
Supplementary information

Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer

In the format provided by the
authors and unedited

Supplementary information to:

Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer

Cameron Herberts^{†1}, Matti Annala^{†1,2}, Joonatan Sipola^{†2}, Sarah W. S. Ng¹, Xinyi E. Chen¹, Anssi Nurminen², Olga V. Korhonen², Asli D. Munzur¹, Kevin Beja¹, Elena Schönlaui¹, Cecily Q. Bernales¹, Elie Ritch¹, Jack V. W. Bacon¹, Nathan A. Lack^{1,3,4}, Matti Nykter², Rahul Aggarwal^{5,6}, Eric J. Small^{5,6}, Martin E. Gleave¹, SU2C/PCF West Coast Prostate Cancer Dream Team, David A. Quigley^{5,7,8}, Felix Y. Feng^{5,6,7,9*}, Kim N. Chi^{1,10,*}, Alexander W. Wyatt^{1,11*}

¹Vancouver Prostate Centre, Department of Urologic Sciences, University of British Columbia, British Columbia, Canada; ²Prostate Cancer Research Center, Faculty of Medicine and Health Technology, Tampere University and Tays Cancer Center, Tampere, Finland; ³School of Medicine, Koç University, Istanbul, Turkey;

⁴Koç University Research Centre for Translational Medicine, Koç University, Istanbul, Turkey; ⁵Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, United States; ⁶Division of Hematology and Oncology, Department of Medicine, University of California San Francisco, San Francisco, California, United States; ⁷Department of Urology, University of California San Francisco, San Francisco, California, United States; ⁸Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States; ⁹Department of Radiation Oncology, University of California San Francisco, San Francisco, California, United States; ¹⁰Department of Medical Oncology, BC Cancer, Vancouver, British Columbia, Canada; ¹¹Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, British Columbia, Canada

¹2660 Oak St, Vancouver, BC V6H 3Z6, Canada; ²Kauppi Campus, Arvo Building, Arvo Ylpön katu 34, 33520 Tampere; ^{3,4}Rumelifeneri, Sarıyer Rumeli Feneri Yolu, 34450 Sarıyer/Istanbul, Turkey; ⁵1450 Third St. Room 387, San Francisco, CA, 95158; ^{6,7}400 Parnassus Ave San Francisco, CA 94143, United States; ⁸550 16th St 2nd floor, San Francisco, CA 94158, United States; ⁹505 Parnassus Ave, San Francisco, CA 94143, United States; ¹⁰600 W 10th Ave, Vancouver, BC V5Z 4E6; ¹¹570 W 7th Ave, Vancouver, BC V5Z 4S6

†These authors contributed equally to this work.

***These authors have jointly supervised this work.**

Correspondence to Dr. Alexander W. Wyatt, Vancouver Prostate Centre, Department of Urologic Sciences, University of British Columbia, 2660 Oak Street, Vancouver, BC, V6H 3Z6, Canada; Tel: +1-604-875-4818; Fax: +1-604-875-5654; Email: awwyatt@mail.ubc.ca

Supplementary Note

Stand Up 2 Cancer (SU2C) / Prostate Cancer Foundation (PCF) West Coast Prostate Cancer Dream Team

Joshi J. Alumkal¹; Tomasz Beer²; Kim N. Chi^{3,4}; Christopher Evans⁵; Felix Y. Feng^{6,7,8,9}; Adam Foye¹⁰; Martin E. Gleave³; Jiaoti Huang¹¹; Primo Lara¹²; David A. Quigley^{6,8,13}; Robert Reiter¹⁴; Matthew Rettig¹⁵; George Thomas¹⁶; Lawrence True¹⁷; Owen Witte¹⁸

¹Department of Medicine, University of Michigan, Ann Arbor, Michigan, United States; ²Knight Cancer Institute, School of Medicine, Oregon Health & Science University, Portland, Oregon, United States; ³Vancouver Prostate Centre, Department of Urologic Sciences, University of British Columbia, British Columbia, Canada; ⁴Department of Medical Oncology, BC Cancer, Vancouver, British Columbia, Canada; ⁵Department of Urology, School of Medicine, University of California, Davis, Sacramento, California, United States; ⁶Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, United States; ⁷Division of Hematology and Oncology, Department of Medicine, University of California San Francisco, San Francisco, California, United States; ⁸Department of Urology, University of California San Francisco, San Francisco, California, United States; ⁹Department of Radiation Oncology, University of California San Francisco, San Francisco, California, United States; ¹⁰University of California San Francisco, San Francisco, California, United States; ¹¹Department of Pathology, Duke University, Durham, North Carolina, United States; ¹²University of California, Davis, Sacramento, California; ¹³Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States; ¹⁴Department of Urology, University of California, Los Angeles, California, United States; ¹⁵Division of Hematology and Oncology, David Geffen School of Medicine, University of California, Los Angeles, California, United States & Division of Hematology and Oncology, VA Greater Los Angeles Healthcare System, Los Angeles, California, United States; ¹⁶Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, United States; Department of Pathology, Oregon Health and Science University, Portland, Oregon, United States; ¹⁷Department of Pathology, University of Washington, Seattle, WA, United States; ¹⁸Department of Microbiology, Immunology, and Molecular Genetics at the David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States.

SU2C / PCF West Prostate Cancer Coast Dream Team representative contact: Dr. Felix Y. Feng, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, 1450 3rd St, Box 3110, San Francisco, CA 94158, USA. Tel. +1 415 502 6355. Email: Felix.Feng@ucsf.edu

Supplementary Methods

Table of Contents

Copy number model fitting and cancer fraction estimation	4
Detection of whole genome duplication	8
Quantifying chromosomal instability and subclonal copy number heterogeneity	11
Quantifying subclone fractions	11
Inference of phylogenetic trees	15
Reconstruction of subclone specific copy number profiles	16
Visualization of tumor burden and subclone fractions over time	17
Mutation-based timing of whole-genome duplication	19
Estimation of metastatic tissue contribution to ctDNA	20
Detection of subclonal whole genome duplication events	21
Evidence for parallel evolution of whole genome duplication	25
Identification of somatic mutations that expanded during treatment	26
Identification of rearrangements that expanded during treatment	27
Selection of patient samples with DNA repair defects	27
Validation	28
Validation of truncal copy number models	28
Validation of somatic mutation calls using GATK Mutect2	28
Validation of structural variants using GRIDSS2	31
Validation of subclonal reconstruction using PyClone and PhyloWGS	31
Validation of mutation CCF clusters using a generative simulation	33
Validation of subclonal reconstruction using a blinded simulation	33
Validation of mutational signature decomposition on bulk data using SigProfileExtractor	40
Validation of evolutionary changes in mutational signature activity using TrackSigFreq	42
Validation of sample identities via SNP profile comparison	45
Simulation of subclonal reconstruction in 5% ctDNA sample	46
Patient truncal copy number models and subclonal reconstructions	47
PyClone and PhyloWGS subclonal reconstruction validation data	222
Battenberg copy number models	233
Supplementary references	251

Copy number model fitting and cancer fraction estimation

Our methodology for cancer fraction and copy number estimation was inspired by contemporary approaches¹⁻³. We divided the genome into regularly spaced 1000 bp windows (with 50% overlap between consecutive windows) and calculated the number of aligned DNA fragments within each window using bedtools-2.25.0⁴. A set of 5 whole genome sequenced ctDNA-negative cfDNA samples and 34 leukocyte DNA samples were used to generate two reference coverage profiles for copy number analysis (one for cfDNA analysis and one for tissue DNA analysis). To generate a reference coverage profile, we calculated the median positional coverage across all reference samples after removing cross-sample differences in GC bias and global sequencing depth. To remove GC bias, we calculated coverage logratios $L_k = \log_2(R_{k,s} / R_{k,1})$ for each reference sample against an index sample (arbitrarily selected as the first reference sample), where $R_{k,s}$ represented the read count within genomic window k in sample s . We then fitted a Loess regression curve to a two-dimensional scatter plot of (G_k, L_k) pairs — where G_k was the GC fraction of genomic window k — and subtracted the Loess curve from the logratio values to correct for GC bias and differences in global sequencing depth. We then inverted the above formula and converted each sample's GC-corrected logratios back to read counts. The reference coverage profile was calculated as the median GC-corrected coverage across all reference samples.

To calculate positional coverage logratios for a tissue or cfDNA sample containing cancer, the sample's read count within each genomic window was divided by the corresponding read count in the reference coverage profile. The resulting coverage logratios were corrected for GC bias using Loess regression using the same approach as above.

