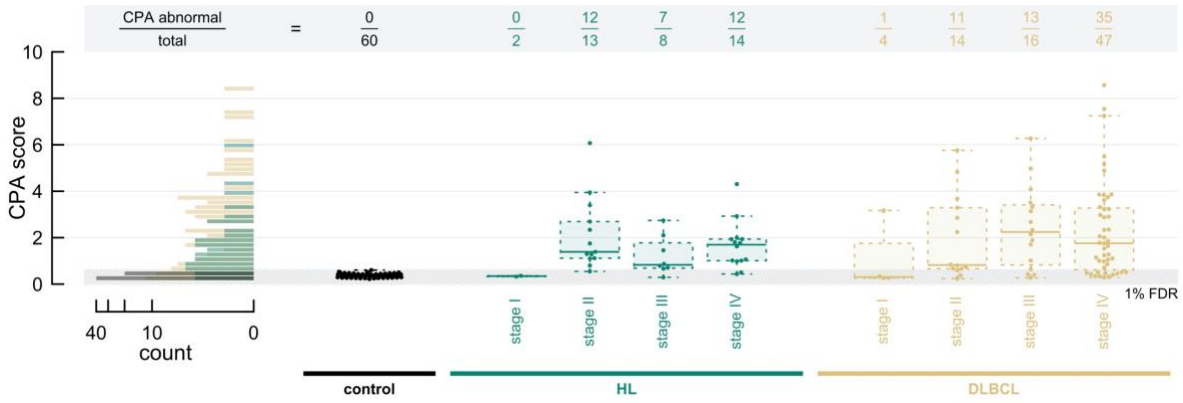


Figure S2. Concordance between clinical parameters and liquid biopsy-derived variables

The matrix' diagonal depicts all evaluated variables; the first two (CPA and tumor fraction (TF)) depend on the copy number profiles; the remainder are clinical parameters. The lower left matrix half contains scatter plots: dots represent liquid biopsies; corresponding least squares fits are shown by green dotted lines. The matrix' diagonal defines the scatter plots' axes. The upper right-hand side matrix half describes the number of available data points (n); the Pearson correlation coefficient (r); and the corresponding P-value (according to a linear model) for each evaluated relation, again specified by the matrix' diagonal. Numbers are shown in bold once significant (P<.05). The background ellipses visualize the tightness of the relation.

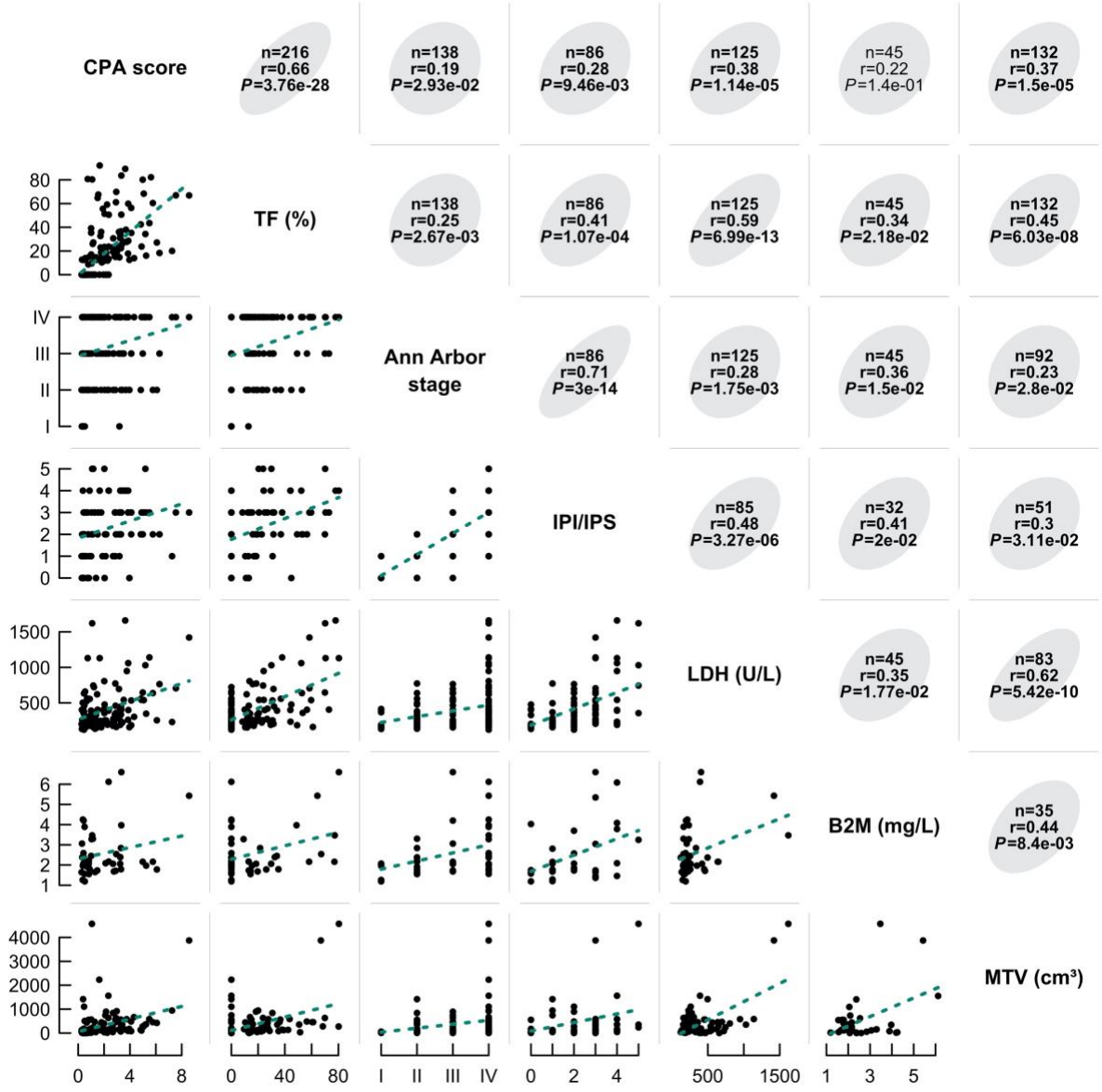


Figure S3. Viral loci filtering

Non-specific regions were filtered out prior to inferring viral read fractions, as executed for locus 37230-37260 in the EBV genome.

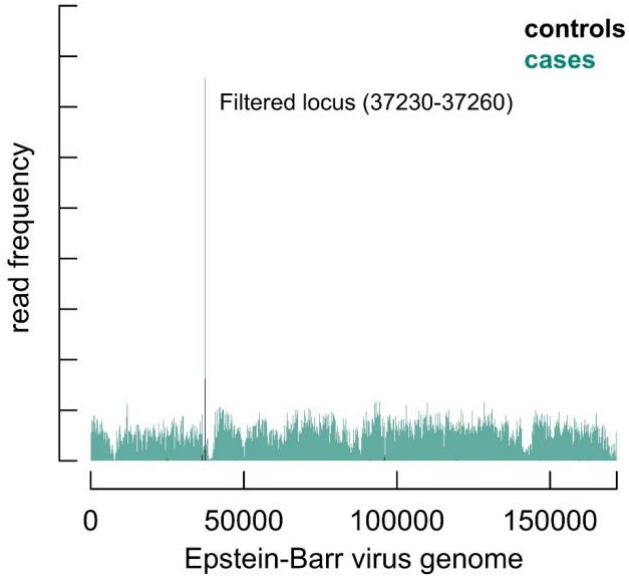
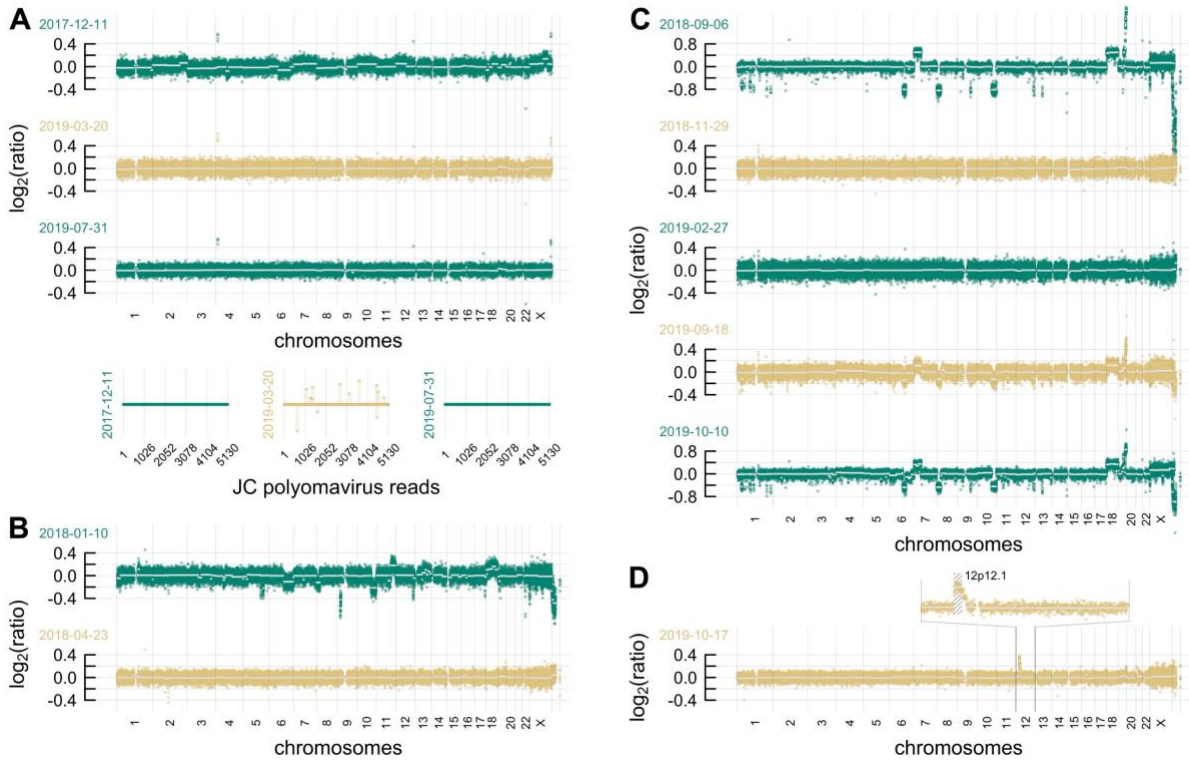


Figure S4. Case reports

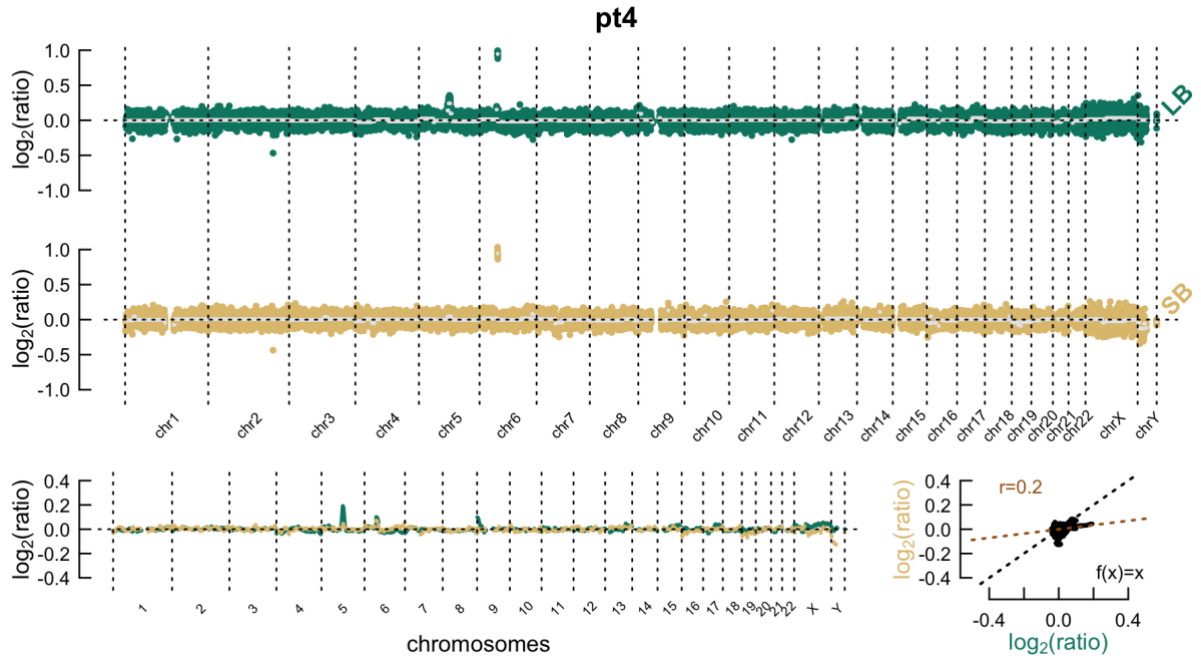
(A) Patient 80, illustrated by three longitudinal liquid biopsies (staging; surveillance; surveillance). Copy number profiles are represented by bins (dots) and segments (horizontal white lines). The three thick horizontal lines at the bottom each illustrate the JC polyomavirus reference genome. Thinner colored lines below and above the latter represent reads, which are connected to the reference at their aligned position. (B) Patient 82, illustrated by two longitudinal liquid biopsies (baseline, interim). Copy number profiles are visualized in correspondence to A. (C) Patient 36, illustrated by five longitudinal liquid biopsies (staging, interim, interim, staging, staging). Copy number profiles are visualized in correspondence to A. (D) The final patient, illustrated by one liquid biopsy. The copy number profile is visualized in correspondence to A. Chromosome 12 is enlarged on top, where the p12.1 region is emphasized.



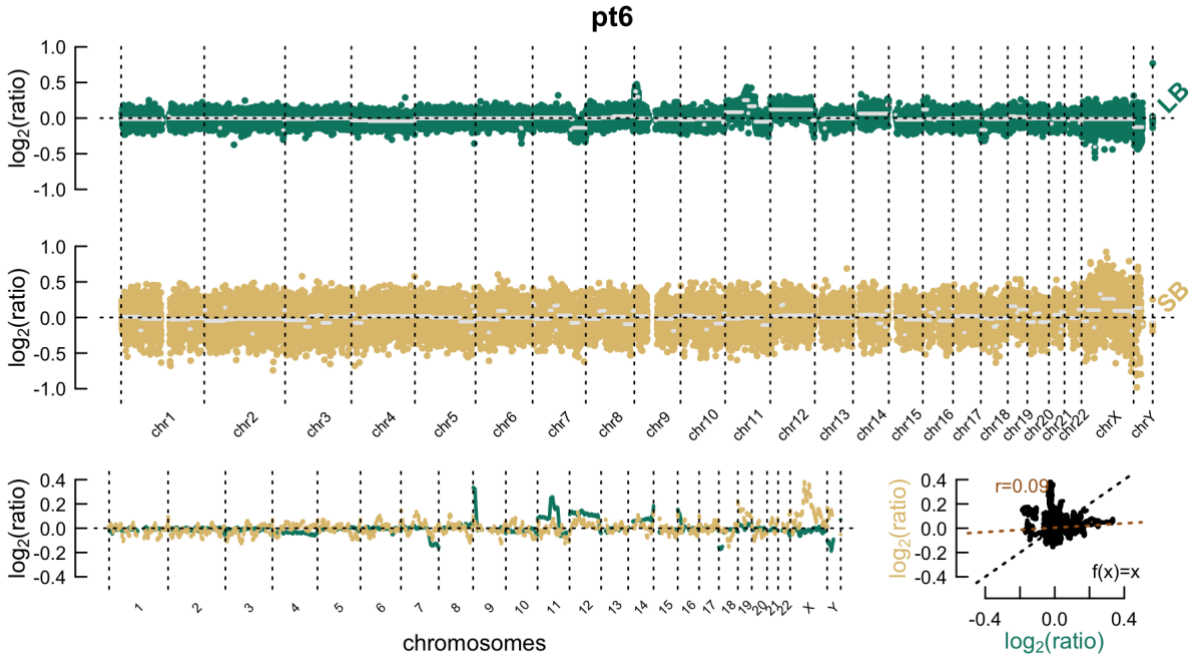
Liquid-solid pairs

Figures below present a copy number profile comparison between the liquid (LB; green) and solid biopsy (SB; yellow) of the named patient. Dots represent bins, whereas horizontal white lines indicate segments, covering bins of expected equal copy number. Underneath, an overlap plot of the smoothened profiles is shown. The right bottom corner contains a correlation scatter plot, using previous smoothened values. The solid line indicates identity (y = x), whereas the dotted line results from a least squares analysis. The Pearson correlation coefficient (r) is given.