To identify genomic regions with distinct copy number states, we searched for coverage logratio discontinuities (i.e. breakpoints) across all samples from a given patient. We used a segmentation procedure incorporating information from all same-patient samples. Each chromosome was represented as a sequence of coverage logratios $L_1, \dots, L_n$ ordered by position. To identify a breakpoint a , we searched for integers a and b defining a region $L_a, \dots, L_b$ such that the statistical difference relative to region $L_1, \dots, L_{a-1}$ was maximized. The difference was defined as $Z = (\mu_{1..a} - \mu_{a..b}) / \sqrt{1/a + 1/(b-a)}$, where $\mu_{a..b} = \text{mean}(L_a, \dots, L_b)$. For each chromosome, the breakpoint position a that produced the highest Z score in any sample from the patient was identified, and the chromosome was split into two segments if the Z score exceeded a specific cohort-wide threshold. This process was repeated recursively for each resulting segment. To account for variable levels of coverage logratio noise between samples, the Z scores calculated for each sample were divided by a sample noise score N , defined as the average difference between consecutive logratios.

To calculate a heterozygous SNP allele fraction (HSAF) track encompassing the entire genome, we identified heterozygous SNPs in each patient based on their leukocyte DNA sample, and calculated the allele fraction of these heterozygous SNPs in their cfDNA/tumor samples. Somatic copy number alterations in cancer samples can cause the parental and maternal alleles of a heterozygous SNP to become unbalanced, resulting in a detectable deviation from the expected 50% allele fraction seen in normal cells. Since a $(50 + X)\%$ allele fraction is equivalent to a $(50 - X)\%$ allele fraction of the other allele, we transformed the HSAF values to fall within the range $[0.5, 1]$ using the formula $0.5 + \text{abs}(\text{HSAF} - 0.5)$. This allowed direct calculation of segment-level median HSAF values. Haplotype information was not utilized.

Coverage logratios contain information about the relative copy number of different genomic regions, but do not directly measure absolute copy number, as they are confounded by other factors, such as the total number of sample DNA fragments sequenced, the fraction of normal cell DNA in a sample, and the presence of whole genome duplication (which can uniformly increase the copy number of all genomic regions). In samples composed of a homogeneous cancer cell population and a normal cell population at fractions F and $1 - F$, the coverage logratio of a genomic segment can be modeled as

$$LR = \log_2 \frac{FC + (1-F)P}{P} + D \text{ where}$$

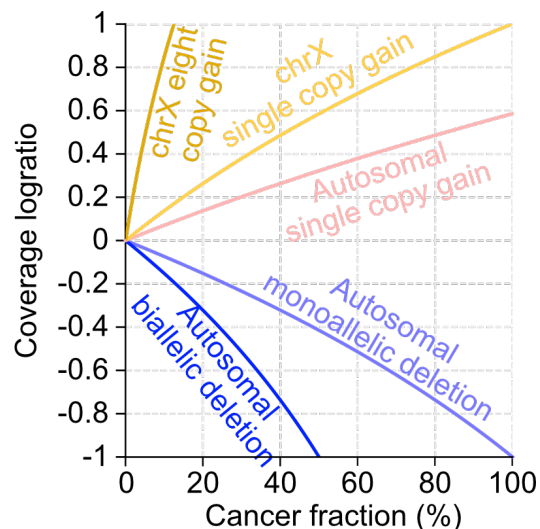
F = cancer fraction in sample

C = copy number in cancer cells

P = copy number in normal cells

D = diploid level (coverage logratio at $C=2$)

$D=0$



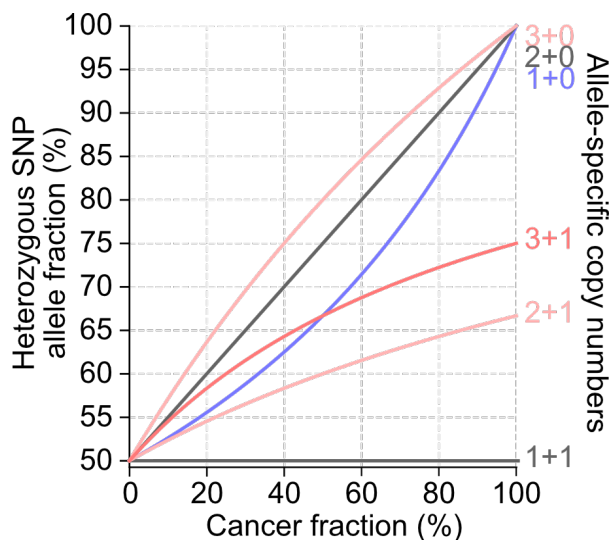
For a heterozygous SNP with allele-specific copy numbers A and B (such that $A+B=C$) in a cfDNA/tumor sample, the B allele fraction can be modeled as

$$HSAF = \frac{BF + (1-F) \cdot 1}{F(A+B) + (1-F) \cdot 2} \text{ where}$$

F = cancer fraction in sample

A = copy number of allele A in cancer cells

B = copy number of allele B in cancer cells



For a somatic mutation present in a sample's cancer cells, the mutation allele fraction can be modeled as

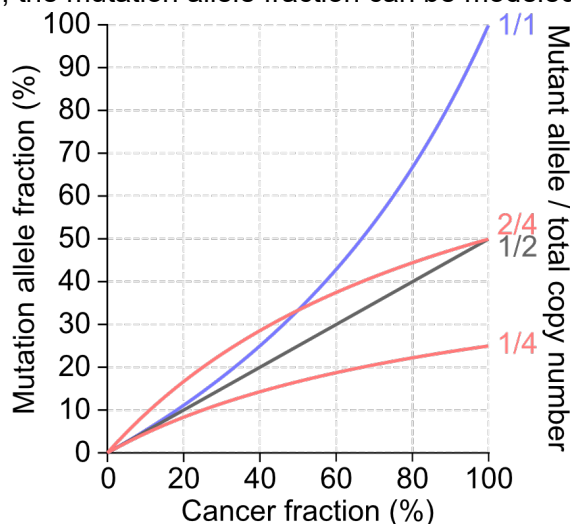
$$MAF = \frac{MF}{F(M+N) + (1-F) \cdot P} \text{ where}$$

F = cancer fraction in sample

M = mutant allele copy number in cancer cells

N = wildtype allele copy number in cancer cells

P = wildtype allele copy number in normal cells



To infer the allele-specific copy number of genomic segments, we searched for a copy number model providing the best fit to the observed coverage logratios, HSAFs, and somatic mutation allele fractions (MAF) in autosomal regions. Each model had only two parameters: the cancer fraction F and diploid level D . There is no *a priori* knowledge about which coverage logratio level corresponds to which copy number state in a sample, so the diploid level must be inferred from the data. For each candidate model, we calculated the expected coverage logratios, HSAFs, and possible MAFs associated with different allele-specific copy number states (e.g. A, AB, AA, AAB, AAA, AABB, AAAB, AAAA, ..., where A and B are the two parental alleles). Information from chromosomes X and Y was not used in model fitting, because HSAF information is not available for these chromosomes in male patients.

In samples containing DNA from two or more cancer cell populations, genomic segments whose allele-specific copy number state differs between the populations will not align perfectly with a single-population model. Since each genomic segment contains multiple coverage logratio and HSAF measurements with noise, we modeled them as distributions. A genomic segment was considered to align with a truncal copy number model if there existed at least one allele-specific copy number state whose mathematically expected coverage logratio fell within the observed coverage logratio distribution's middle 70%, and expected HSAF fell within the observed HSAF distribution's middle 70%. Additionally, we calculated a raw copy number for the segment (as a real number), and required that this copy number had to be within 0.25 units of the nearest integer copy number state. We rejected models that required any cancer cell population to have more than 50 Mb of the genome at a zero copy state (due to biological implausibility), or any genomic segment at less than zero copies. Our automated procedure for copy number model fitting makes no assumptions about the number of distinct cancer cell populations comprising a sample. Instead, our procedure assumes that each sample has abundant genomic regions with no significant subclonal copy number heterogeneity (i.e. the regions have the same copy number state in all cancer cell populations present in the sample).