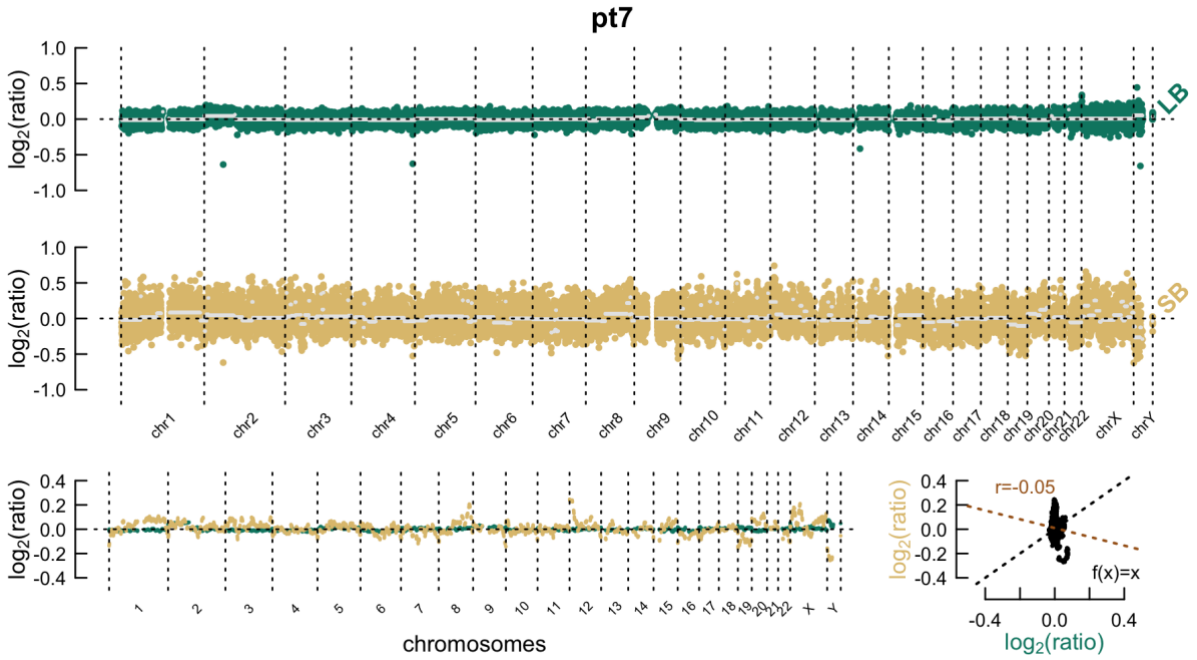
Patient 4



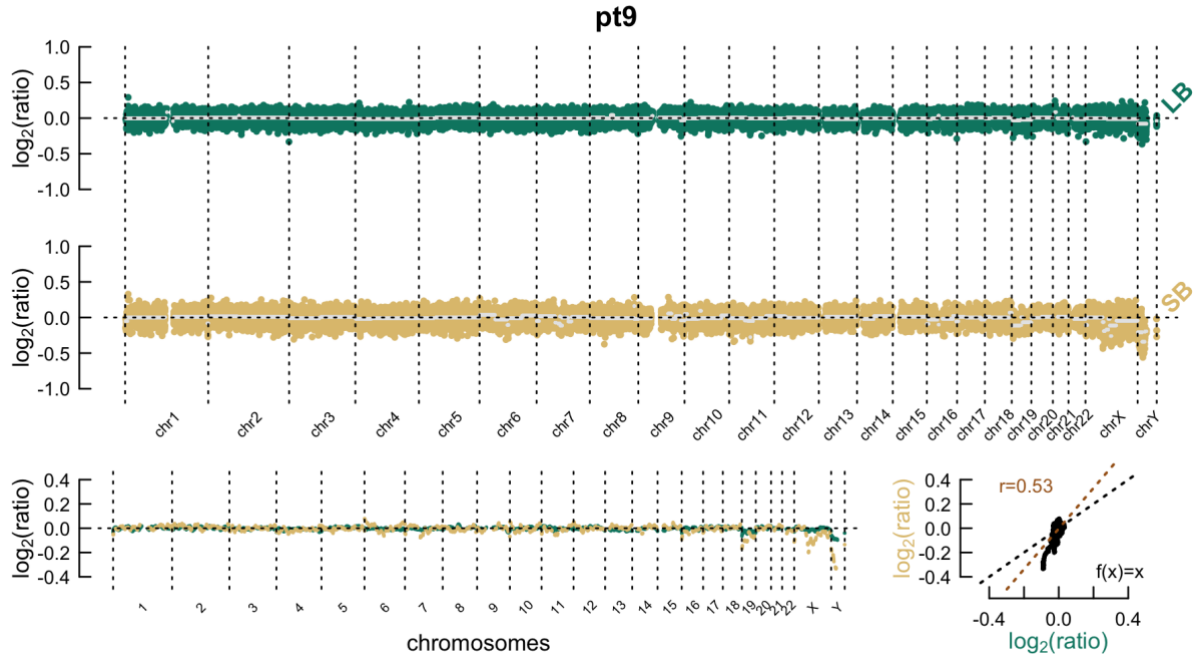
Patient 6



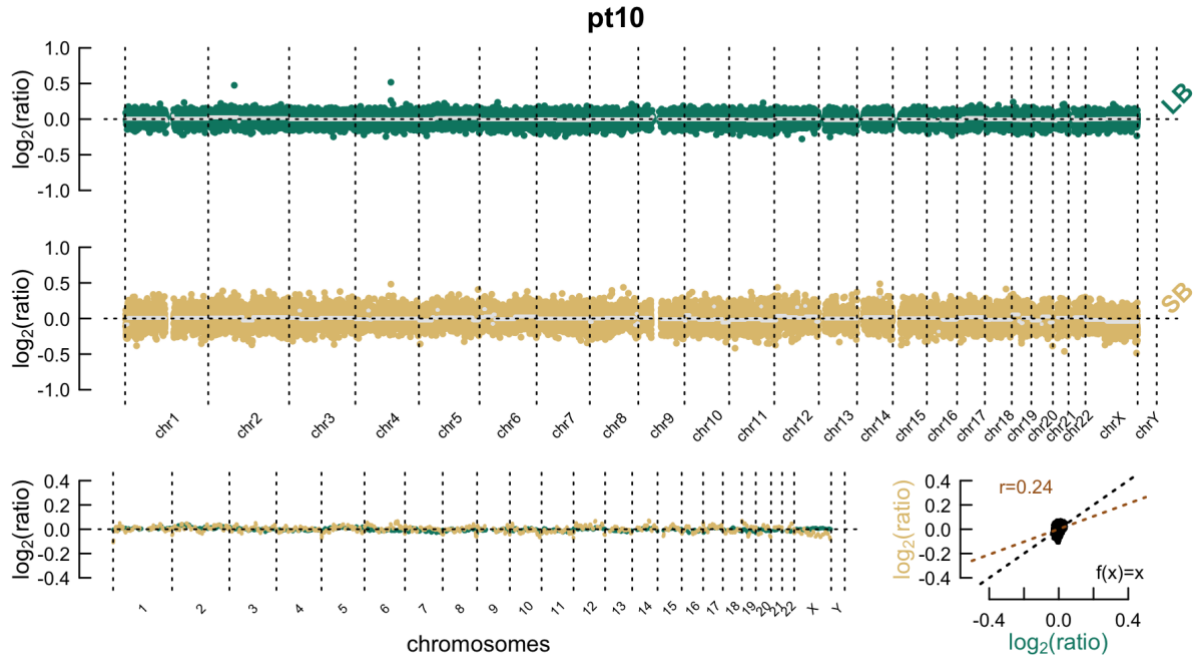
Patient 7



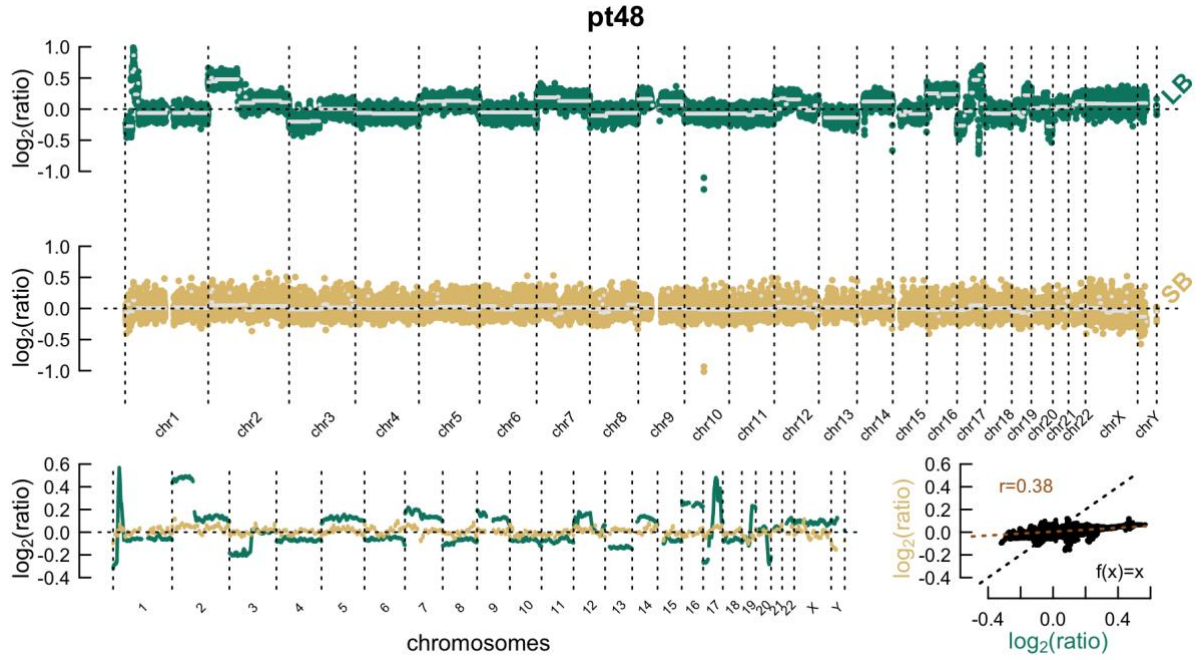
Patient 9



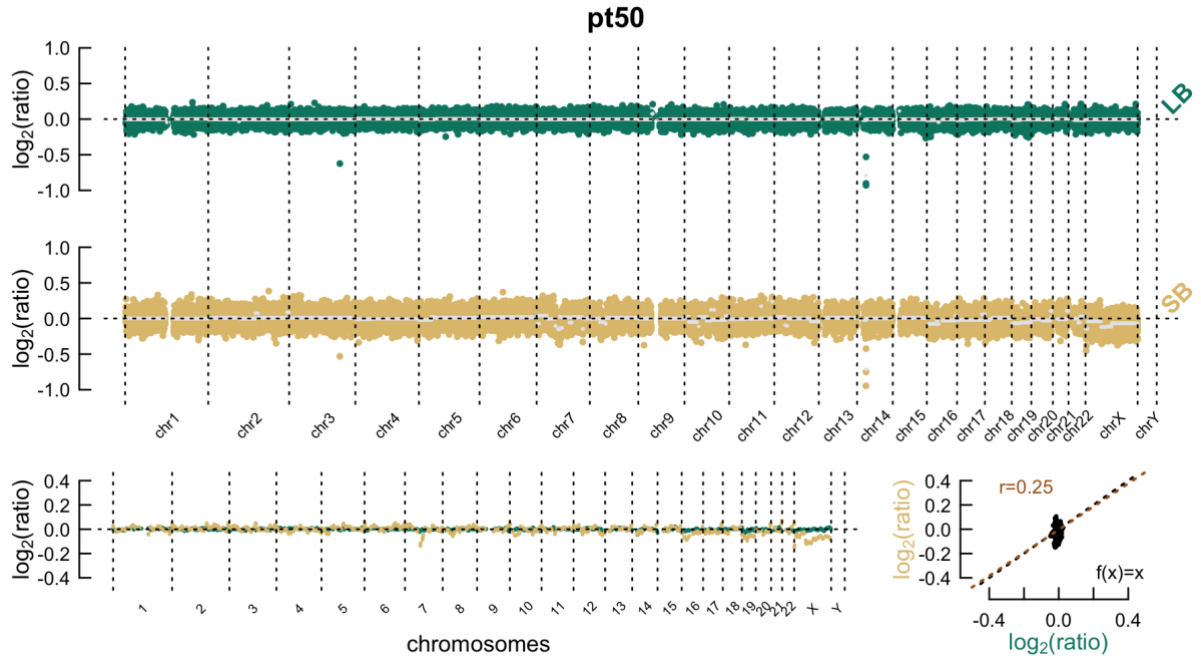
Patient 10



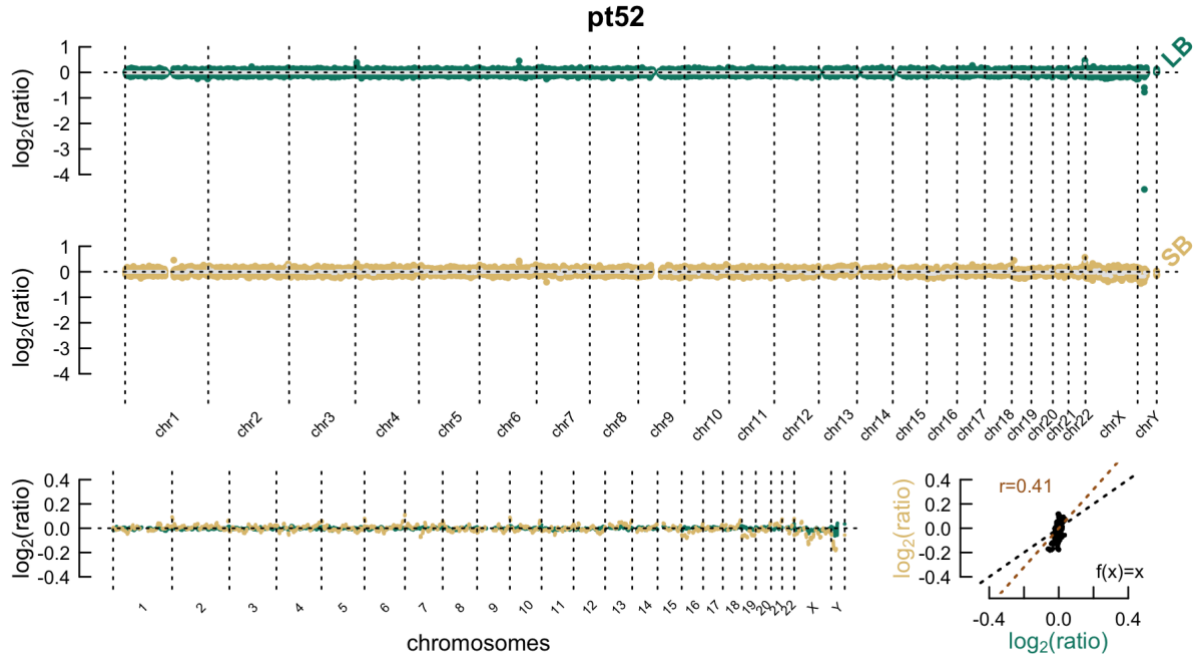
Patient 48



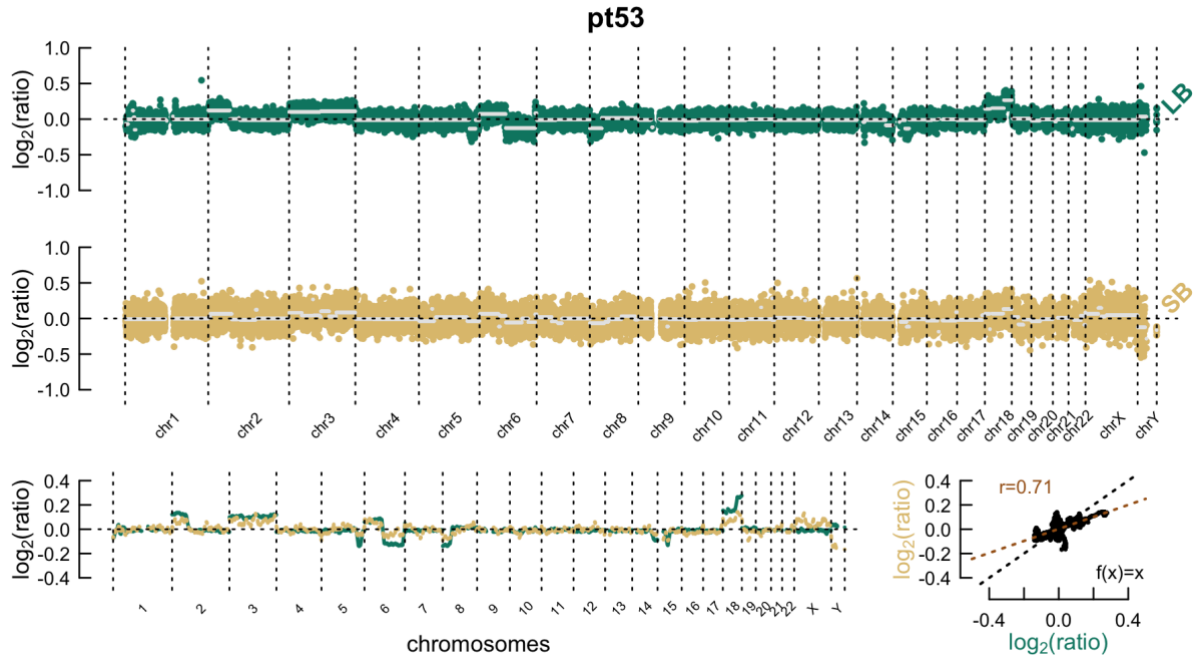
Patient 50



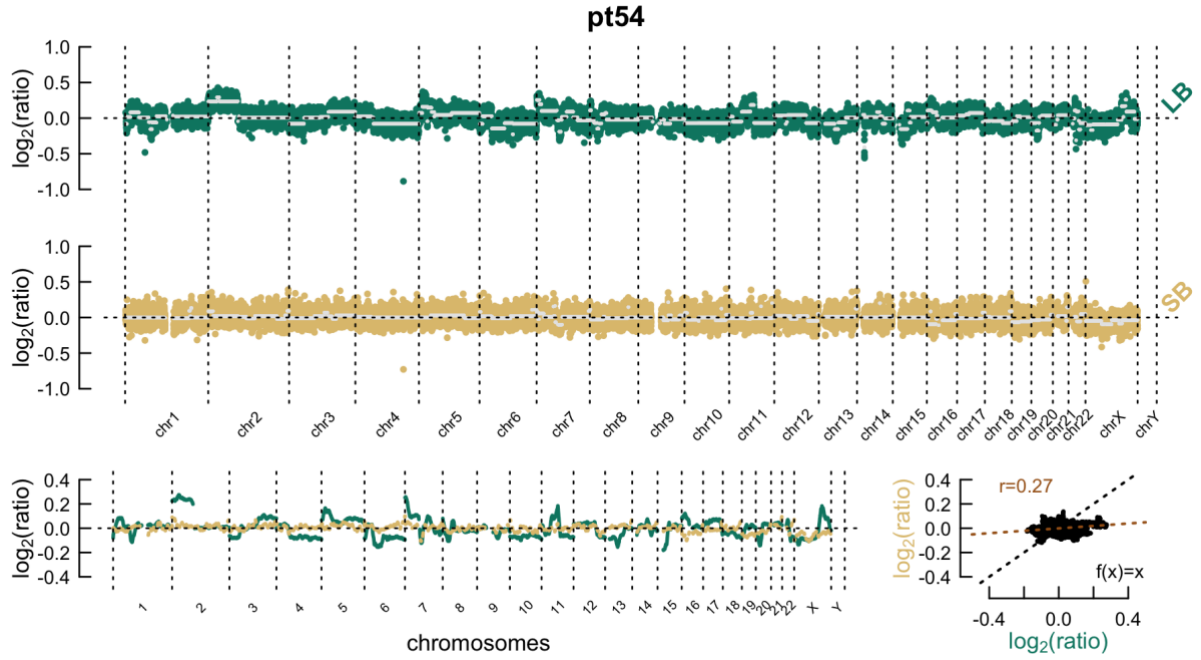
Patient 52



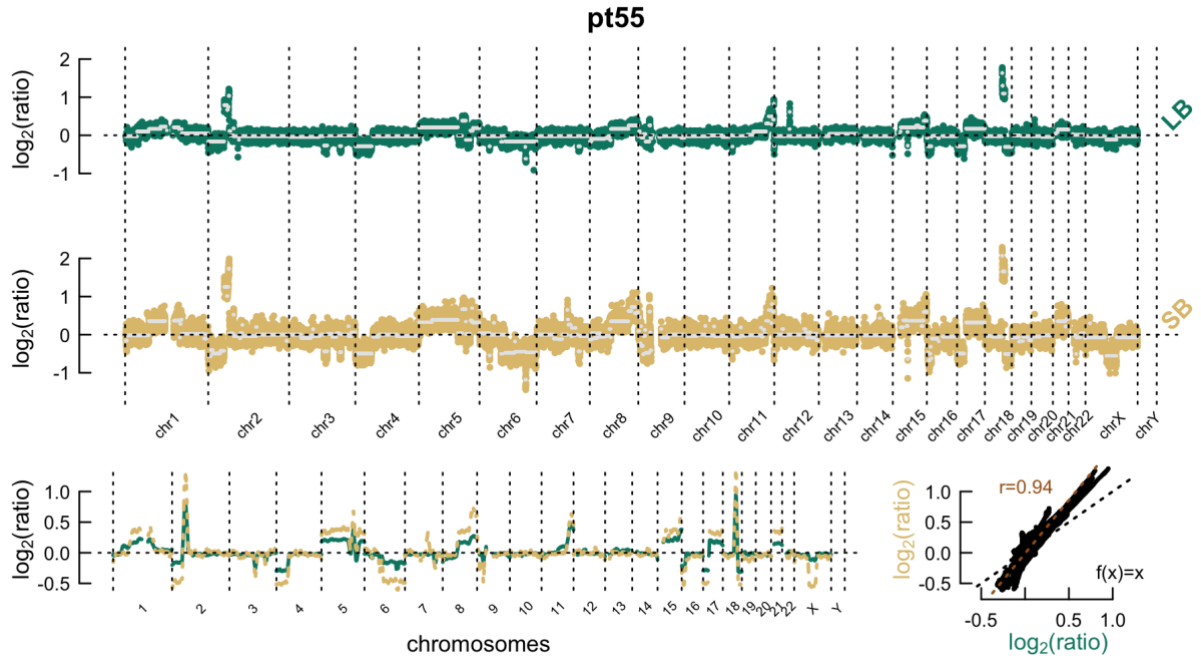
Patient 53



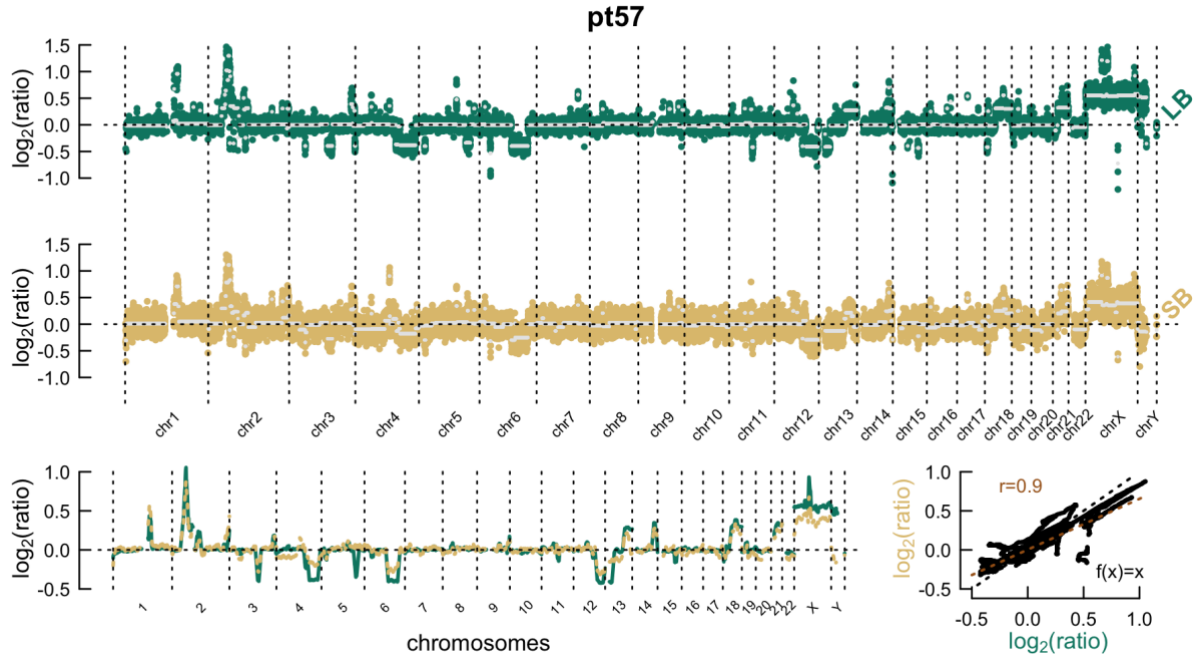
Patient 54



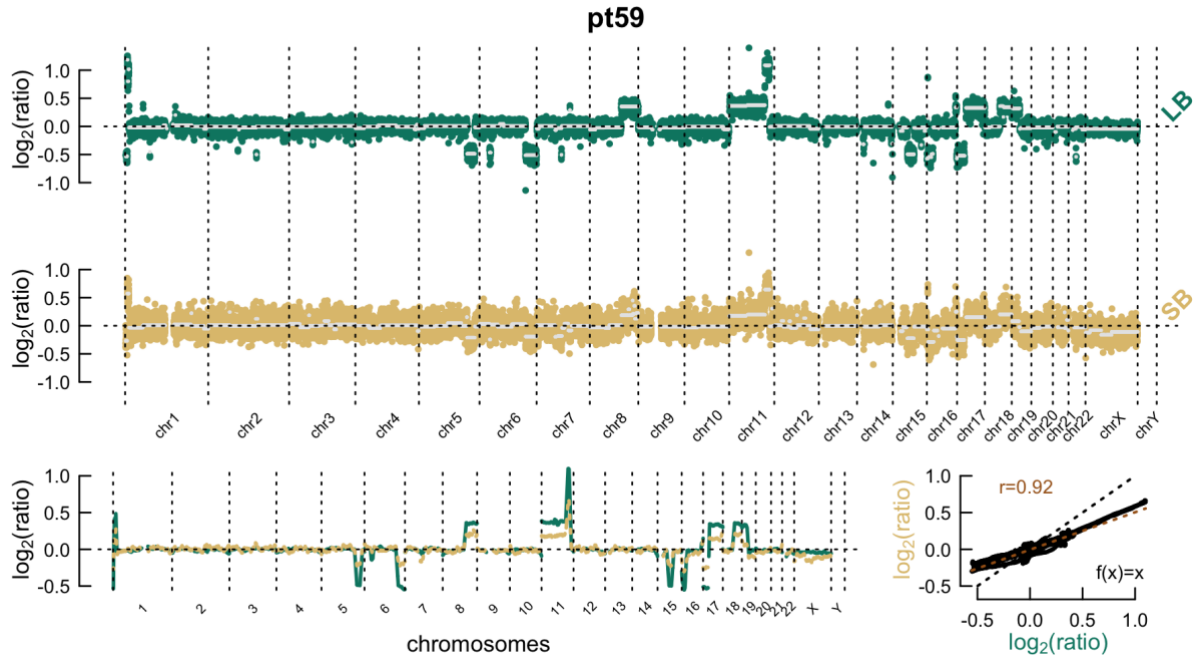
Patient 55



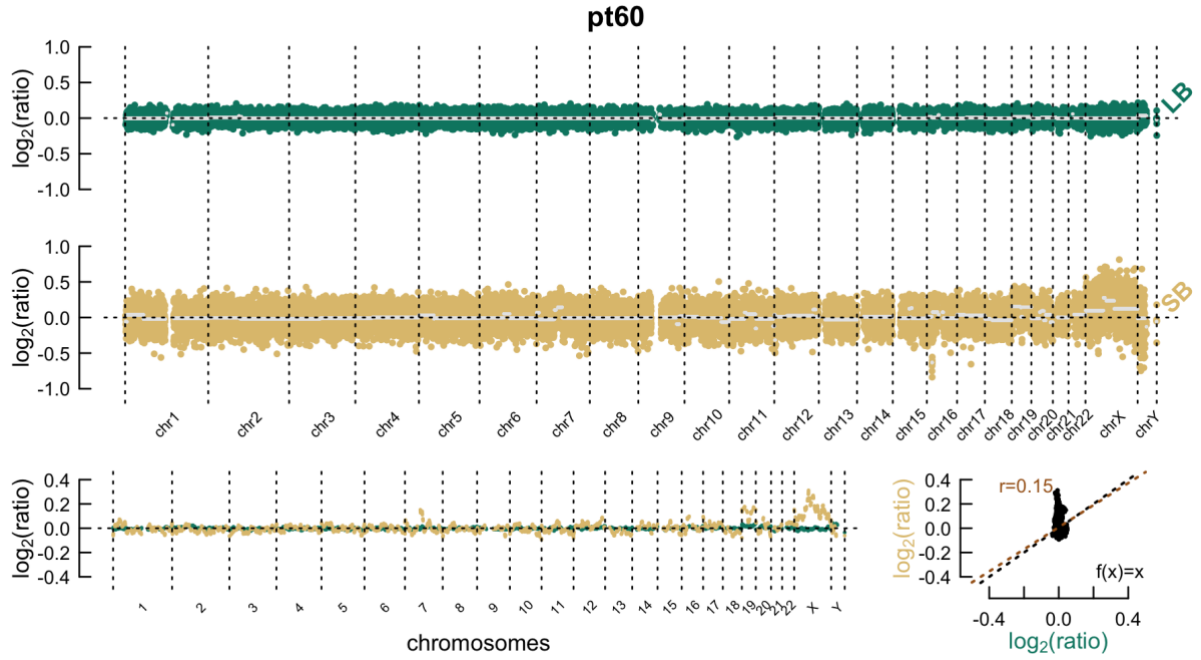
Patient 57