We searched for the best fitting model using an automated algorithm, which reports 1 - 3 models with different whole genome duplication states and suggests which one is most likely to be correct. First, the algorithm identifies candidate diploid level ranges enriched for genomic segments with balanced HSAFs close to 50%. The algorithm then tests all possible models with a cancer fraction between 0 - 100% and a diploid level within the candidate ranges, and finds a copy number model minimizing the error function

$$E = 1 / F + D + aM$$

where F is the fraction of segments aligning with the candidate truncal copy number model, and

$$D = \sum L_s \sqrt{(R_s - R_{\text{nearest}})^2 + (H_s - H_{\text{nearest}})^2}$$

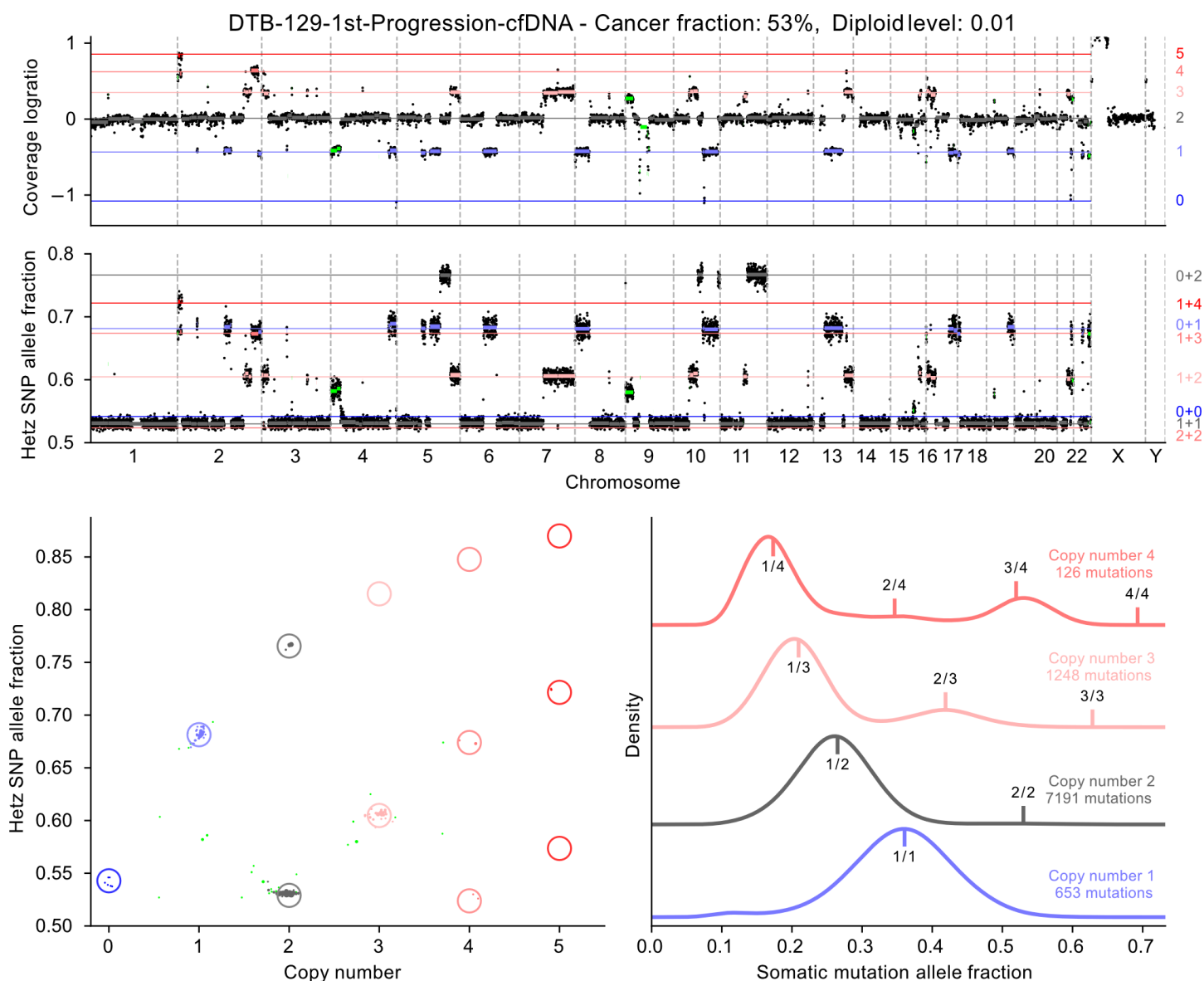
is the summed Euclidean distance of segment coverage logratios R_s and HSAFs H_s from the nearest truncal copy number state in the candidate model, weighted by segment length L_s . Segments that did not align with any truncal copy number state were not included in calculating D . Finally,

$$M = \sum L_s [\min(H_s / H_{\text{max}}, 1) - 1]^2$$

is a penalty term applied to a model if some segments show HSAF values H_s greater than the maximum possible HSAF value H_{max} . For a segment with a copy number of C in a sample with a cancer fraction of F , the maximum possible HSAF was calculated as $(CF + 1 - F) / (FC + 2 - 2F)$, representing the expected HSAF for the $C+0$ allele-specific copy number state.

After identifying the best truncal copy number model for a sample, we generated two additional models representing a genome duplicated and triplicated version of the model. The best errors of the duplicated and triplicated models are multiplied by 1.1 and 1.2, respectively. The models are reported in order from lowest to highest error.

A new visualization showing the complete coverage logratio, HSAF, and MAF data overlaid with the proposed model fit was used to validate the technical and biological plausibility of each model (**Supplementary Figure 1**).

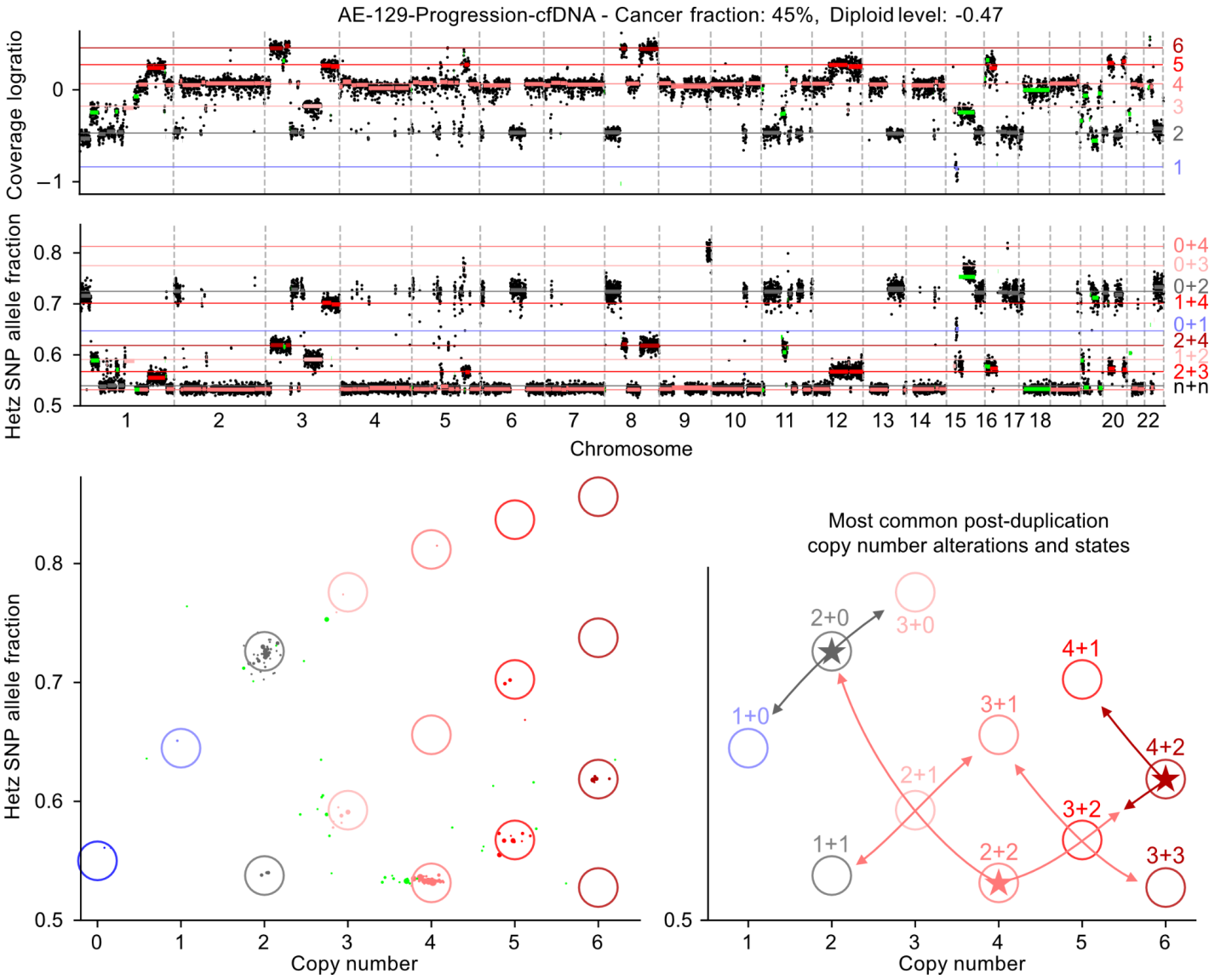


Supplementary Figure 1. Representative example of a best-fit copy number model for a plasma cfDNA sample. Upper plot shows segmented genome-wide coverage logratios (each dot is a 300 kb window). Horizontal lines indicate expected logratios for truncal copy number states, based on the best-fit model. The lower half displays heterozygous SNP allele fractions used to determine allele-specific copy numbers. Segments are colored by copy number (see **Methods—Tumor content and copy number estimation**). Segments marked in green do not coincide with integer copy numbers and represent regions harboring subclonal CN differences. Bottom left: allele-specific copy number states of individual genomic segments (colored dots), overlaid with open circles indicating values expected based on the best-fit copy number model. Bottom right: distribution of somatic mutation allele fractions inside genomic segments with specific truncal copy numbers (each row represents a copy number state). Expected positions of distribution peaks were calculated based on the best-fit copy number model, and are indicated with vertical ticks.