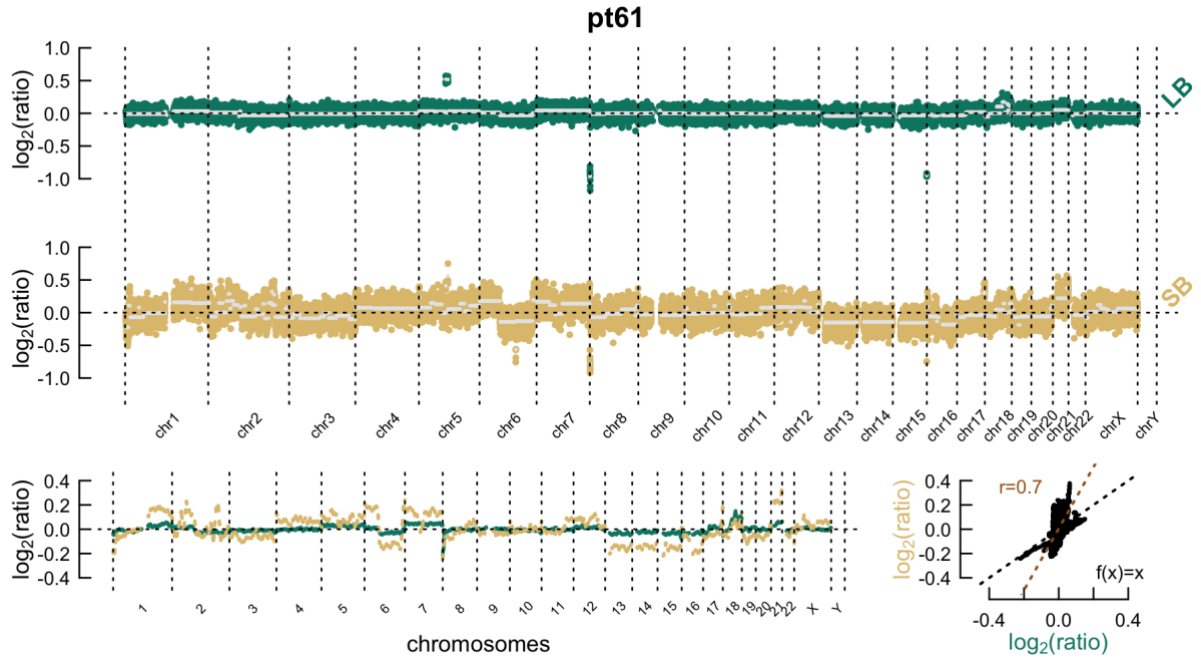
Patient 59



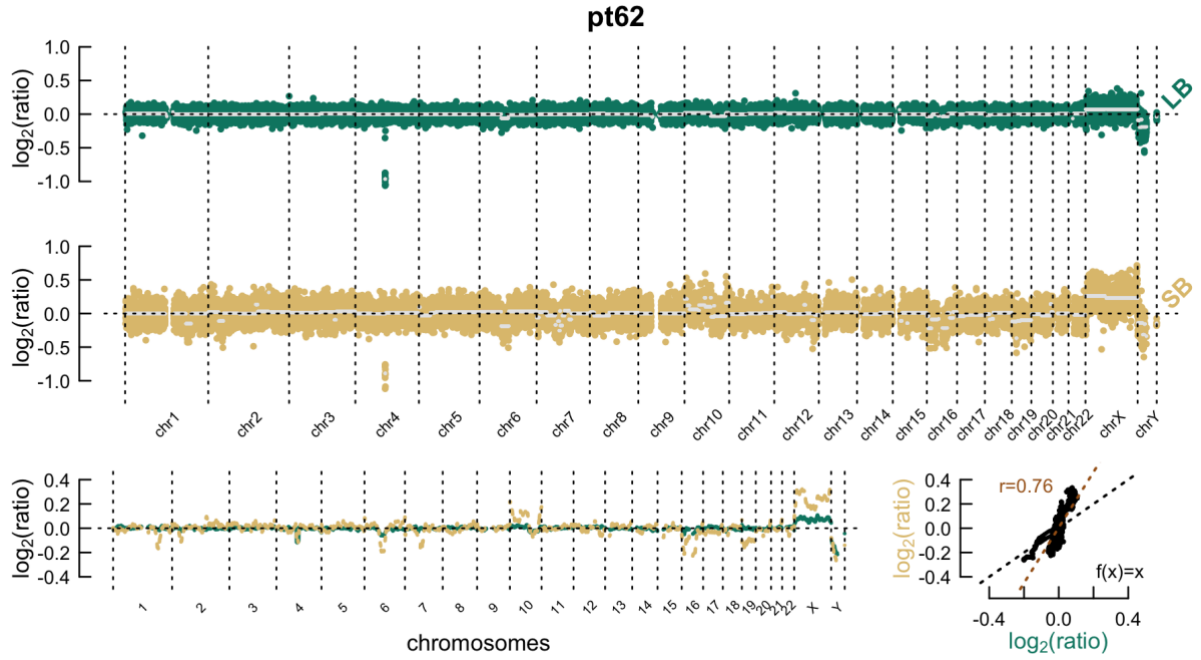
Patient 60



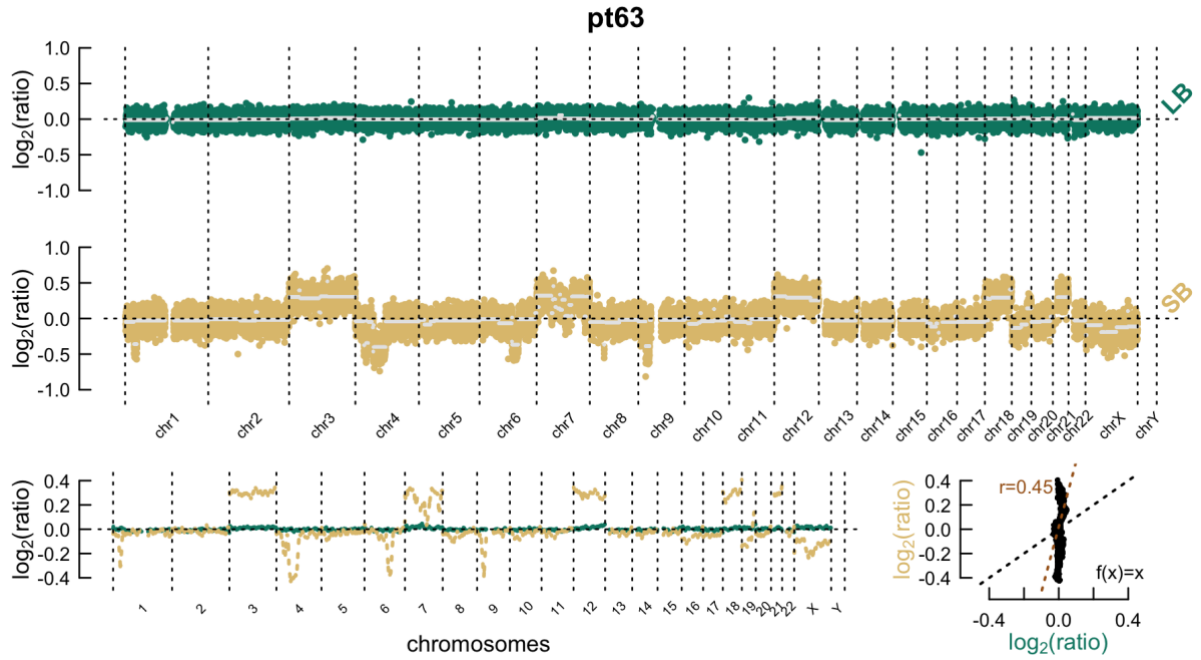
Patient 61



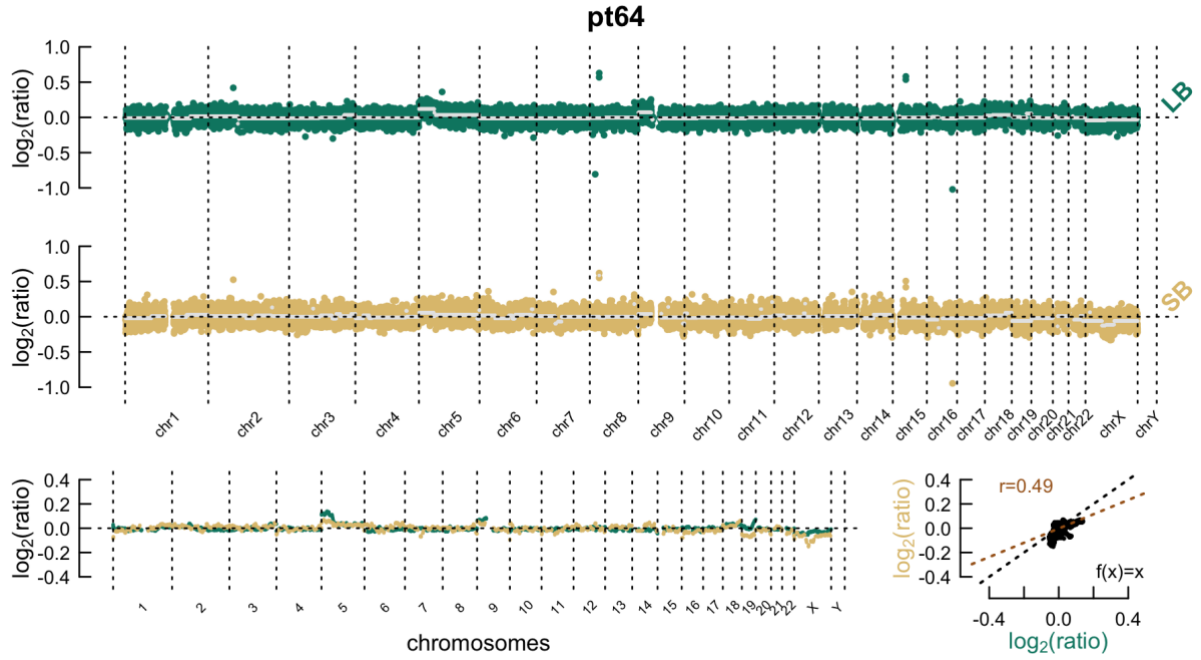
Patient 62



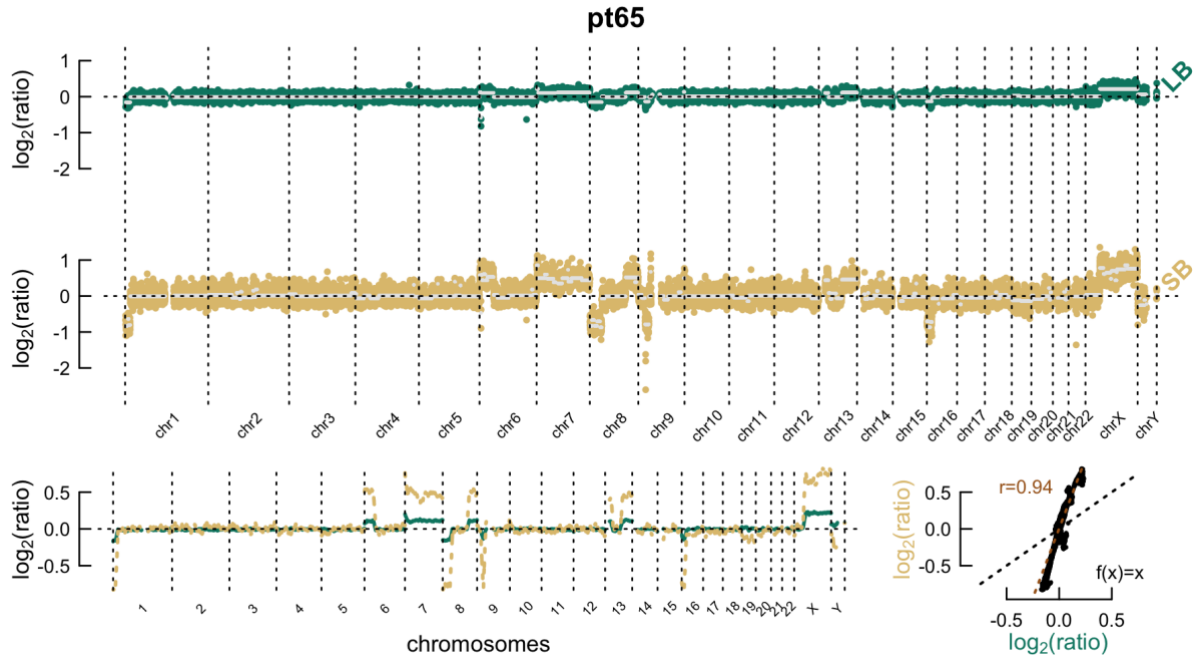
Patient 63



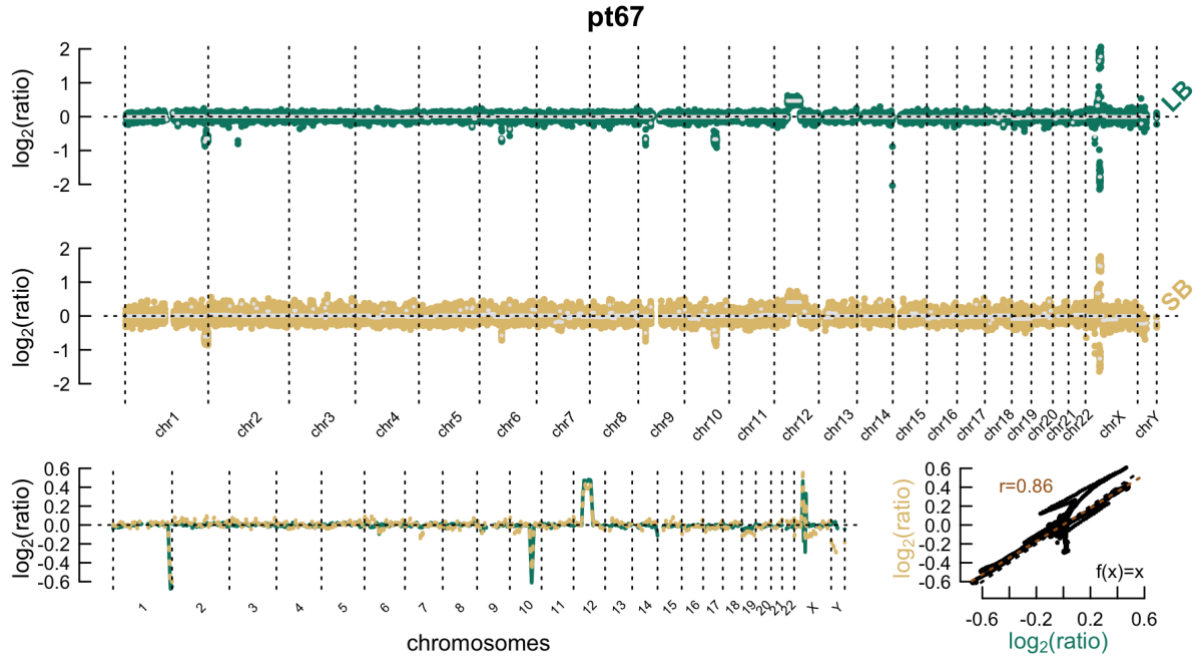
Patient 64



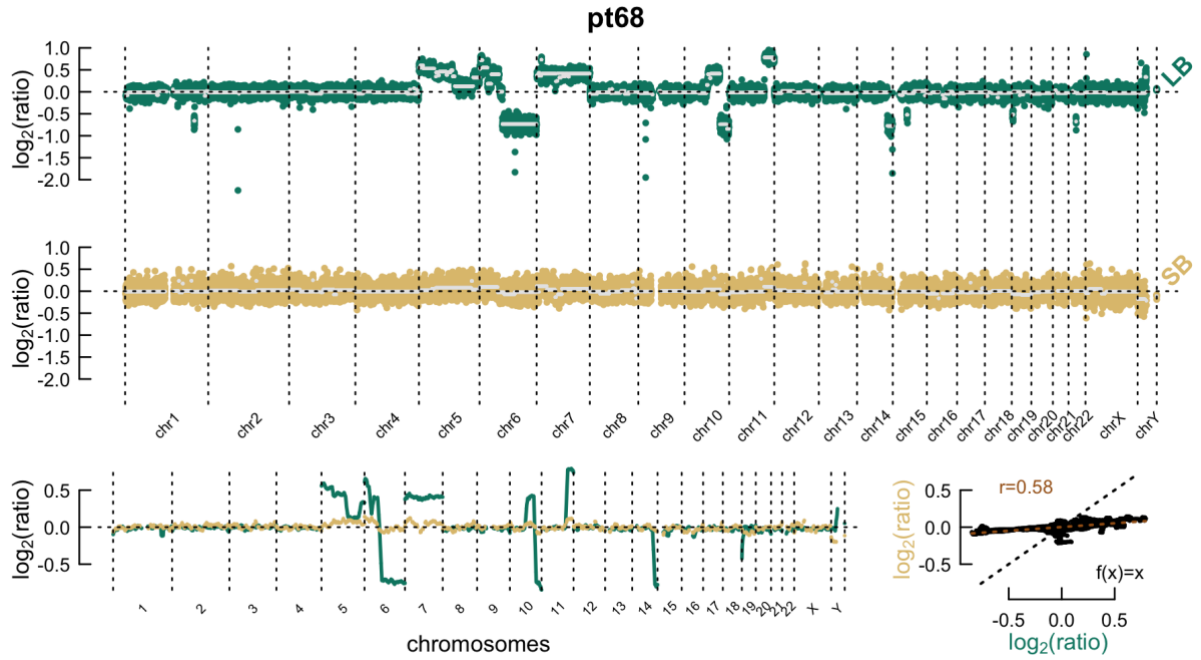
Patient 65



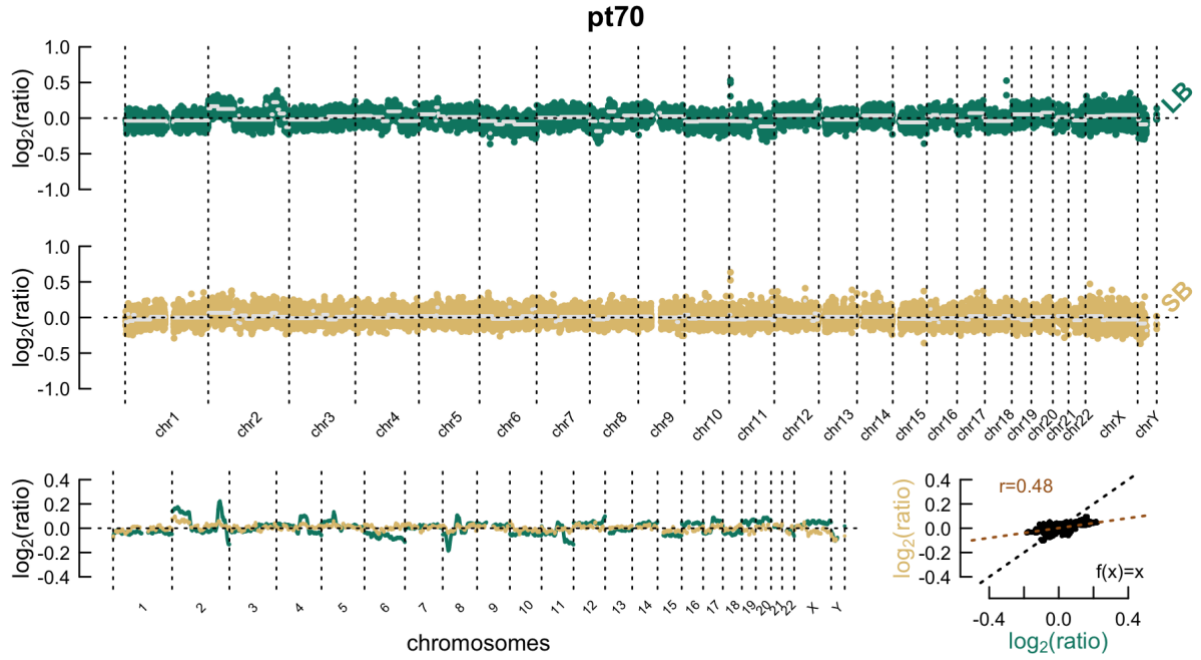
Patient 67



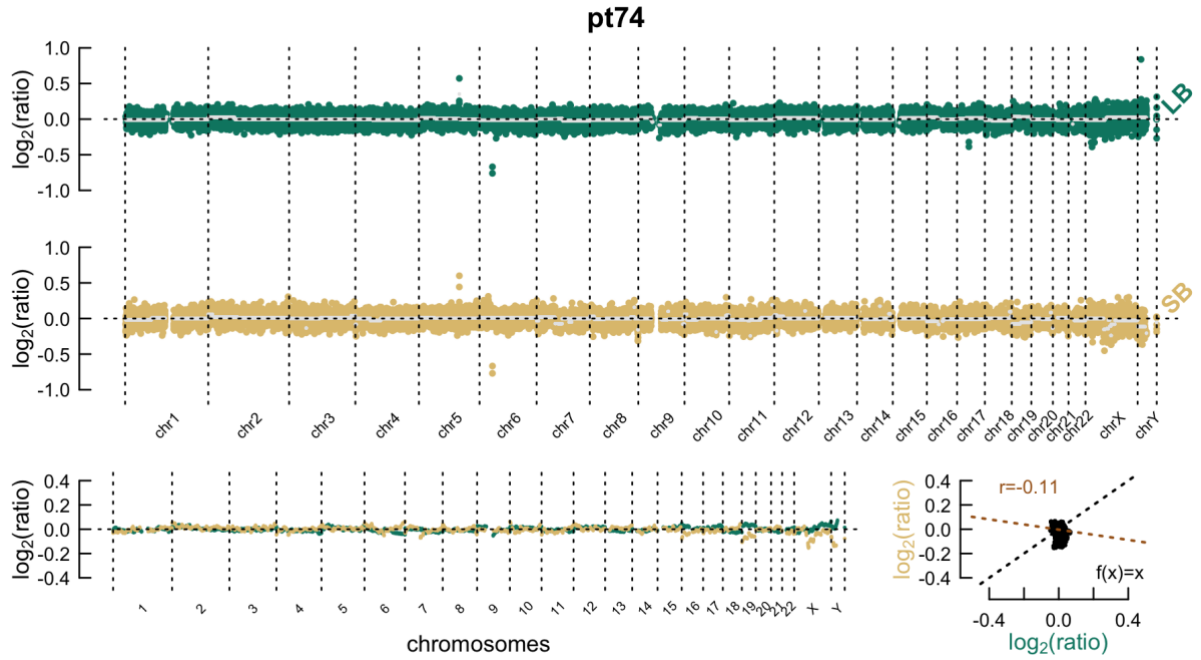
Patient 68



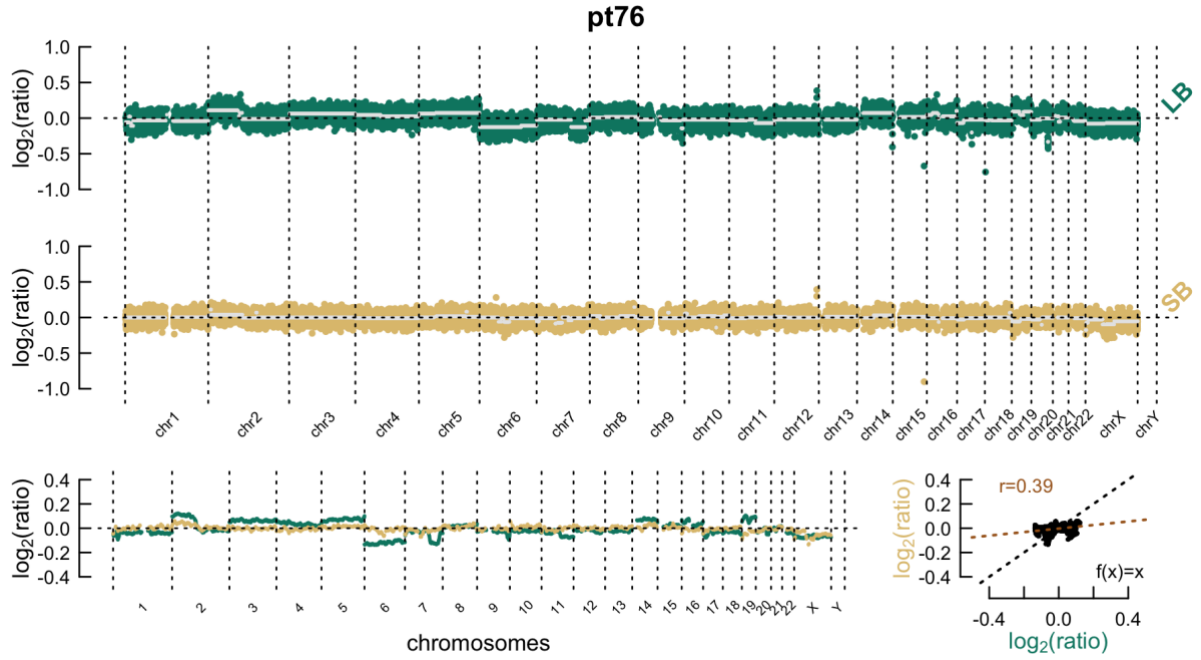
Patient 70



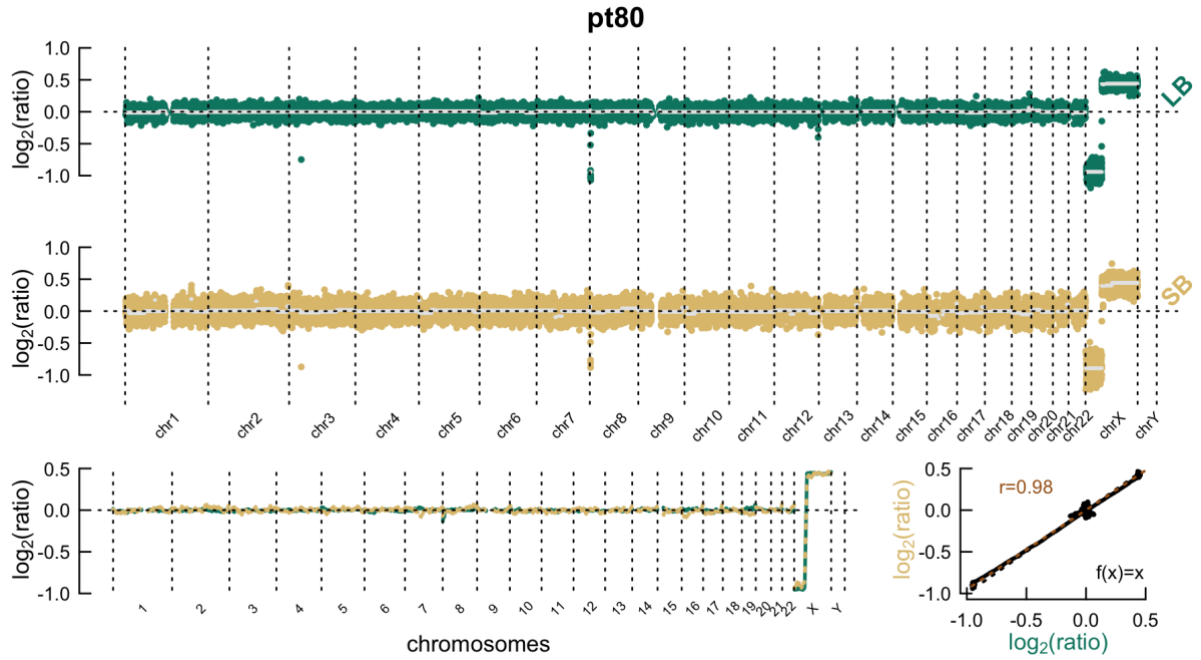
Patient 74



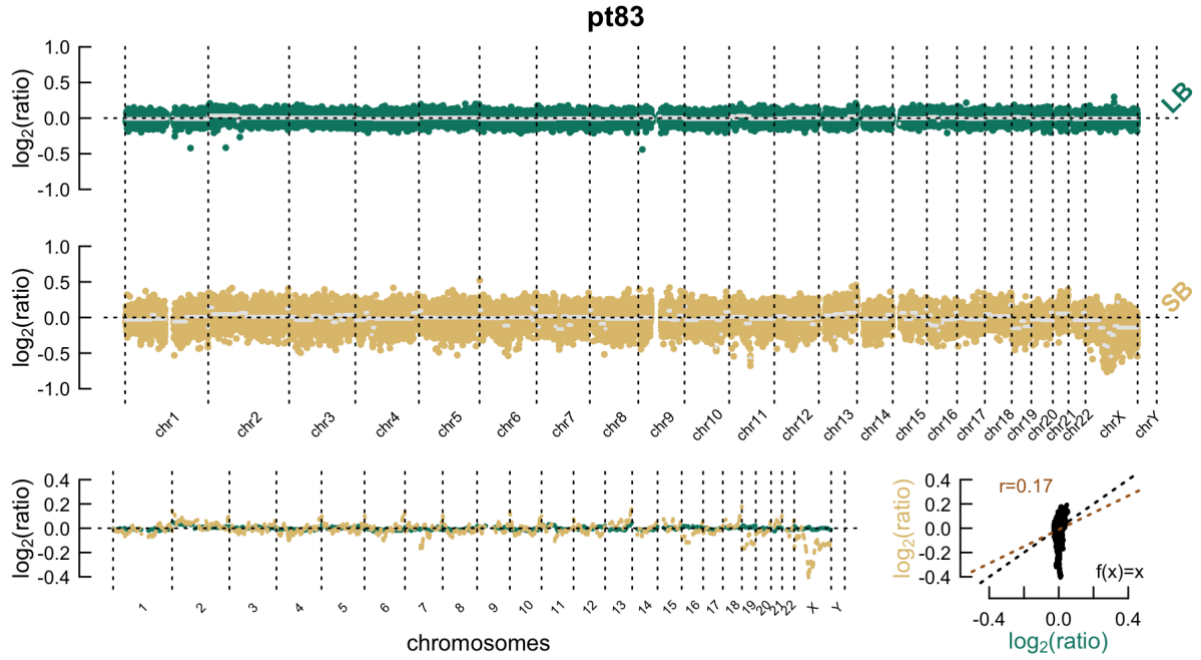
Patient 76



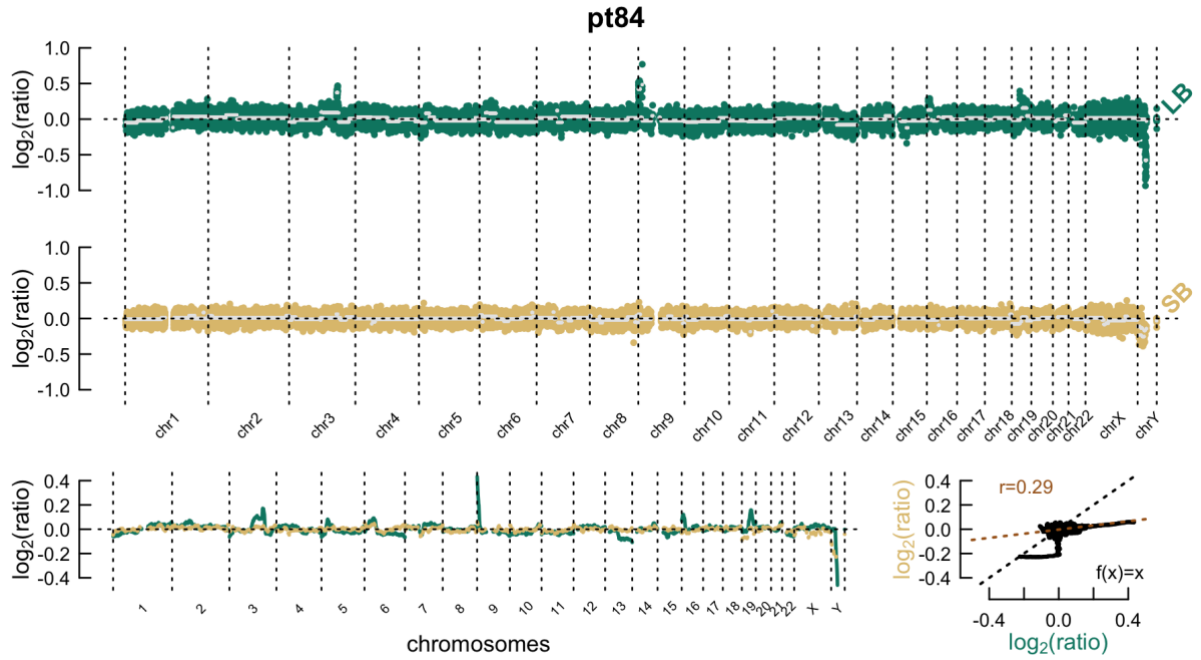
Patient 80



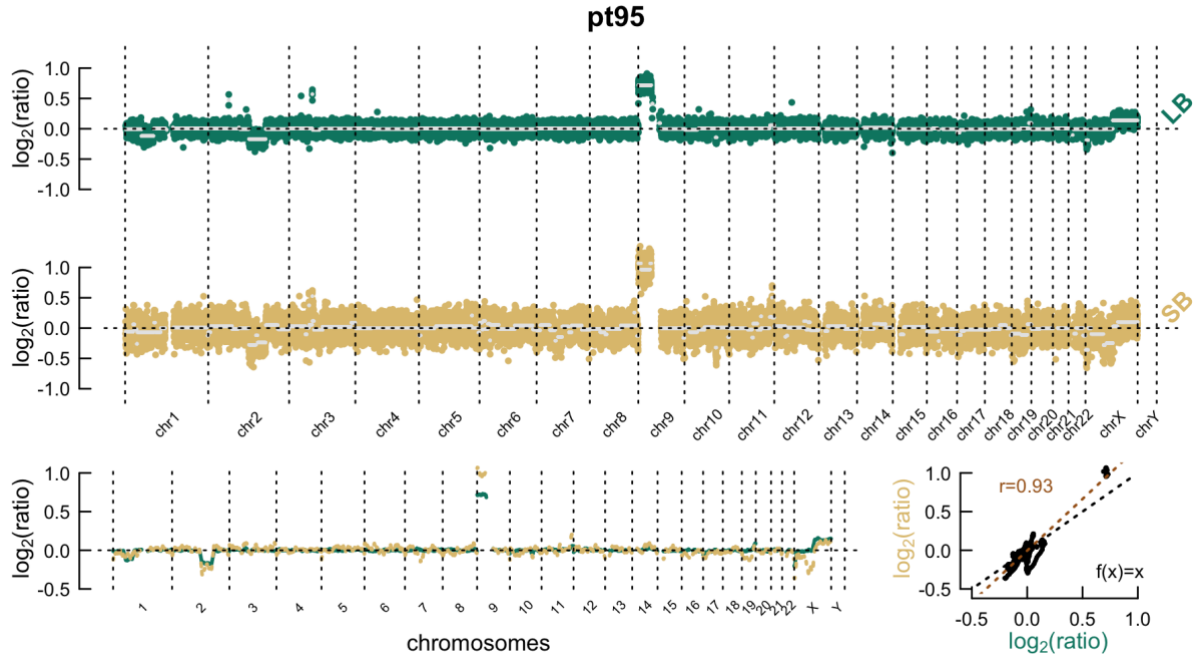
Patient 83



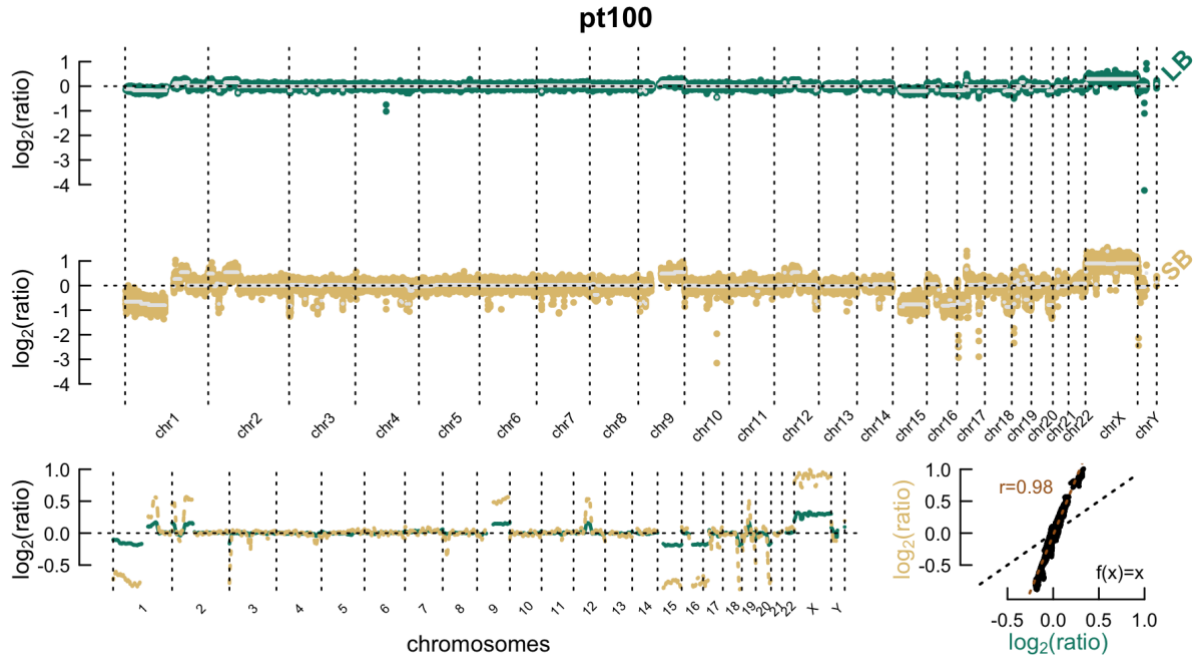
Patient 84



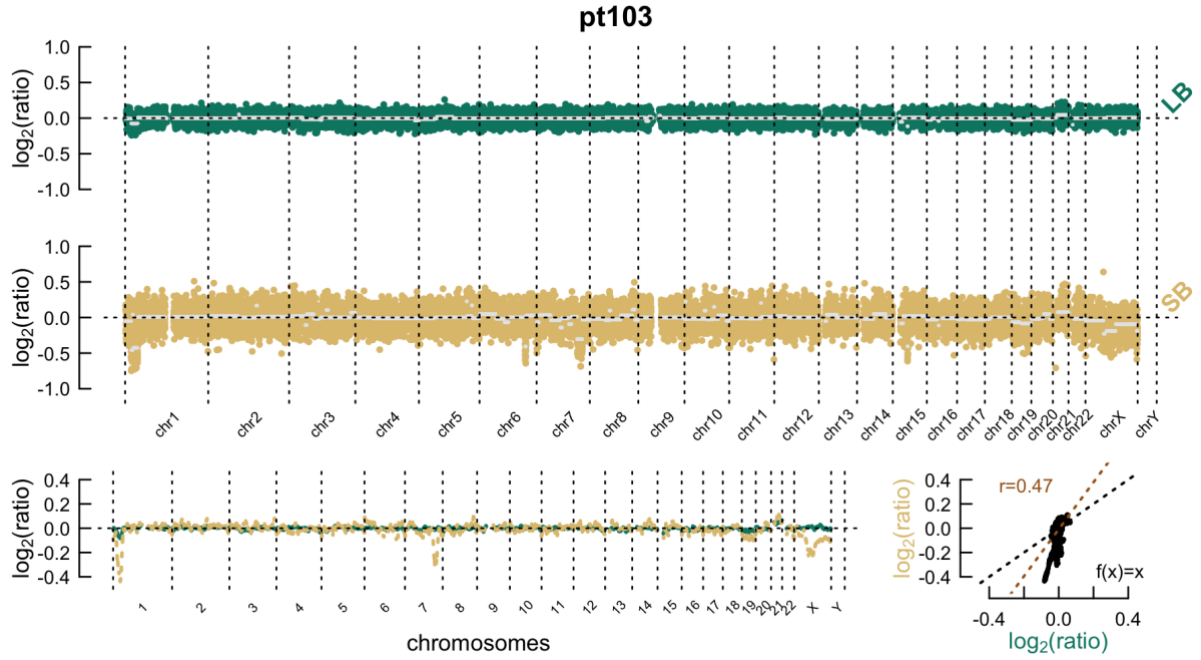
Patient 95



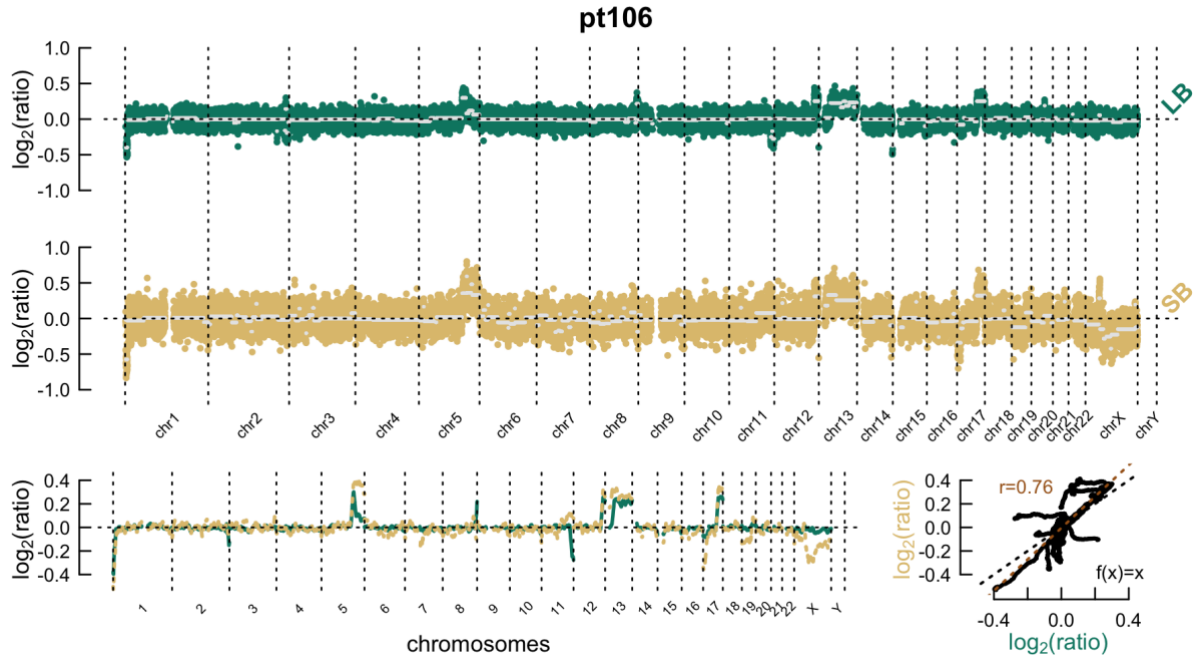
Patient 100



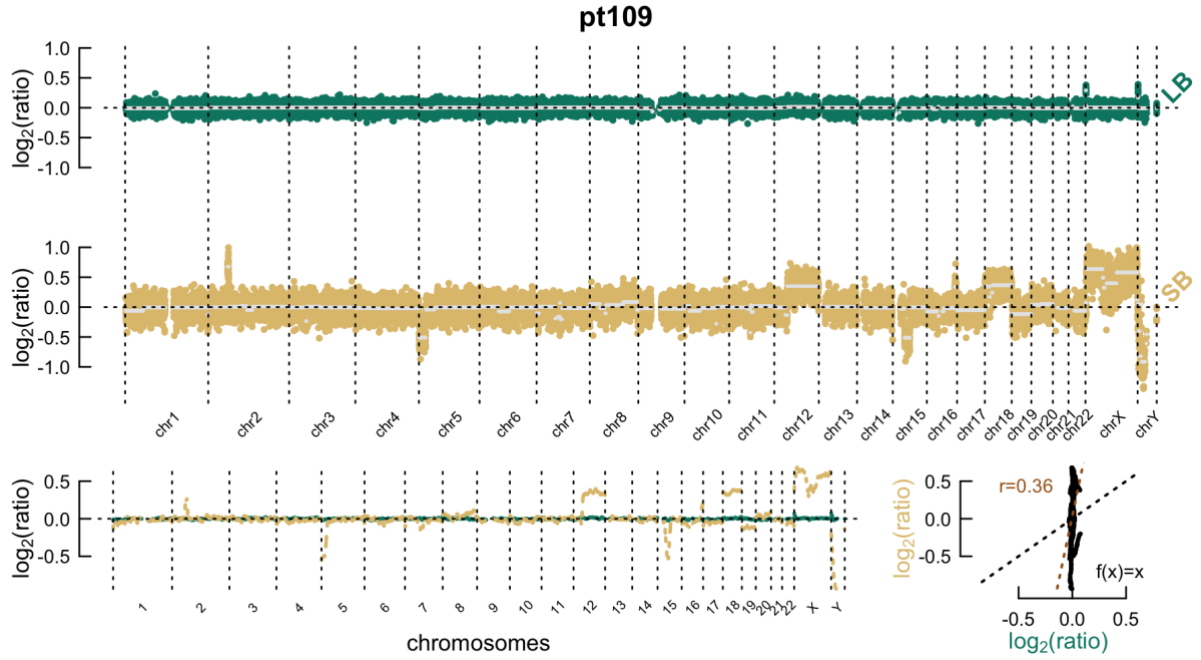
Patient 103



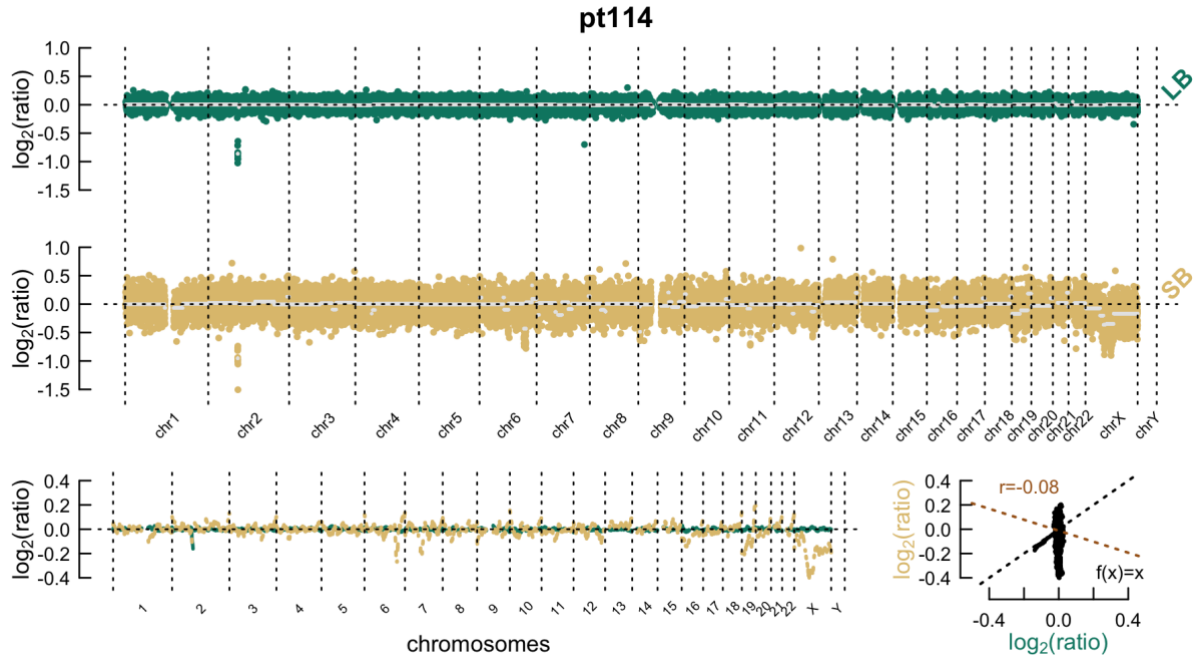
Patient 106



Patient 109



Patient 114



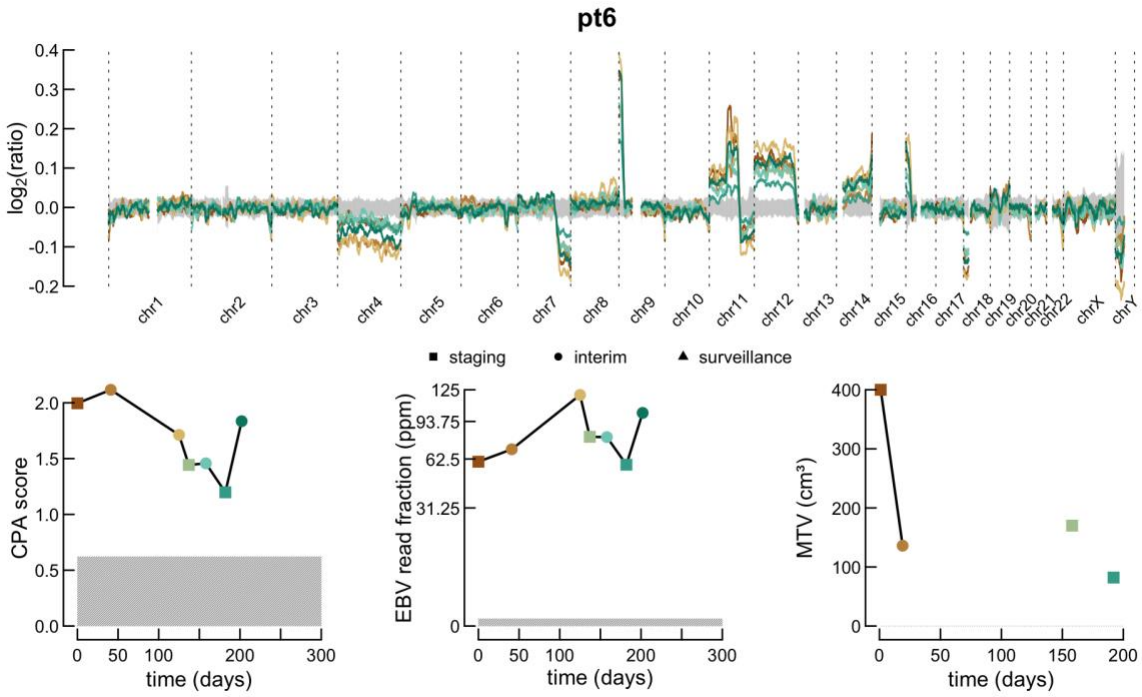
Longitudinal samples

Figures below present a copy number profile comparison between the longitudinal liquid biopsies of the named patient. The smoothened profiles at each point in time is