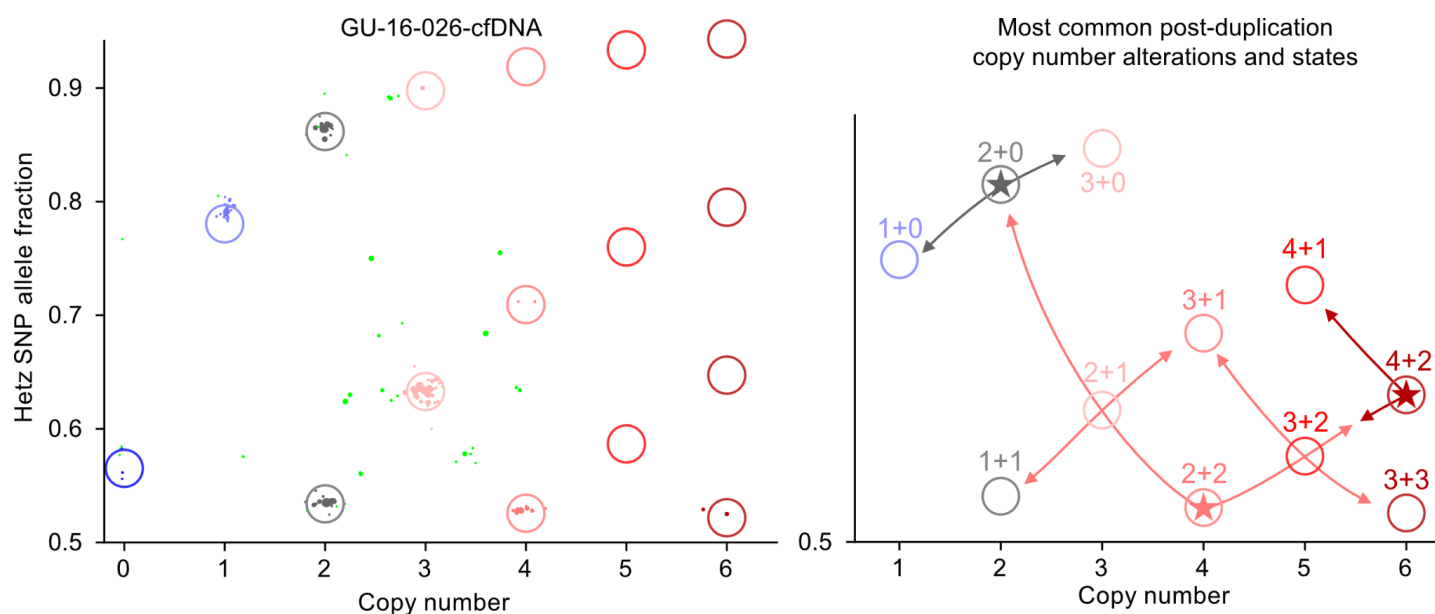
Detection of whole genome duplication

Whole genome duplication (WGD) adds complexity to model fitting because a cell population with WGD is indistinguishable from two non-WGD populations at 1:1 ratio in bulk DNA sequencing (i.e. when DNA fragments from all cells are mixed together). We used a heuristic where a cancer cell population was considered to have undergone WGD if all evaluable samples showed evidence for a 50% subclone of that population. Furthermore, if a sample was initially fitted with a diploid model where at least 50 Mb of the genome had $CN = 1$ with HSAFs in balance or $CN < 1$, a WGD model was adopted instead. Such regions would require at least one subclone to carry extensive biallelic losses, which is biologically implausible.

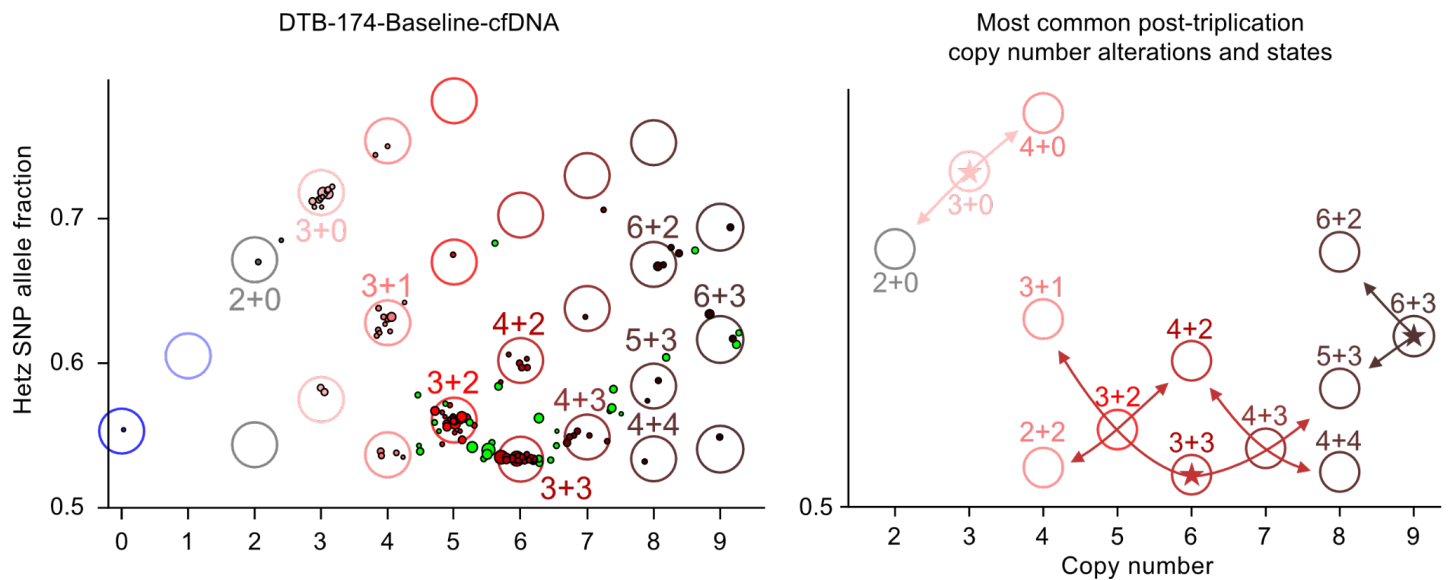
Additional evidence for whole-genome duplication, triplication, and quadruplication can be gleaned from the distribution of allele-specific CN states in individual samples (see right-hand plots below). In all examples below, the left-hand plots illustrate which allele-specific CN states are expected to be the most common in cells that have acquired a given number of extra genome copies. Stars indicate dominant allele-specific CN states (i.e. those arising from the 1+0, 1+1, and 2+1 pre-WGD CN states) after whole-genome duplication, triplication, or quadruplication. Arrows indicate possible single copy perturbations to these base ploidy states (deletions or gains) that can give rise to other frequently observed allele-specific CN states. If a genomic locus is not in allelic balance, any subsequent copy number alterations may be more likely to affect the major allele. For the sake of simplicity, we do not visually indicate these differences in probabilities in the visualizations below.