given as a graph line, where the grey background represent the 95% confidence interval of healthy control profiles. The line graphs are colored and their matching timepoint can be found in the three plots below, indicating CPA score, EBV read fraction and MTV according to PET/CT.

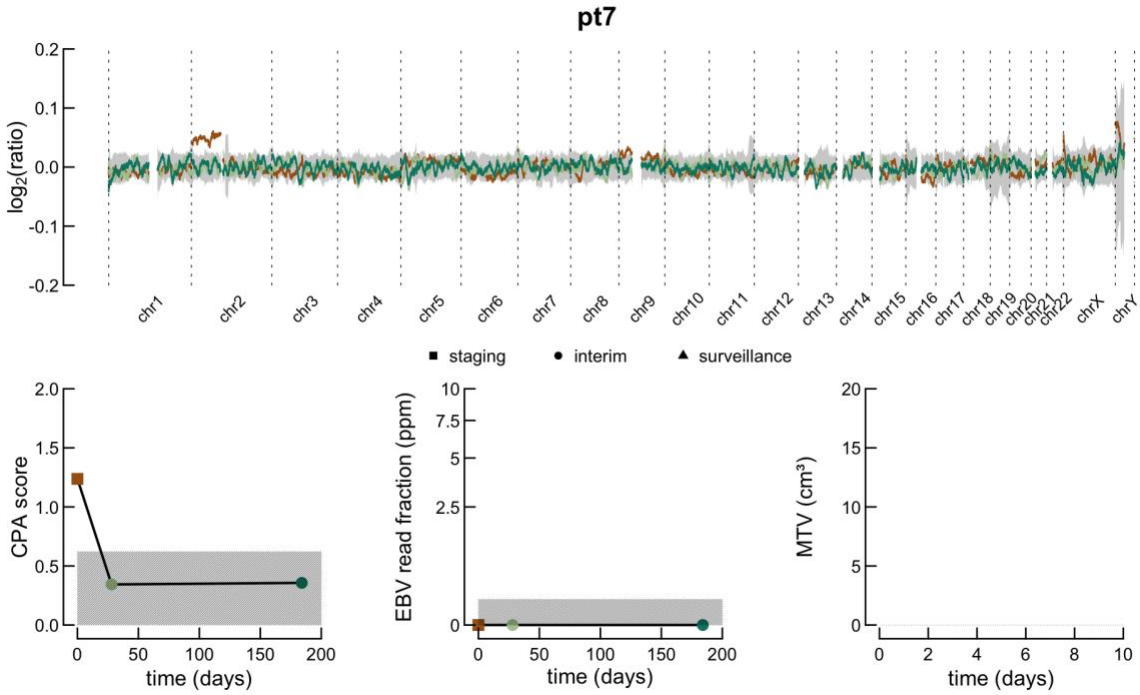
Remark: MTV data is missing when …

- There was no PET/CT scan executed.
- There was a larger than 30-day interval between PET/CT scanning and collecting blood samples.
- The PET/CT could not be accurately interpreted.
- Therapy (other than corticosteroids) was administered between PET/CT scanning and collecting blood samples.

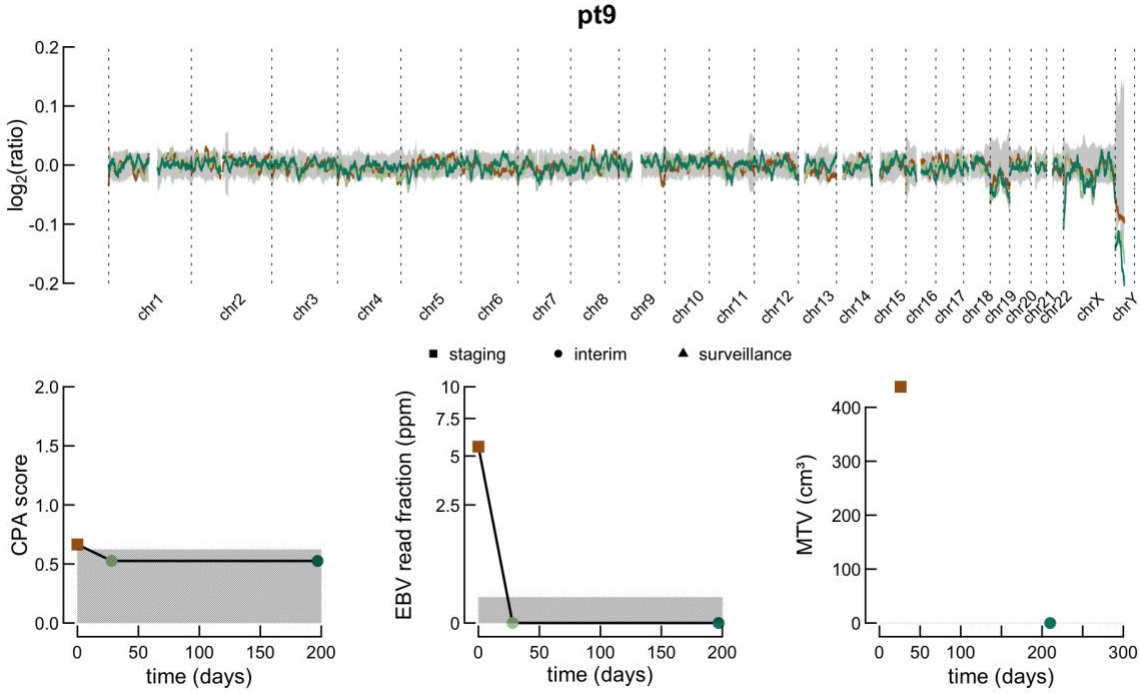
Patient 6



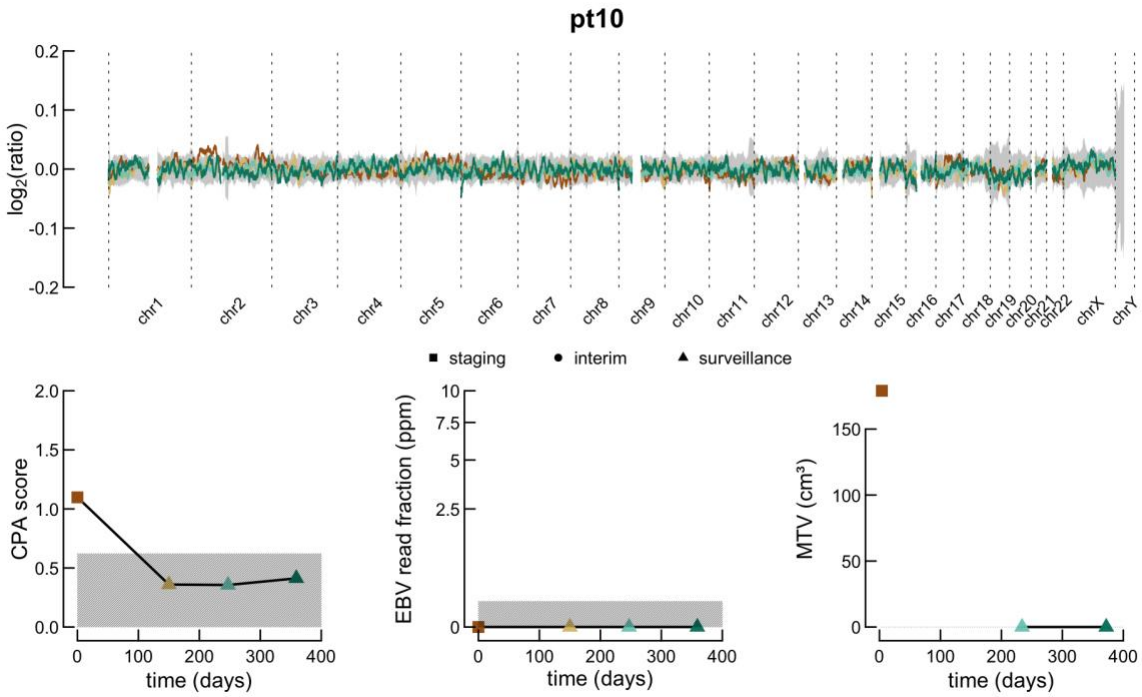
Patient 7



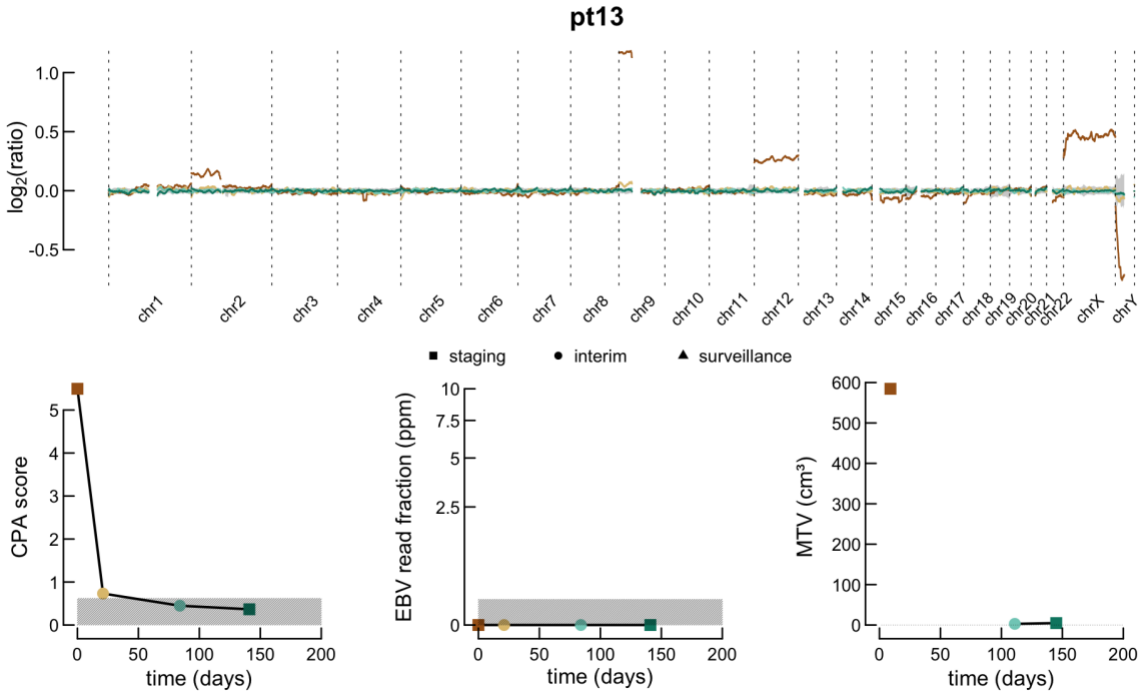
Patient 9



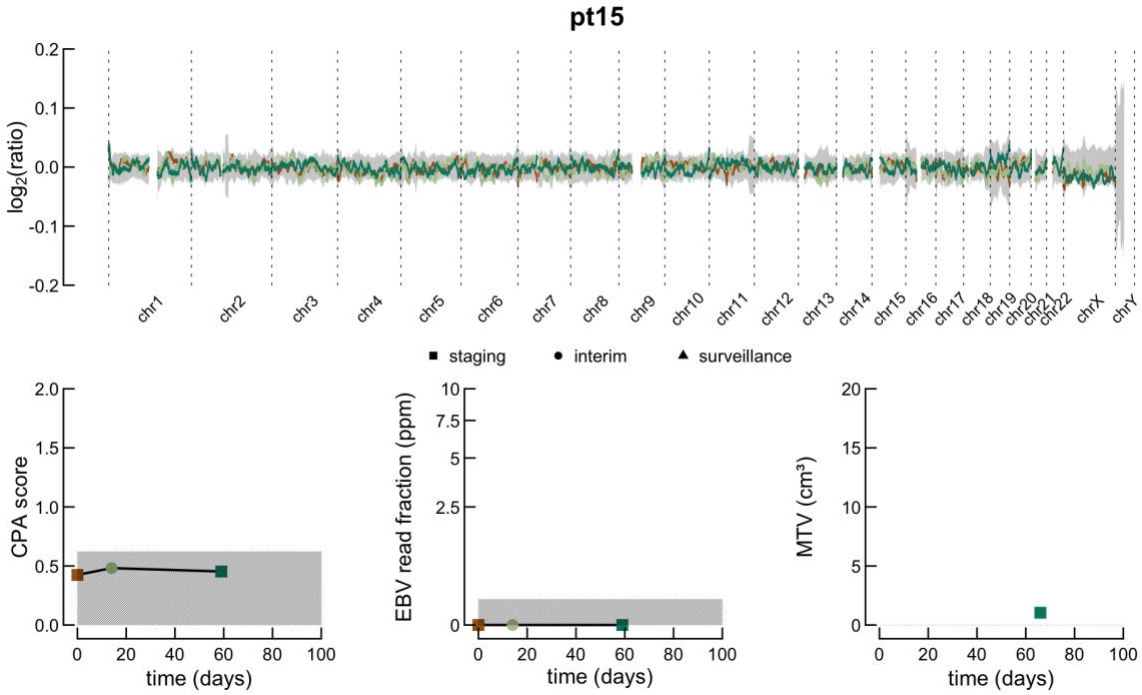
Patient 10



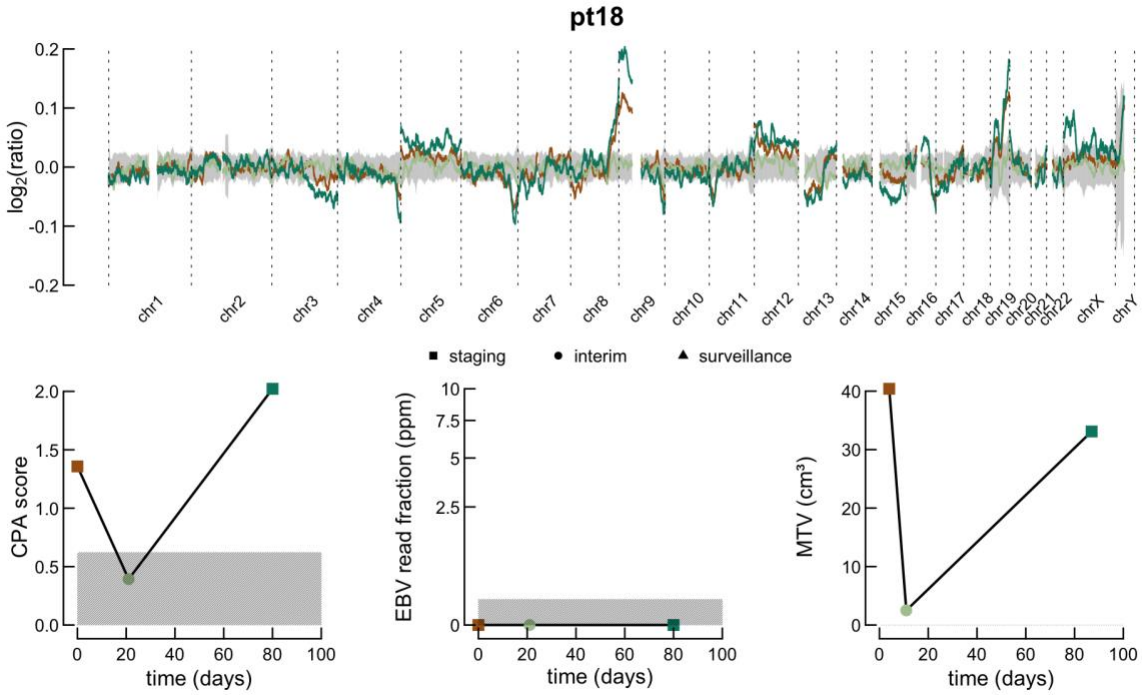
Patient 13



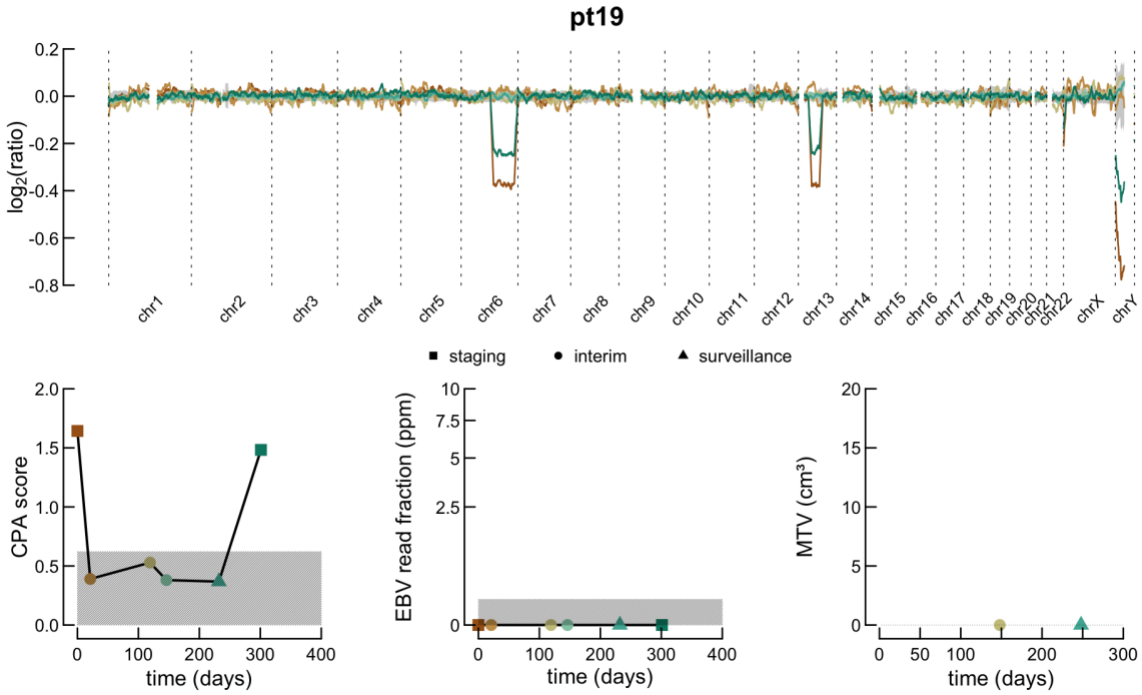
Patient 15



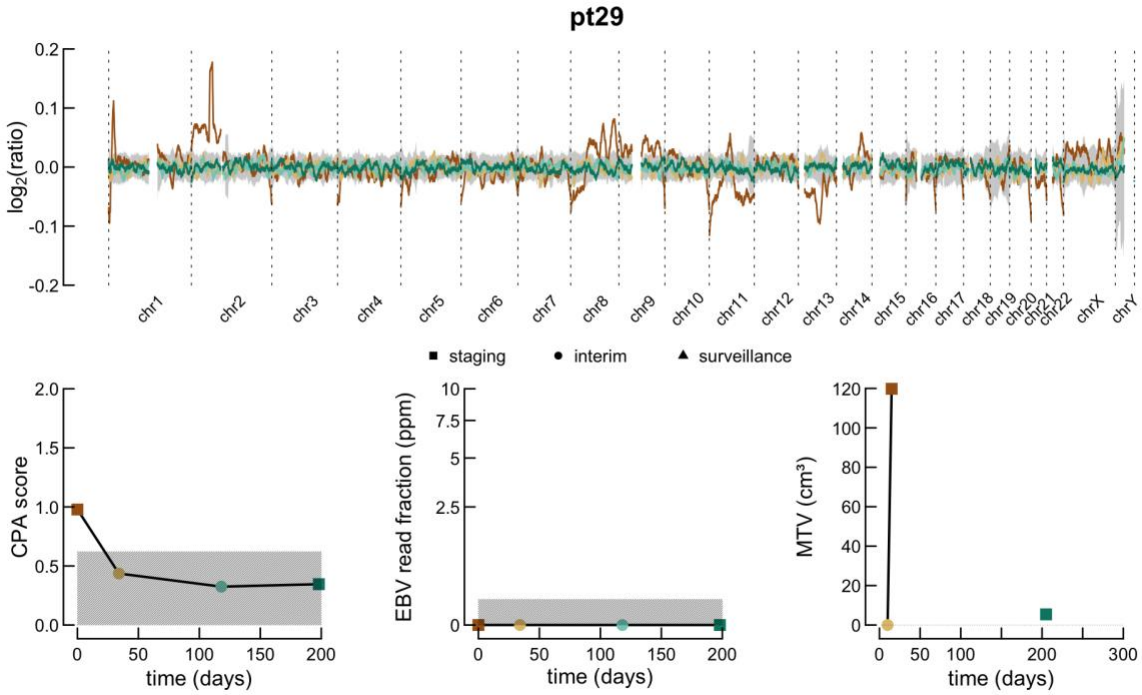
Patient 18



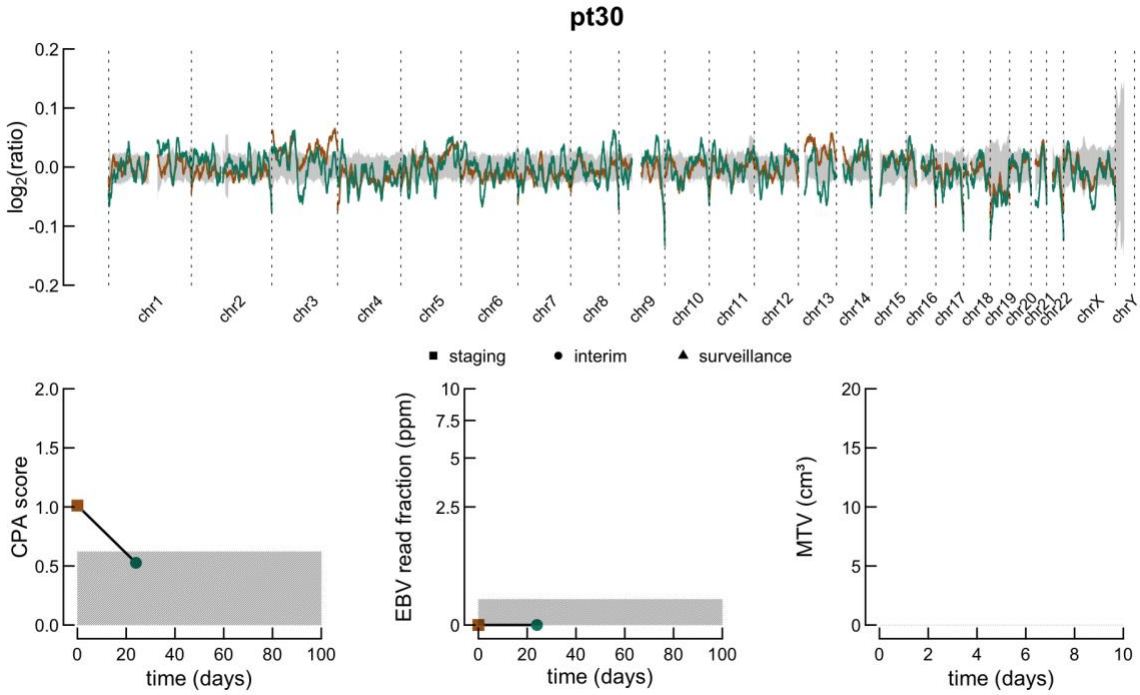
Patient 19



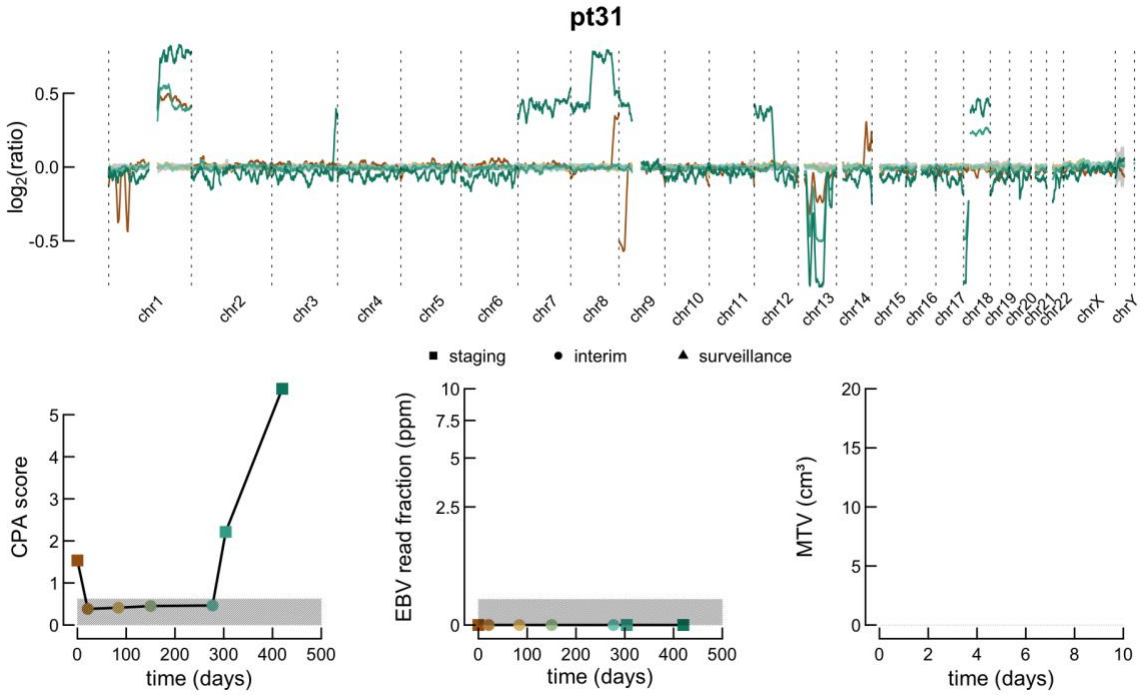
Patient 29



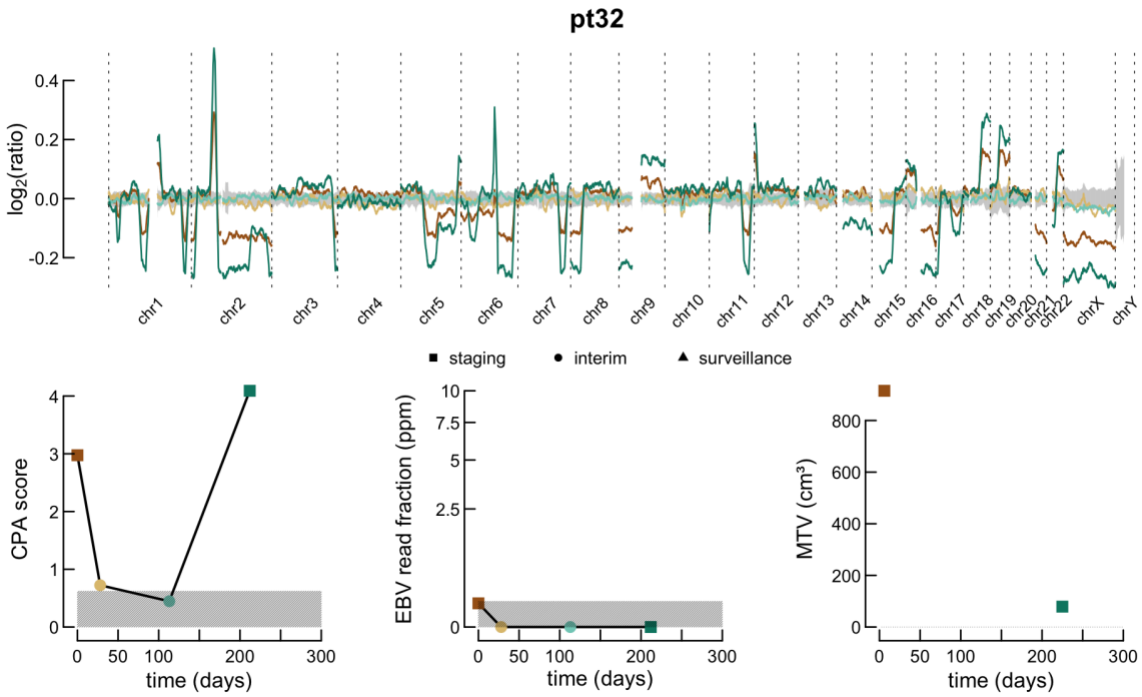