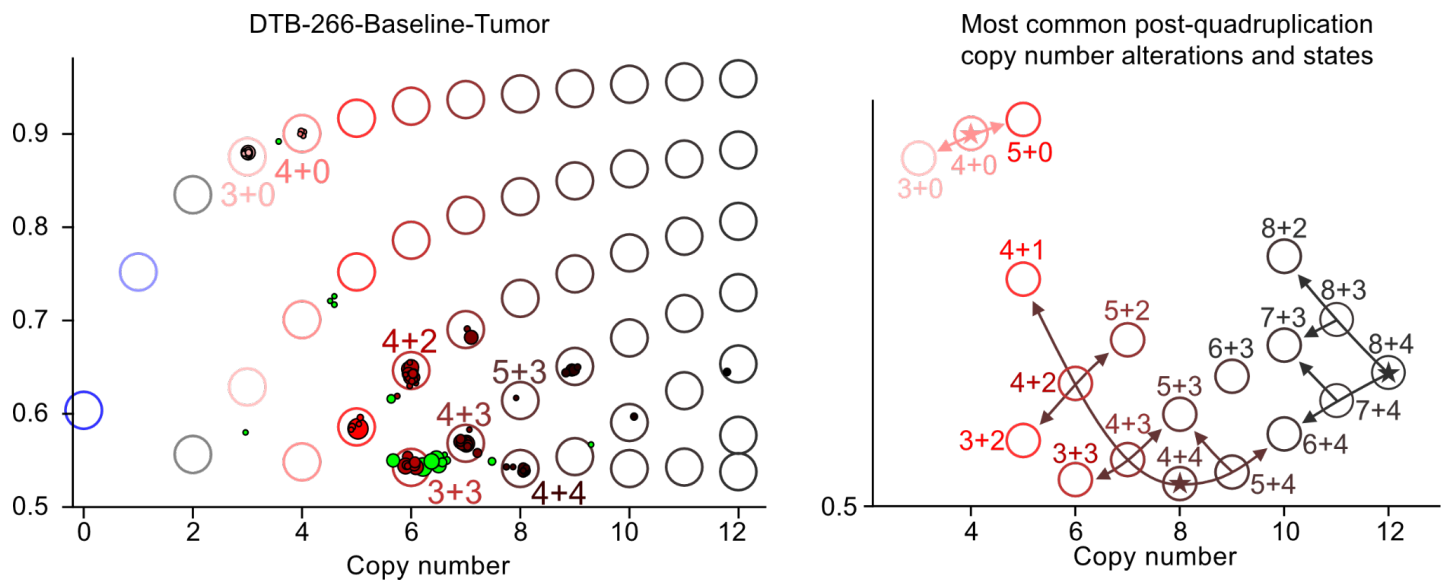
Supplementary Figure 2. Whole genome duplication (simple case). For this representative sample, the distribution of allele-specific copy number states implies whole genome duplication (WGD). This is supported by the higher prevalence of 2+2 copy-number regions (i.e. 4 copies total with balanced alleles) than 3+1 copy regions (i.e. allele-imbalanced 4 copy regions) based on the empirical density of dots representing individual copy number segments (right-hand plot). Alternatively, CN=4 regions could arise from consecutive duplications of lower CN states in the context of a diploid genome (rather than via a single WGD). However—assuming that consecutive copy duplications are independent—this alternative mechanism for producing CN=4 regions has a higher probability of generating imbalanced 3+1 allelic states than 2+2 states (which is not observed). Although it is theoretically possible that multiple independent duplication events (in a diploid context) could give rise to a predominantly CN=4 genome (with balanced haplotypes), a single WGD is a more parsimonious explanation and therefore the preferred interpretation.



Supplementary Figure 3. Whole genome duplication (complex case). In this sample, the average copy number is between 2 and 3. However, similarly to the sample above, segments representing 2+2 allelic states are significantly more common than segments representing 3+1 allelic states, implying whole genome duplication (WGD), rather than sporadic duplications from lower CN states, as the more likely explanation for the observed data. The low mean copy number is explained by the sample featuring a significant number of post-WGD single-copy deletions.



Supplementary Figure 4. Whole genome triplication. The distribution of allele-specific copy number states in this sample implies whole-genome triplication. This is supported by the higher prevalence of 3+3 copy-number regions (i.e. 6 copies total with balanced alleles) than 4+2 copy regions (i.e. allele-imbalanced 6 copy regions) based on the empirical density of segment CN states (right-hand plot). Although CN=6 regions could arise from sporadic duplications of lower CN states (rather than WGD), this alternative mechanism for generating CN=6 regions would result in an approximately equal likelihood of generating 3+3 and 4+2 allele-specific copy number states, which is not observed. Therefore, whole-genome triplication is a more parsimonious explanation for the observed distribution of allele-specific CN states.



Supplementary Figure 5. Whole genome quadruplication. The distribution of allele-specific CN states for this sample is most parsimoniously explained by whole-genome quadruplication. If instead CN=8 regions predominantly arose from multiple sporadic duplications of lower copy number states, we would expect to see an almost balanced distribution of 4+4 and 5+3 copy regions. However, since 4+4 regions are significantly more common than 5+3 regions, whole-genome quadruplication is more likely.

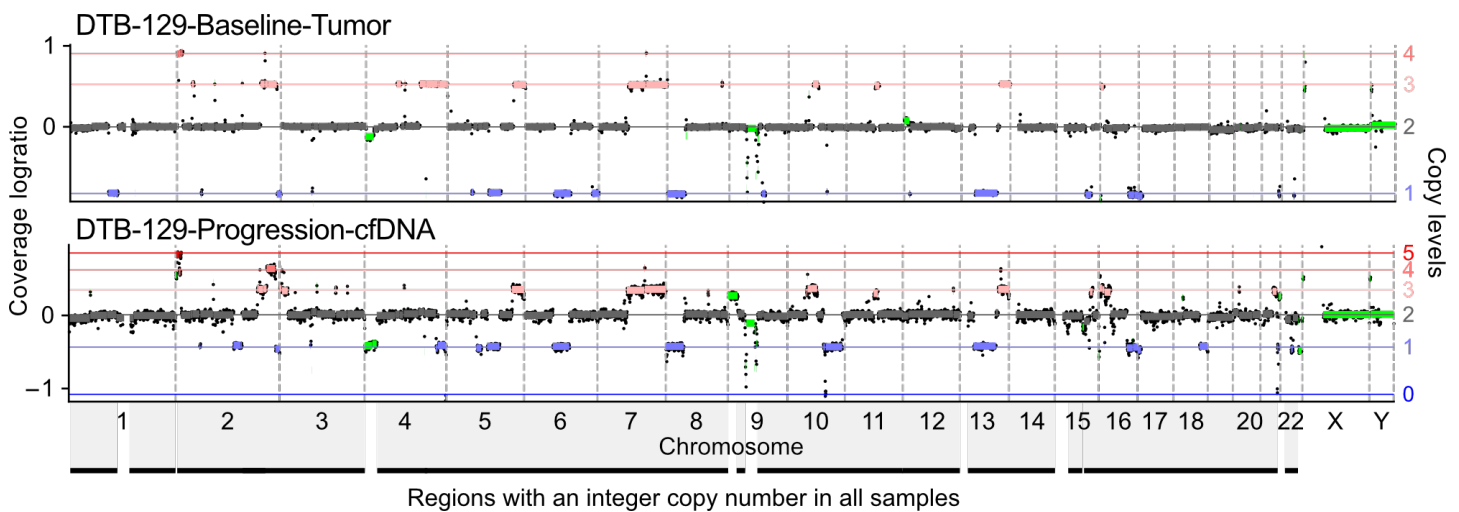
Quantifying chromosomal instability and subclonal copy number heterogeneity

Chromosomal instability was quantified as the percent of the genome with a copy number that deviated from a sample's baseline ploidy (2 for diploid genomes, 4 for tetraploid WGD genomes, 0.5× for sex chromosomes) by more than 0.15. Samples representing admixtures of two or more major clonal populations with different ploidies were excluded from this statistic.

Subclonal copy-number heterogeneity was quantified as the difference in segment absolute copy number and its nearest integer value (i.e. $\text{abs}(\text{round}(\text{CN}) - \text{CN})$), averaged over all same-sample segments produced from truncal copy number fitting ⁵.

Quantifying subclone fractions

A critical step in subclonal reconstruction is the identification of distinct cancer cell subpopulations present in a sample, and subsequent quantification of their relative abundances in all same-patient samples. To do this, we start with the sample-specific truncal copy number models inferred in section “Copy number model fitting and cancer fraction estimation”, and identify all genomic segments that show no evidence for subclonal copy number differences in any sample from the patient (**Supplementary Figure 2**).



Supplementary Figure 6. Truncal copy number model fits for two samples of patient DTB-129. Genomic regions having an integer copy number in all samples are highlighted with black bars and gray shading at the bottom of the figure.

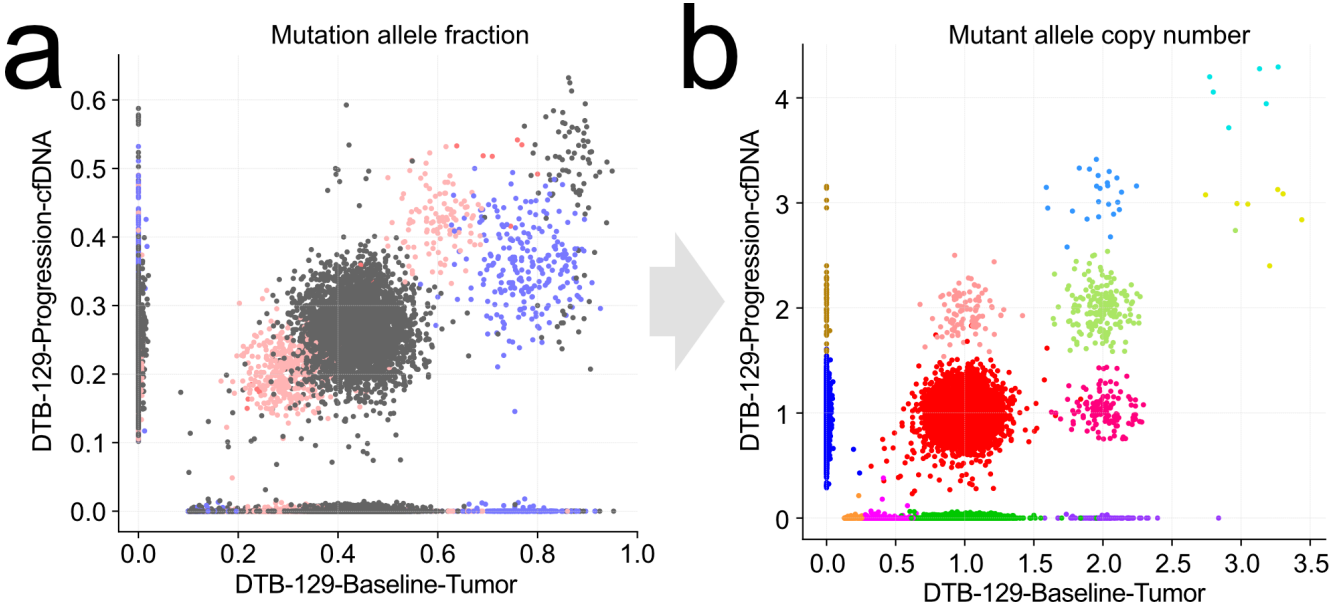
We then identify all somatic mutations that were detected in at least one of the patient's samples, and are located within these genomic regions. We measure the allele fractions of the somatic mutations in all patient samples (including samples where the mutation has a low allele fraction and would not have been independently detected), and visualize the allele fractions as a multi-dimensional scatter plot.

For each mutation identified, we solve its mutant allele copy number M in every sample (**Supplementary Figure 7**) from the formula:

$$A = \frac{FM}{FC + (1 - F)P}$$

A = mutation allele fraction
 F = cancer fraction
 M = mutant allele copy number
 C = region copy number in cancer cells
 P = region copy number in normal cells

The mutant allele copy number M is defined as the average number of mutant allele copies in the cancer cells represented in the DNA sample.



Supplementary Figure 7. Example of inferring mutant allele copy numbers from mutant allele fractions. **(a)** Scatter plot of mutant allele fractions in two samples from patient DTB-129. Mutations (dots) are colored according to the copy number of their genomic locus (blue = 1, black = 2, red = 3). **(b)** Mutant allele copy numbers (MACN) calculated based on the allele fractions and regional copy numbers. The mutations were clustered according to their multi-sample MACN values, and colored based on cluster assignment.

We identify mutation clusters in the N-dimensional MACN space (where N is the number of samples from the patient; **Supplementary Figure 7b**) using an automated algorithm based on a Bayesian procedure, similar to previous work ¹. Our algorithm first randomly selects 1000 mutations and creates new clusters with centroids at those locations within the MACN space. The algorithm then assigns every mutation to one of the clusters with assignment probabilities based on the likelihood L_m of the mutation m originating from that cluster (based on the position of the mutation and the cluster centroid in the MACN space), multiplied by the total number of mutations assigned to the cluster. The algorithm repeats this stochastic reassignment of mutations to clusters a total of 2000 times, discarding clusters that are reduced to zero mutations. This iterative cluster assignment process exhibits symmetry-breaking winner-takes-all behavior where redundant clusters are eliminated, leaving only clusters unambiguously supported by a set of observable mutations. The number of clusters does not need to be defined in advance. Our algorithm differs from Bayesian Dirichlet process-based clustering in that we only create clusters in the beginning (not during subsequent iterations), and can finish the clustering process very rapidly (within a few minutes on one CPU with $N = 3$ and 1000 initial clusters).

To determine the likelihood of mutation m originating from a cluster c , we calculate the centroid of cluster c in MACN space, then use the formula

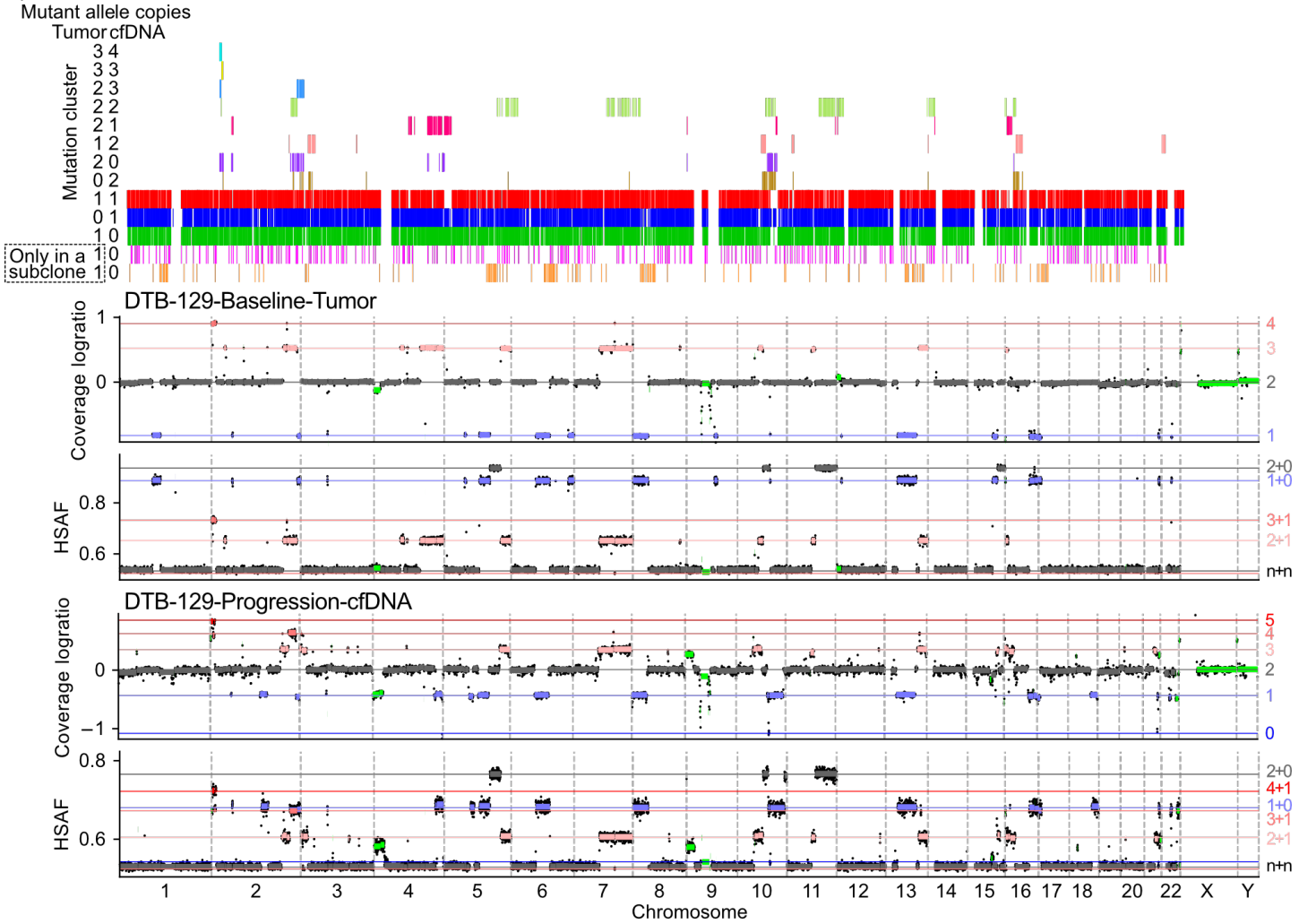
$$MAF = \frac{MF}{F(M + N) + (1-F) \cdot P} \text{ where}$$

F = cancer fraction in sample
M = mutant allele copy number in cancer cells
N = wildtype allele copy number in cancer cells
P = wildtype allele copy number in normal cells

to calculate expected MAF (in each sample) for a mutation with a MACN matching the centroid, and regional total copy number matching mutation m . For a mutation originating from cluster c , the observed mutant DNA fragment count in a sample is binomially distributed and independent from other samples. We can therefore calculate the likelihood of mutation m originating from cluster c as $L_c = \prod \text{Bin}(n_s, p_s)$, where n_s is the total

number of DNA fragments overlapping the mutation in sample s , and p_s is the expected MAF of a mutation originating from cluster c . The final cluster assignment probabilities are defined as Kn_cL_c , where n_c is the number of mutations already in the cluster, and K is a normalization constant chosen so that $\sum Kn_cL_c = 1$ (i.e. the assignment probabilities define a probability vector). After the last iteration of random assignment, the mutations are assigned to the identified clusters based on highest likelihoods (without randomization). When calculating the likelihoods, it is important that n_s is the number of DNA fragments rather than the number of sequencing reads, since paired end reads display the same allele and are therefore not independent observations.