Patient 30



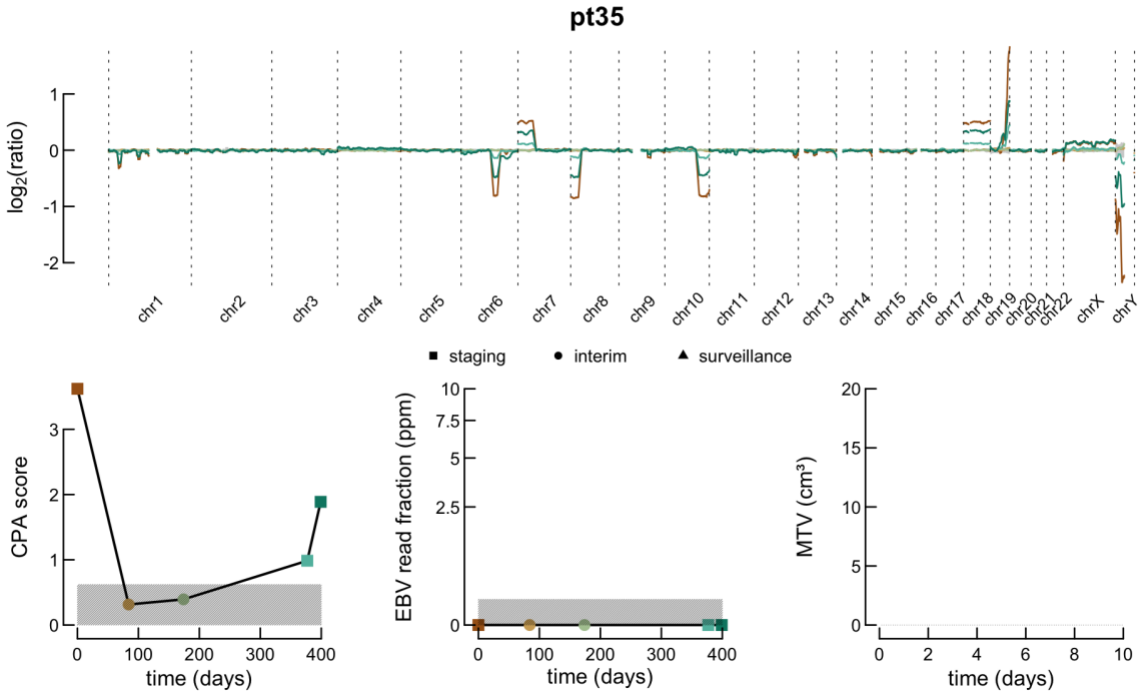
Patient 31



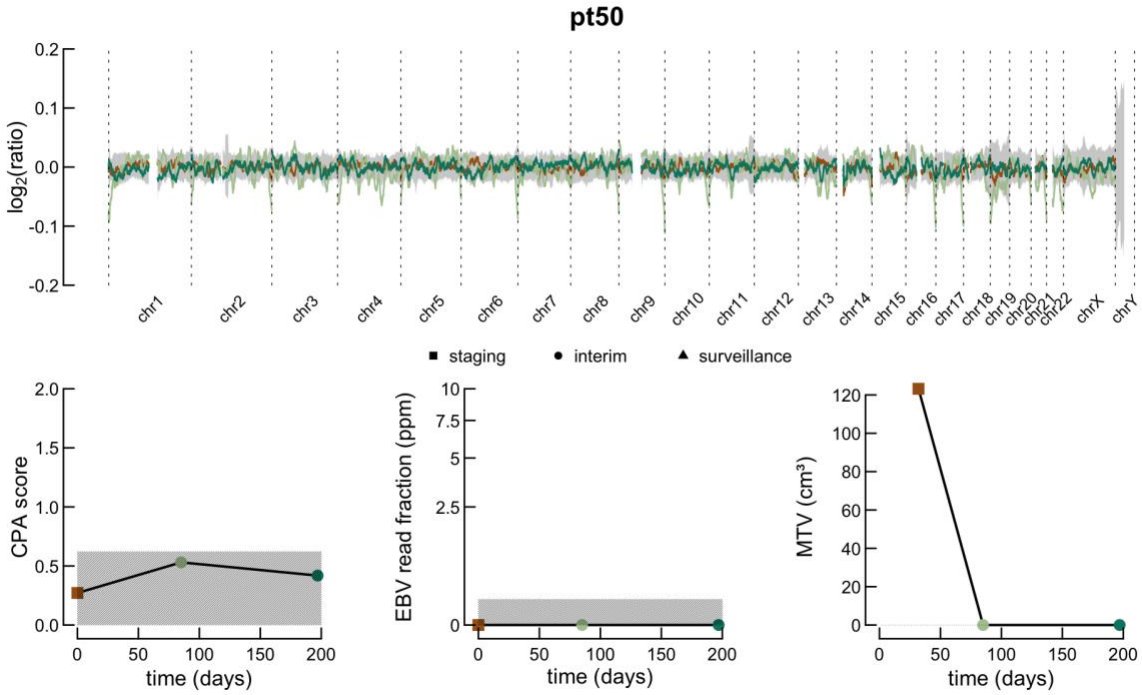
Patient 32



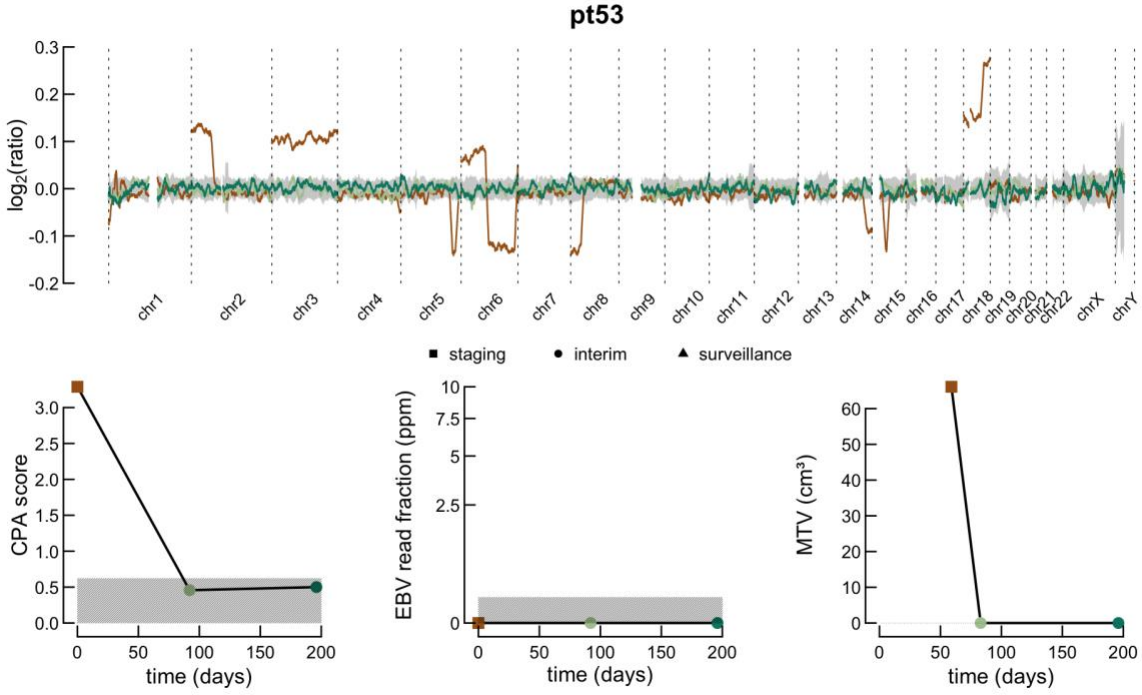
Patient 35



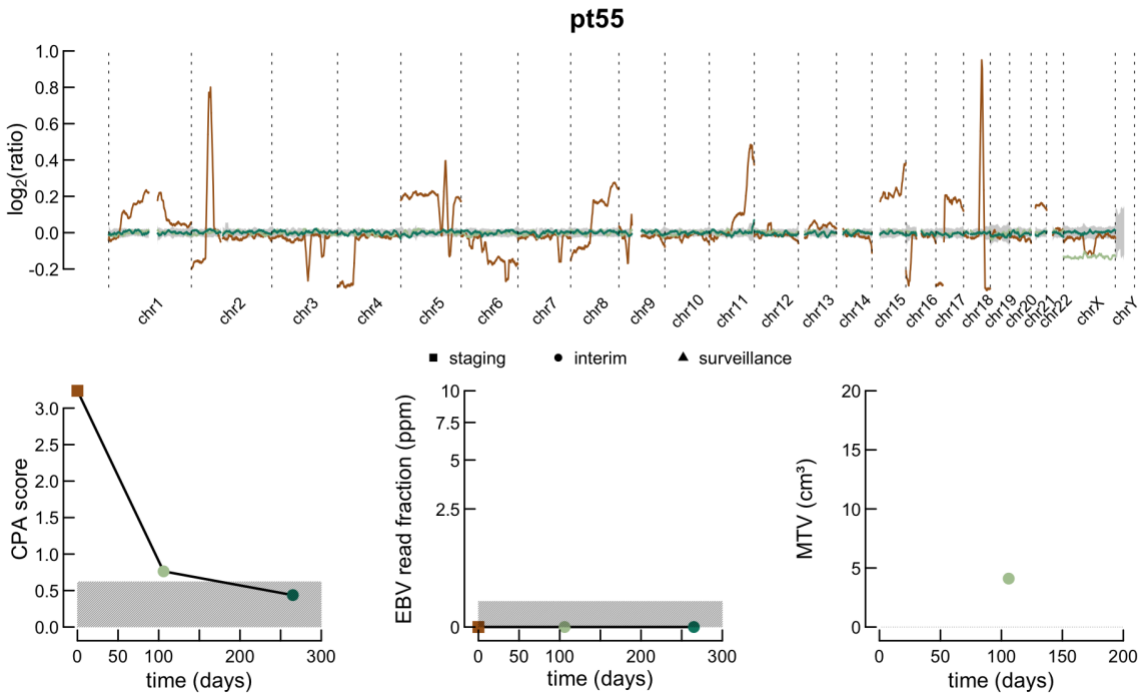
Patient 50



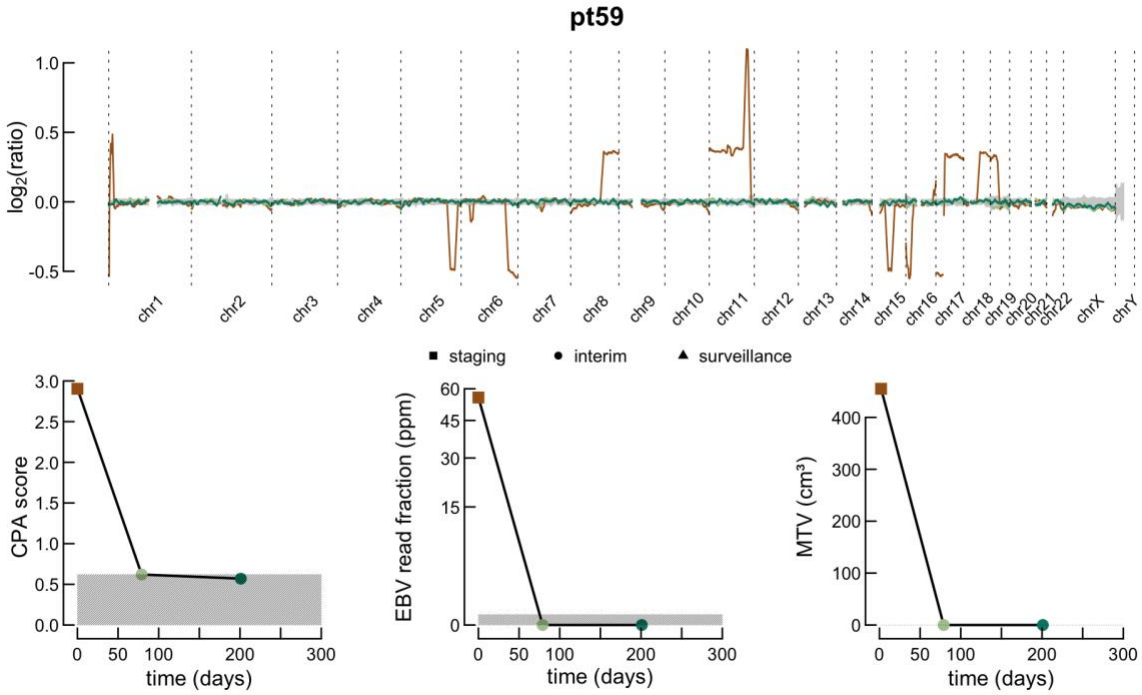
Patient 53



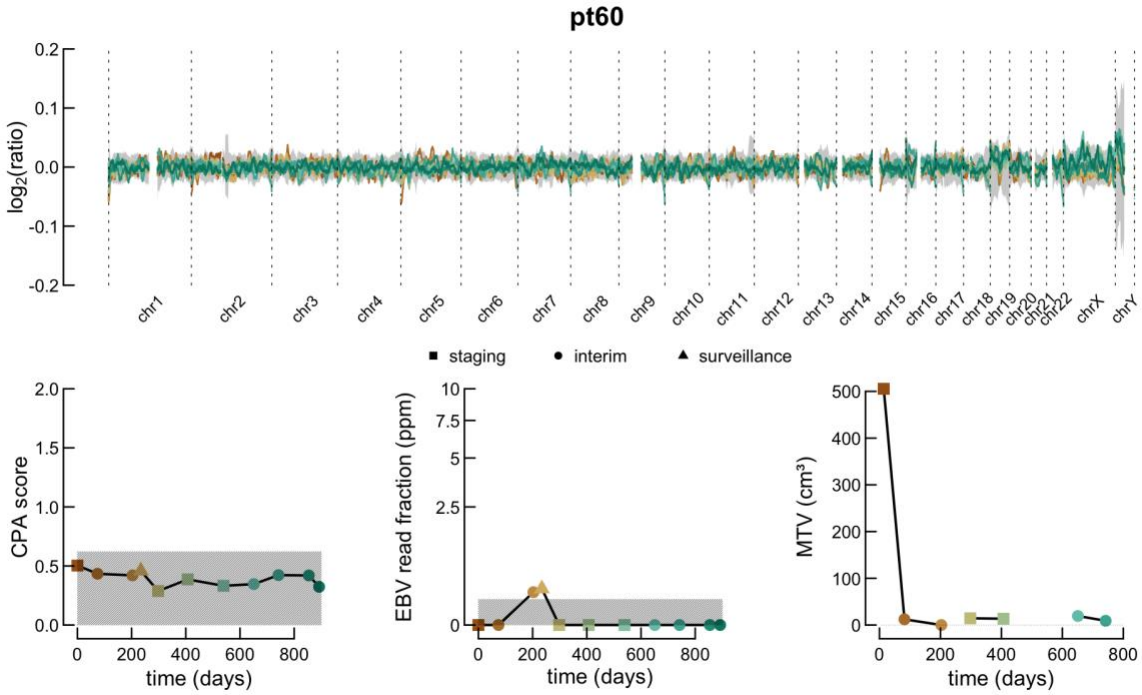
Patient 55



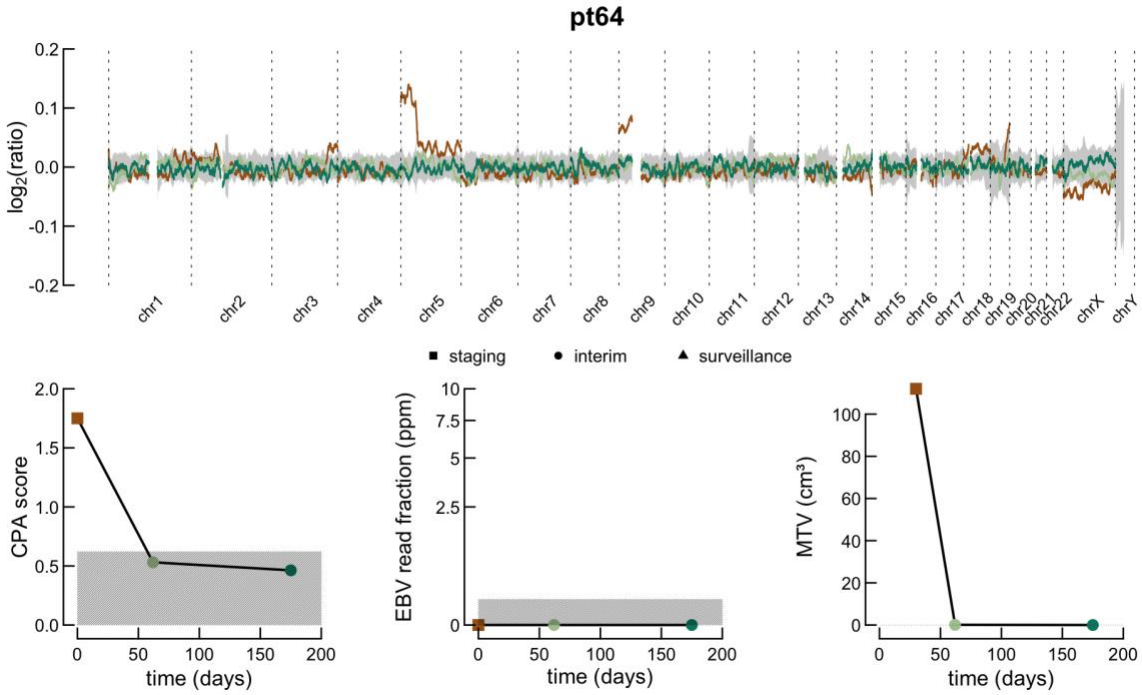
Patient 59



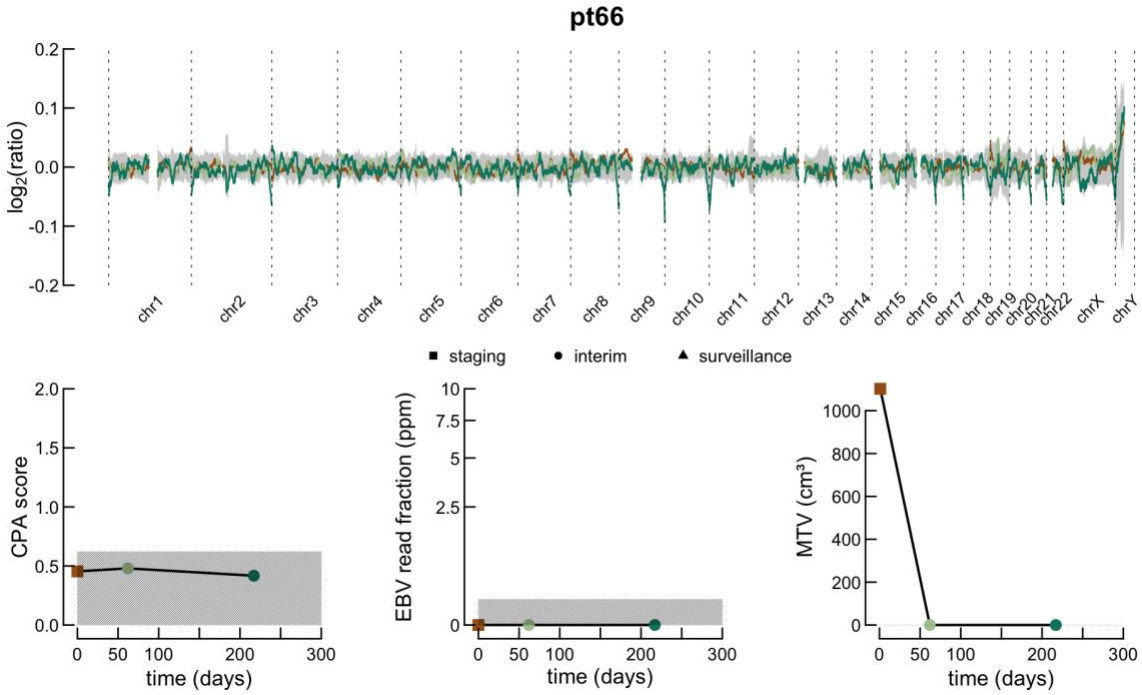
Patient 60



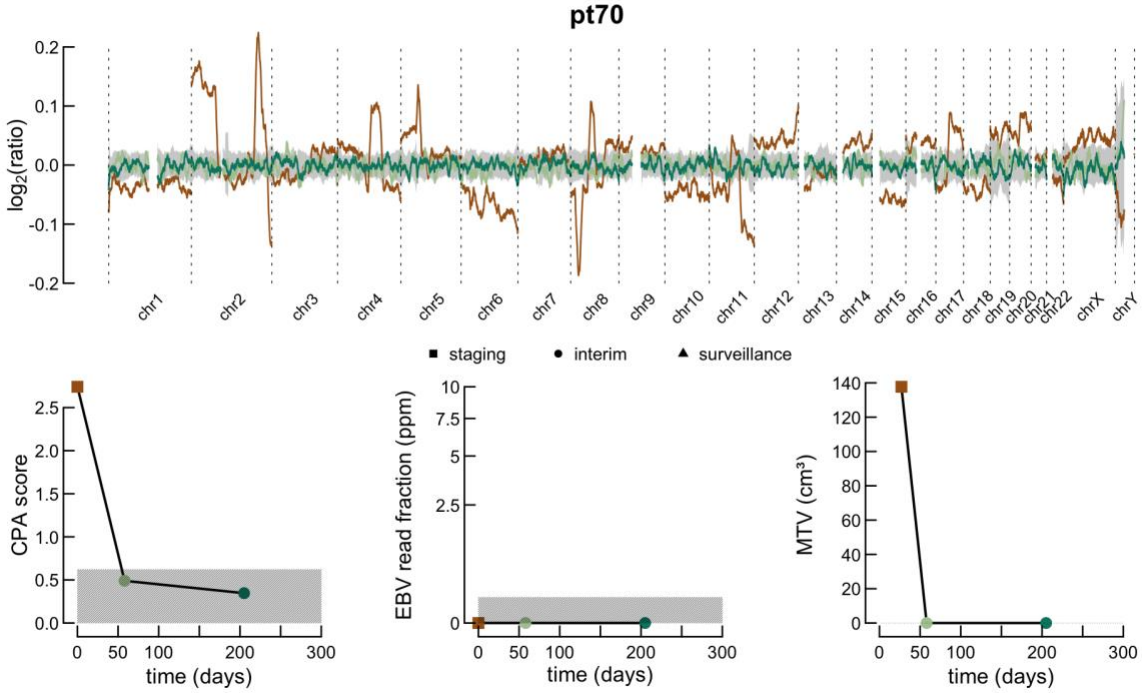
Patient 64