Next, we visualize the genomic positions of mutations in each cluster to validate that the mutations localize to genomic regions whose copy number state is consistent with the observed MACNs (**Supplementary Figure 8**).

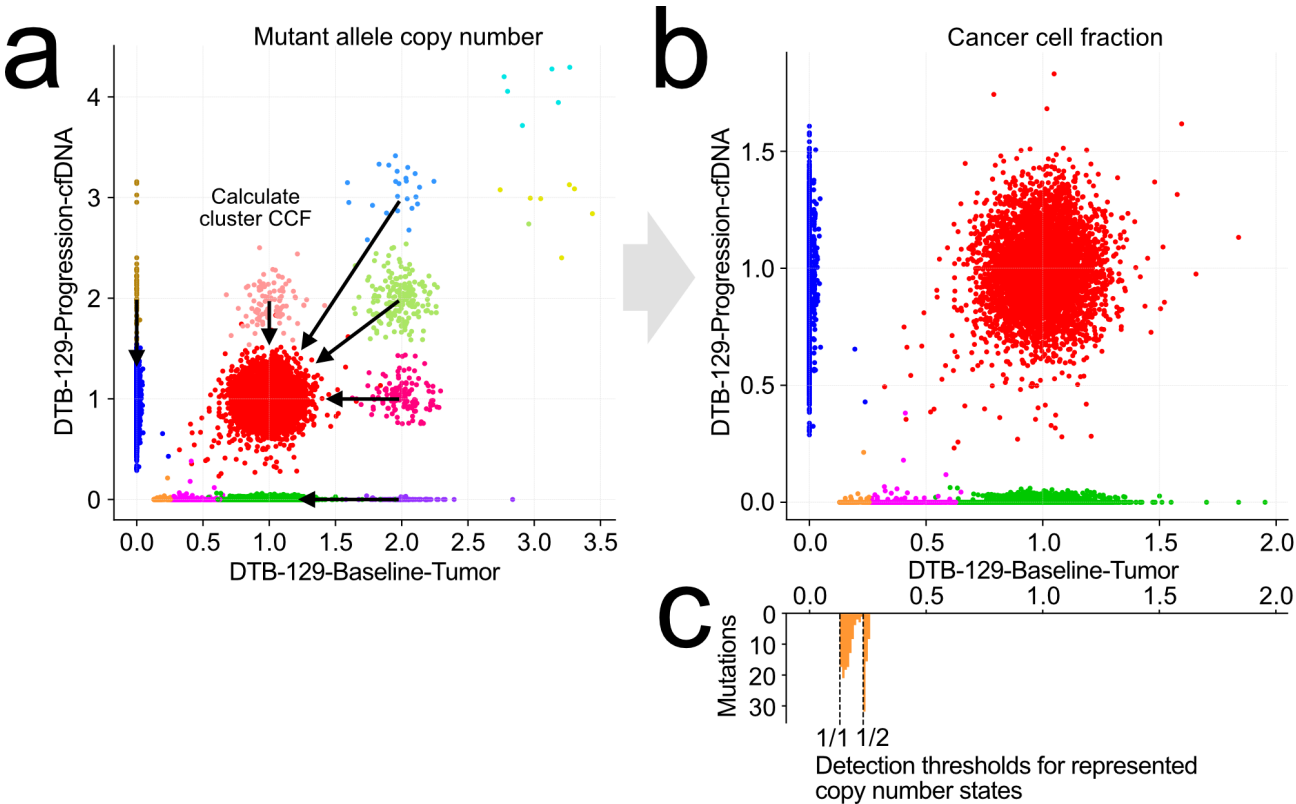


Supplementary Figure 8. Positions of patient DTB-129 mutations in the genome, grouped according to MACN clustering. The mutation clusters show expected genomic localization: mutations with >1 MACN come from regions with >1 major allele copy number, and mutations with differing MACNs between the samples come from regions whose allele specific copy numbers differ between samples. Mutations in the pink and orange cluster have a MACN of 1 in a subclonal population.

For any given sample, we calculate the cancer cell fraction (CCF) of a mutation cluster (i.e. the fraction of cancer cells carrying those mutations) as $CCF = \min(M, 1)$, where M is the average MACN of mutations in that cluster (**Supplementary Figure 9a-b**). Because our analysis only considers mutations in regions without subclonal copy number changes, all mutations can be categorized as either:

1. Subclonal single copy mutations that have arisen after truncal CN changes ($M < 1$), for which $CCF = M$
2. Truncal mutations with an integer M , for which $CCF = 1$

Each identified mutation cluster represents a distinct cancer cell population originating via the clonal expansion of a single ancestor cell carrying a unique set of somatic mutations. The CCF of a population in a given sample is estimated by calculating the average CCF of its mutation cluster in that sample.



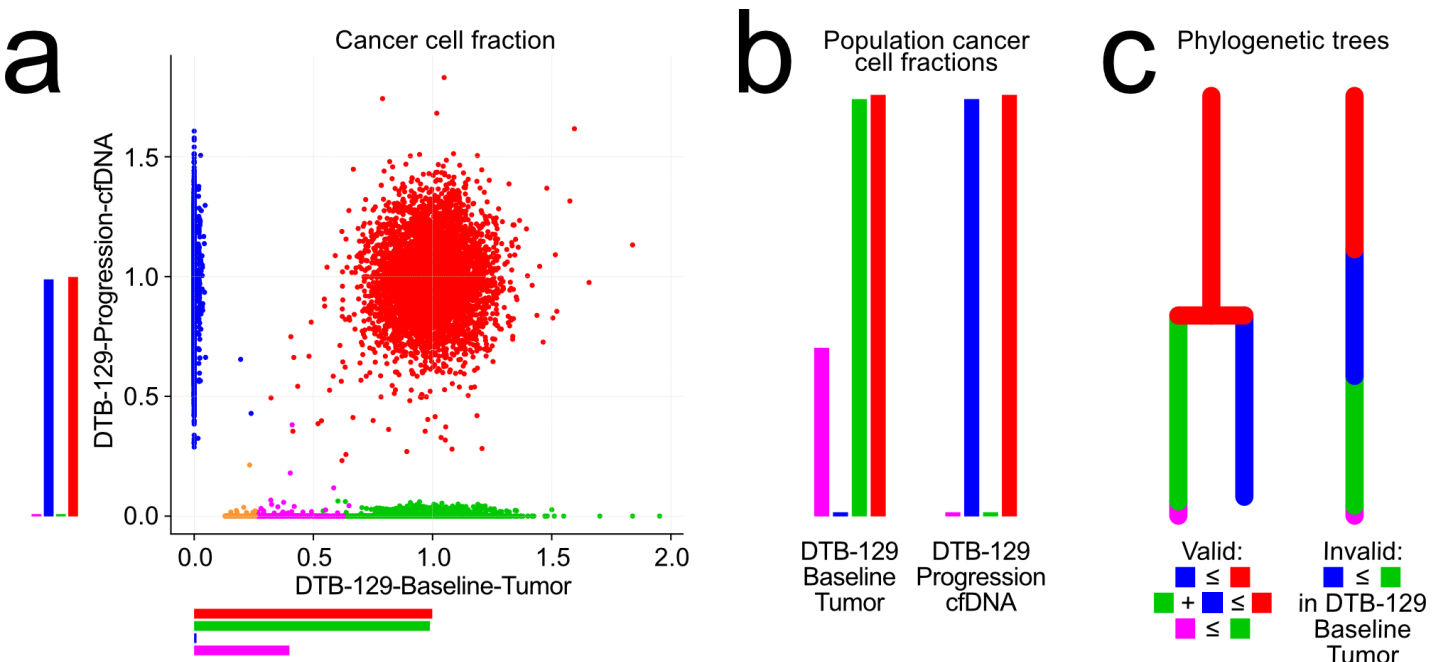
Supplementary Figure 9. Example of calculating mutation cancer cell fractions based on mutant allele copy numbers. Panel c shows the histogram of the orange clusters CCFs. Vertical lines indicate detection thresholds for mutations originating in regions with the indicated copy number states.