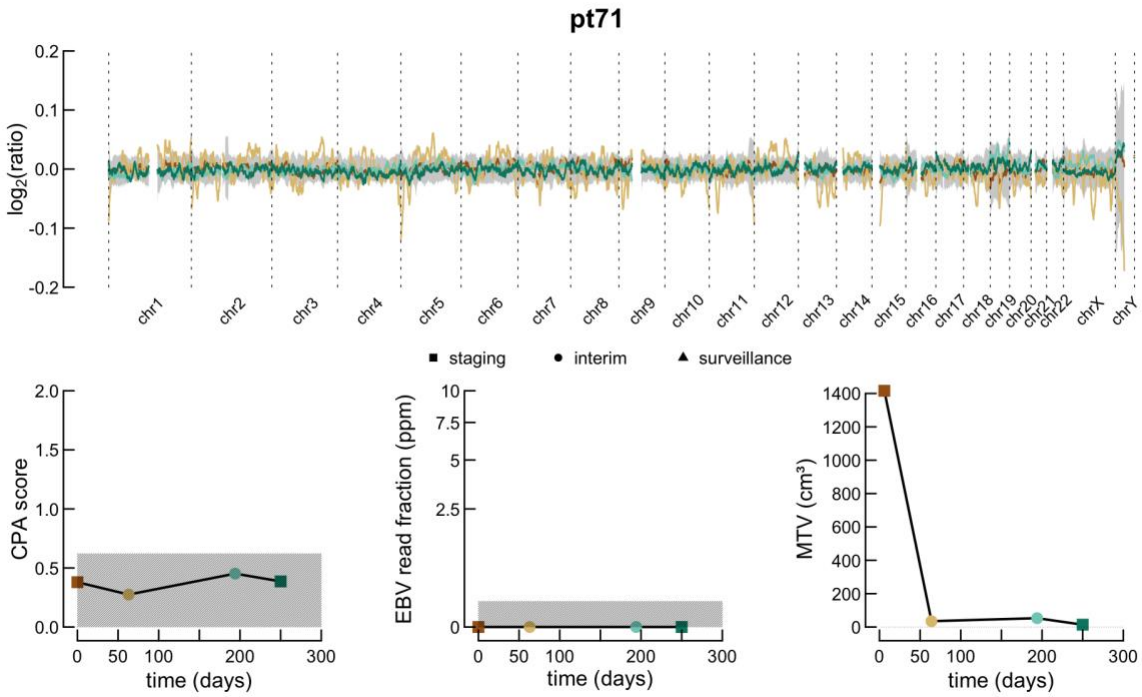
Patient 66



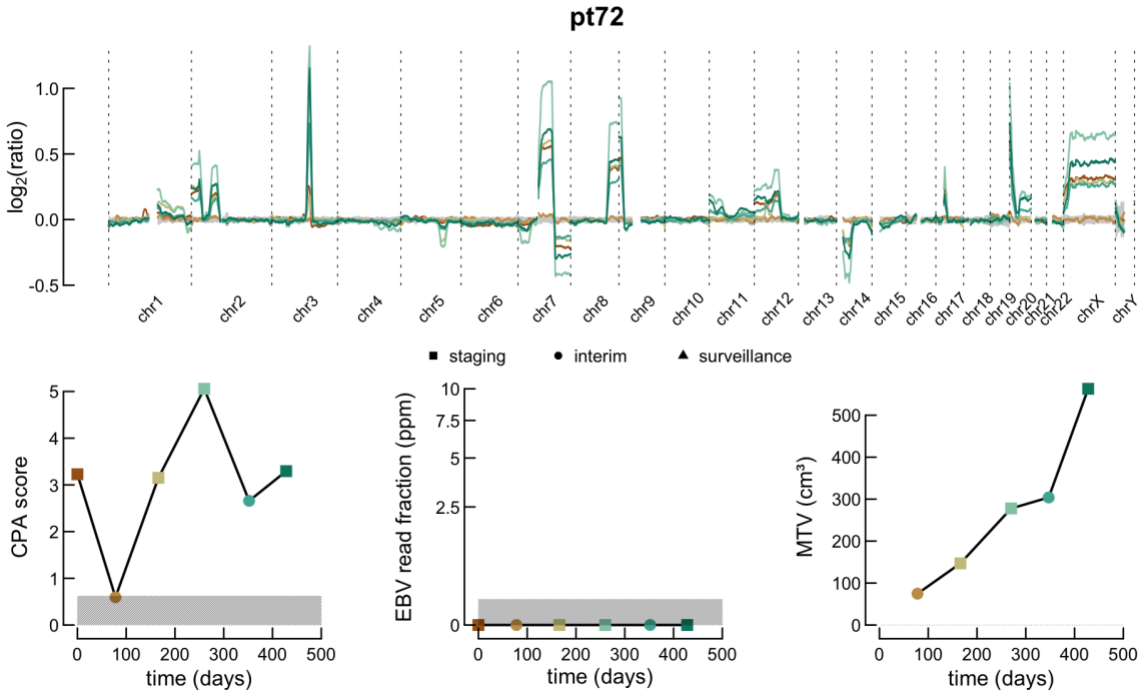
Patient 70



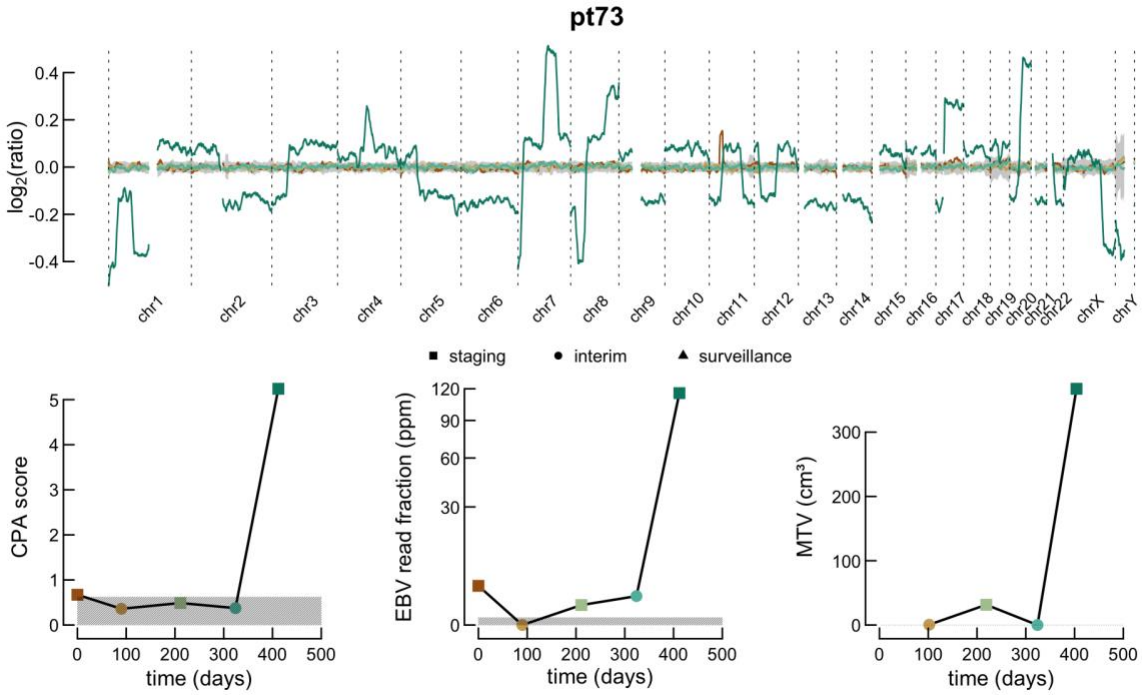
Patient 71



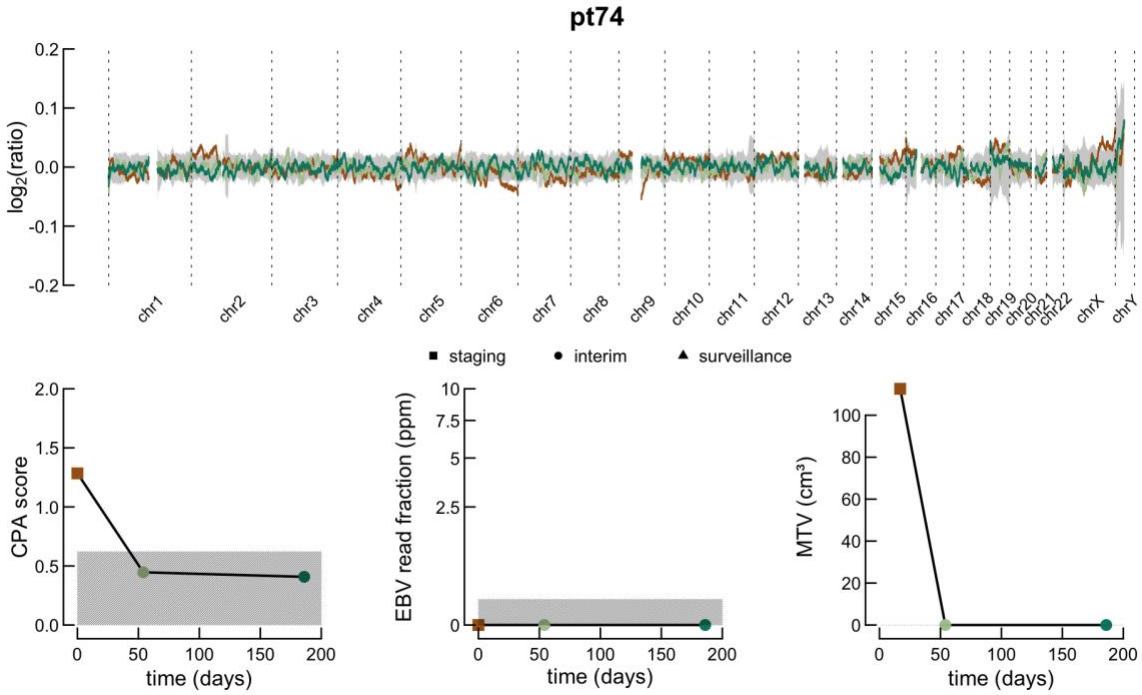
Patient 72



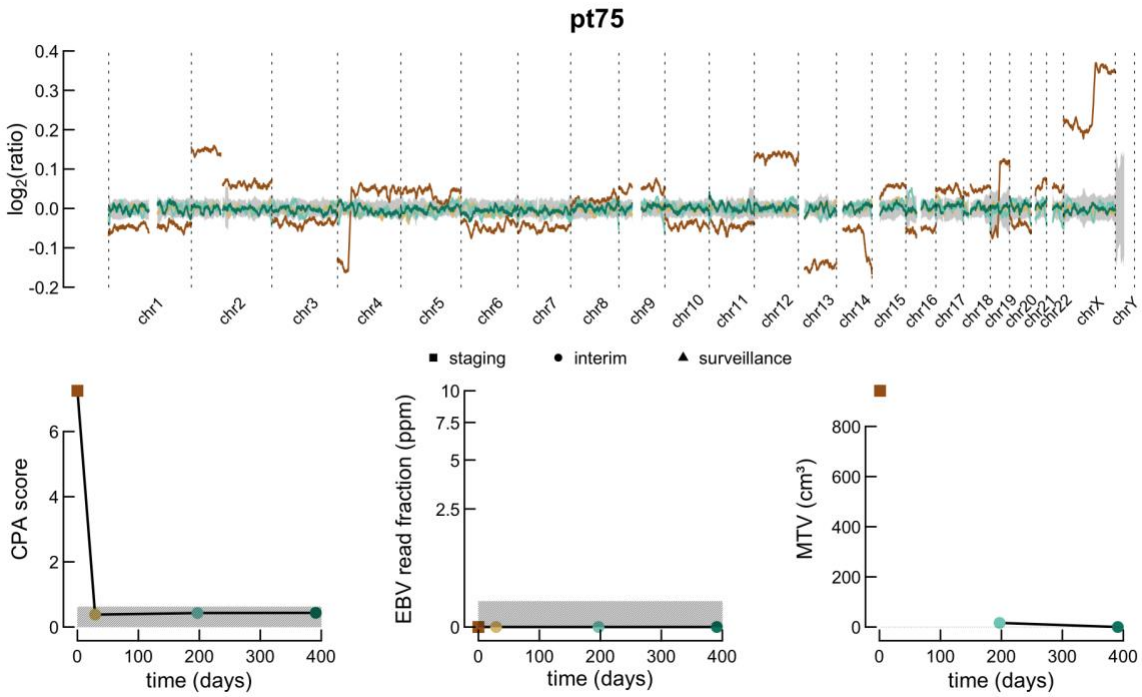
Patient 73



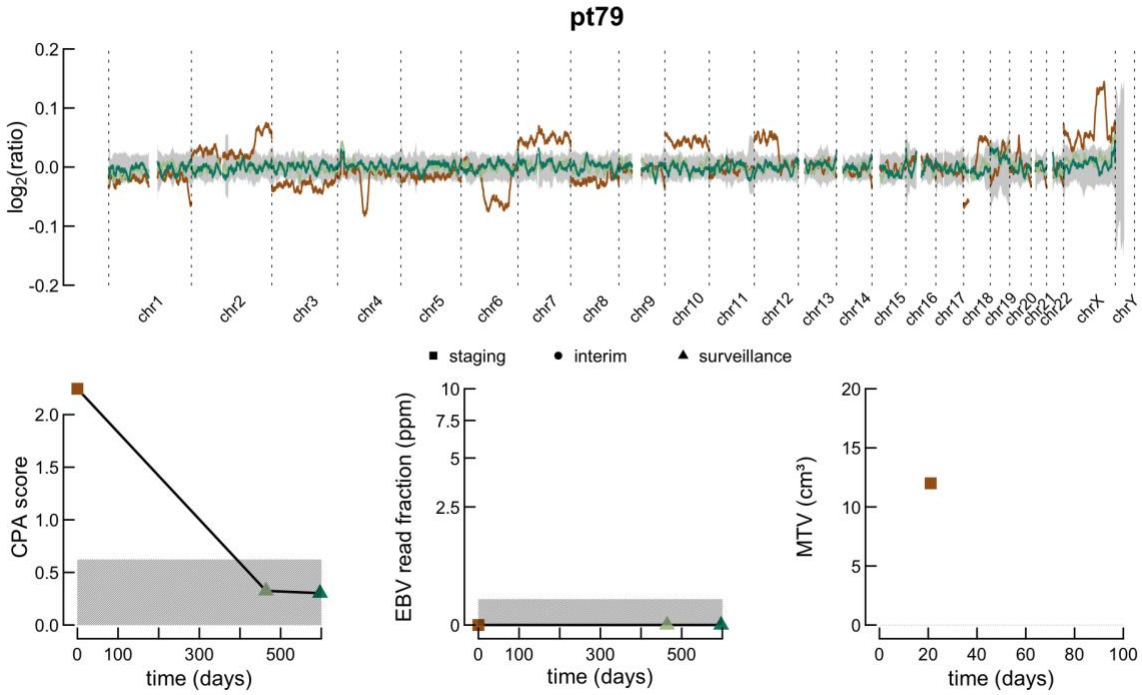
Patient 74



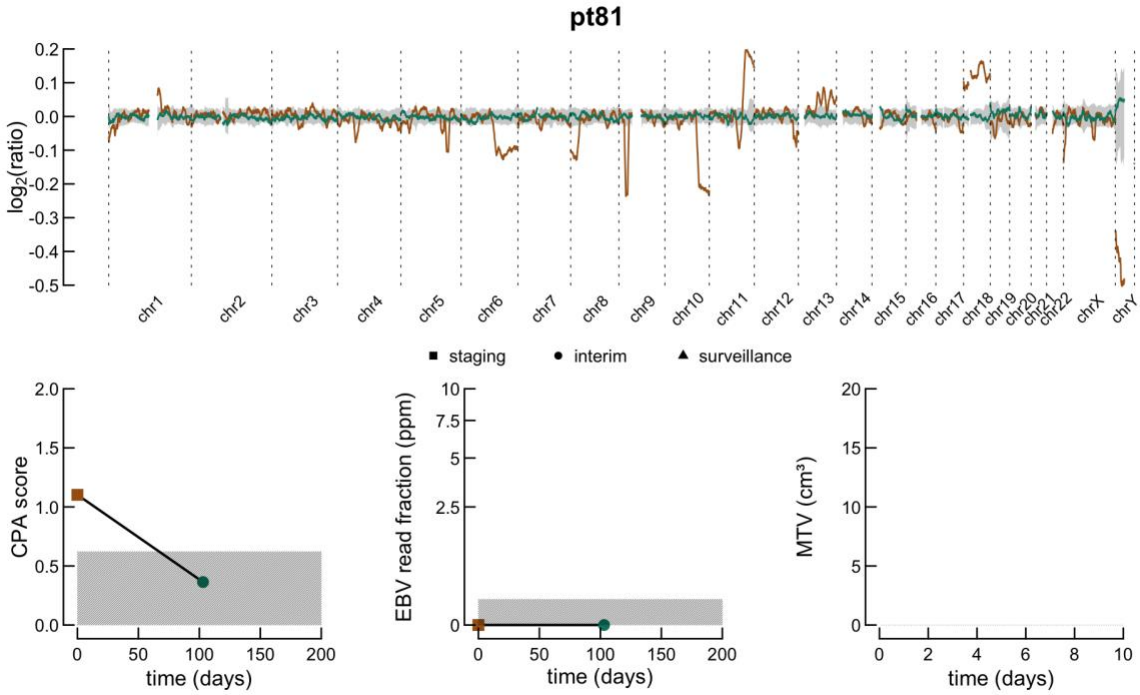
Patient 75



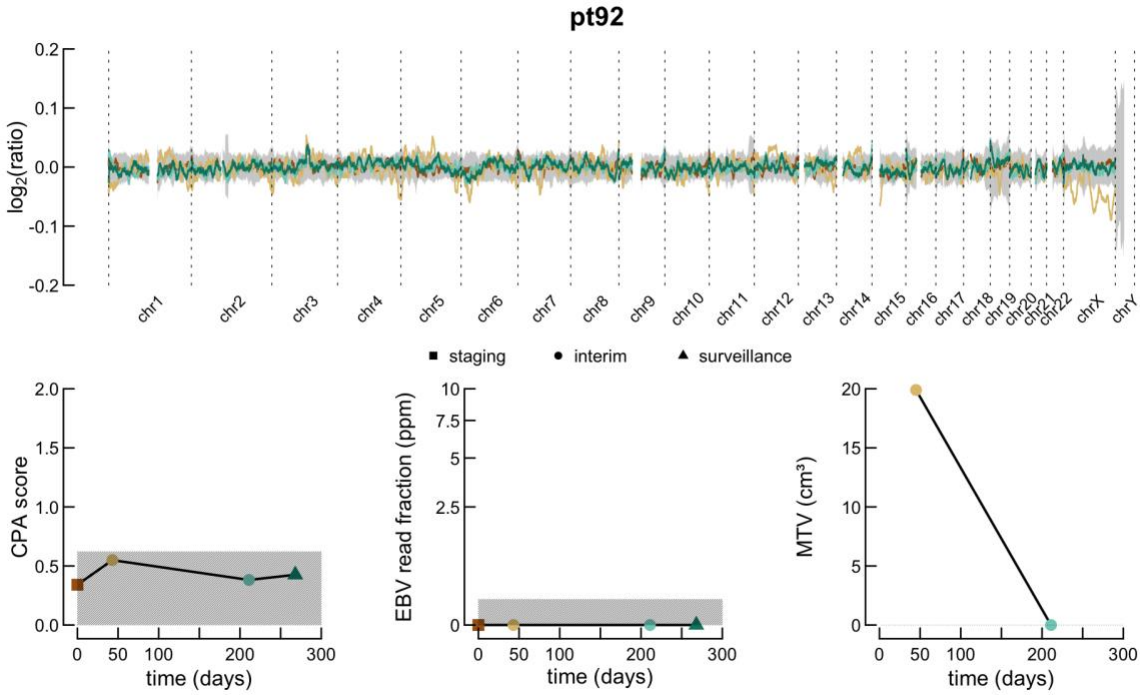
Patient 79



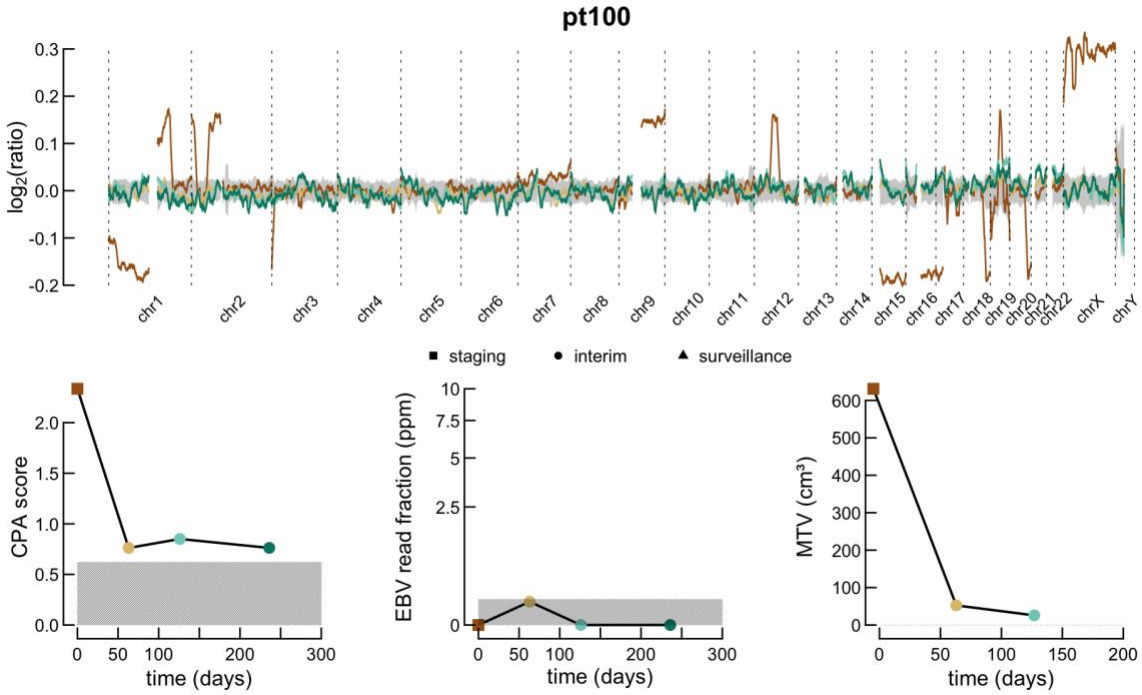
Patient 81



Patient 92



Patient 100



Patient 108