To validate that the identified clusters are not truncated by mutation detection thresholds, and consequently, that the cluster centroids can be confidently localized in the CCF space, we plot mutation histograms for each identified cluster to visually inspect the density of CCF data points. Without the histograms, this information can be difficult to interpret from the two-dimensional scatterplots, especially for axis-adjacent clusters; **Supplementary Figure 9c**). Truncated clusters are omitted from downstream phylogenetic and subclonal analyses, because these analyses are highly reliant on correct subclonal fractions.

For some samples, we manually adjusted the automated clustering result. For example, patient DTB-156 carried a false cluster composed of mutations that were completely deleted in a subclone and thus only found in deleted regions. As a second example, many samples had subclonal (or post-WGD) mutation clusters that had low MAF values in all samples and were partially truncated by the mutation detection process. Due to the partial truncation, the cluster centroid was artificially biased towards higher MACN (and CCF) values, and we adjusted the centroids based on the simulations described in section “Validation of mutation CCF clusters using a generative simulation” in this Supplement.

Inference of phylogenetic trees

After identifying the distinct cancer cell populations and estimating their CCFs in each sample from a given patient (**Supplementary Figure 9, Supplementary Figure 10a-b**), we infer a phylogenetic tree that indicates subclonal relationships between the populations (i.e. which populations are descendants of other populations; **Supplementary Figure 10c**). We do this by eliminating invalid phylogenetic trees that violate the *pigeonhole principle* (also known as the *sum rule*)⁶. This rule dictates that if clone A in the phylogenetic tree has direct descendant subclones B₁, ..., B_n, then in a valid phylogenetic tree the combined CCF of populations B₁, ..., B_n cannot be greater than the CCF of population A.



Supplementary Figure 10. Phylogenetic tree (c) of cancer cell populations is inferred from population cancer cell fractions (a-b) by applying the pigeonhole principle.

In some scenarios it is possible to find multiple phylogenetic trees satisfying the pigeonhole principle. In our cohort only patient AE-136 had multiple (two) valid phylogenetic trees. All inferred phylogenetic trees are shown in Supplement section “Patient truncal copy number models and subclonal reconstructions”. For patient AE-136 we carried out subclonal CN profile deconvolution on both of the two valid phylogenetic trees, and selected the tree that yielded a better fit to the observed data.

Reconstruction of subclone specific copy number profiles

After inferring the subclonal composition and phylogenetic tree of all samples from a patient, we infer the genome-wide copy number profiles of each subclone by applying a phylogeny-aware deconvolution to each genomic segment. Specifically, we search for combinations of allele-specific copy numbers (for each subclone) that—when weighted by the subclone’s cellular prevalence in every sample—best recreates the coverage logratio and HSAF values observed in the samples. We consider all possible combinations of subclone allele-specific copy numbers between 0 - 8 copies. Each combination is scored using an error function which aims to find an optimal copy number profile that fits the observed coverage logratios and HSAFs, while simultaneously minimizing the total number of copy number alterations (CNAs) in branches of the phylogenetic tree. The error function is

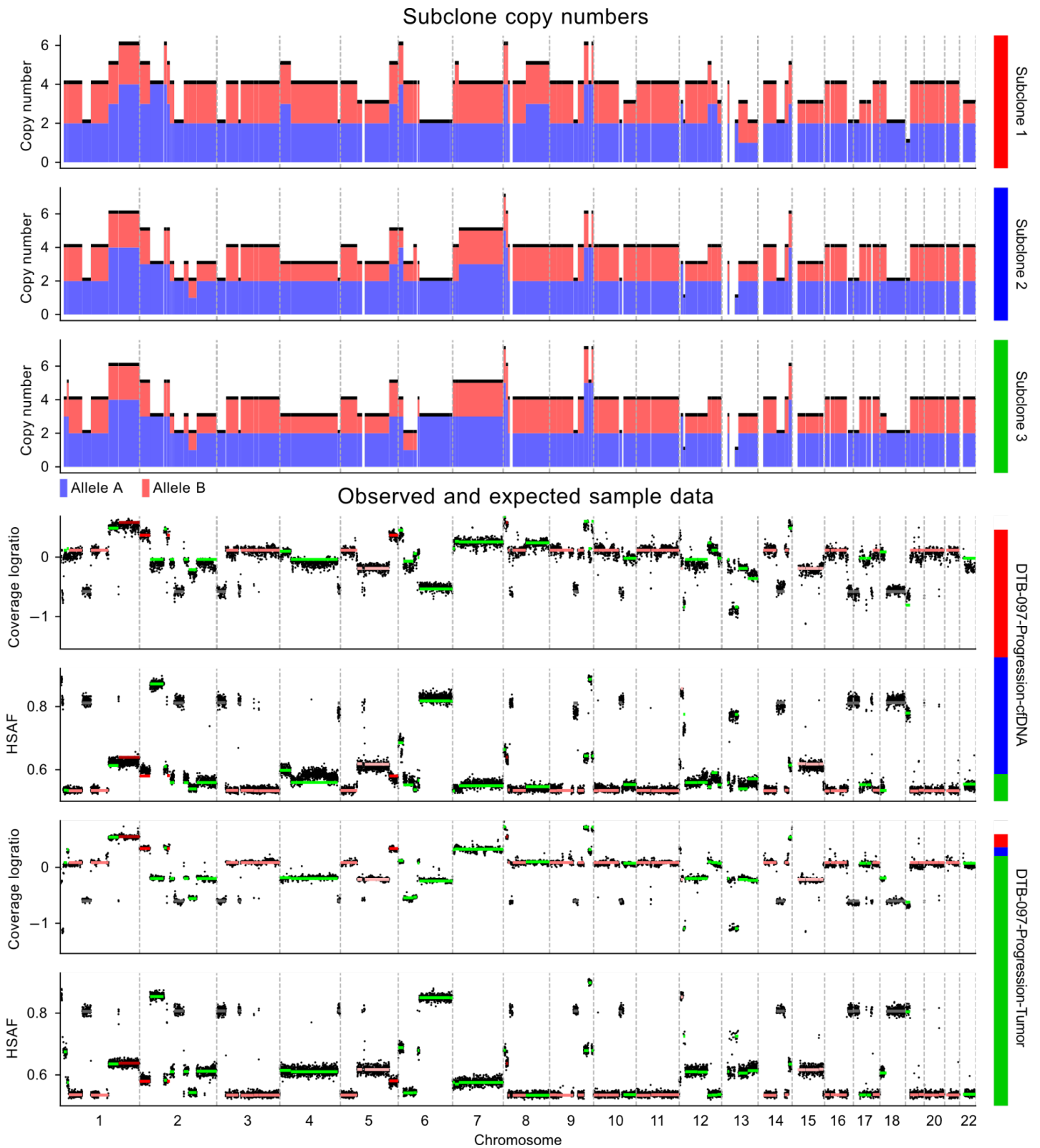
$$E = \max_s(aR_s) + \max_s(bH_s) + cK$$

where R_s and H_s are the absolute difference between observed and expected coverage logratio and HSAF z-score in sample s , respectively (detailed definitions below). K is the total number of CNAs in all branches of the phylogenetic tree. Optimal values for constants a , b and c were selected based on empirical benchmarking as $a = 1$, $b = 0.15$, $c = 0.05$. **Supplementary Figure 11** shows an example reconstruction of subclone copy number profiles.

Since cancer cell populations can give rise to new subpopulations carrying new CN alterations (including alterations that fully or partly revert ancestral CN changes), we model each sample’s coverage logratio and HSAF data as arising from a mixture of *homogeneous expanded populations*, i.e. cancer cell populations that exclude known subpopulations. In other words, we solve the CN profiles at each node of the phylogenetic tree. These nodes represent the cells of origin that gave rise to detected clonal expansions. It is important to note that our use of the term *homogeneous* in this context refers only to lack of detectable subpopulations and is not intended to suggest complete lack of cell-to-cell variability.

Definition of R_s and H_s in a single sample (sample index s omitted)

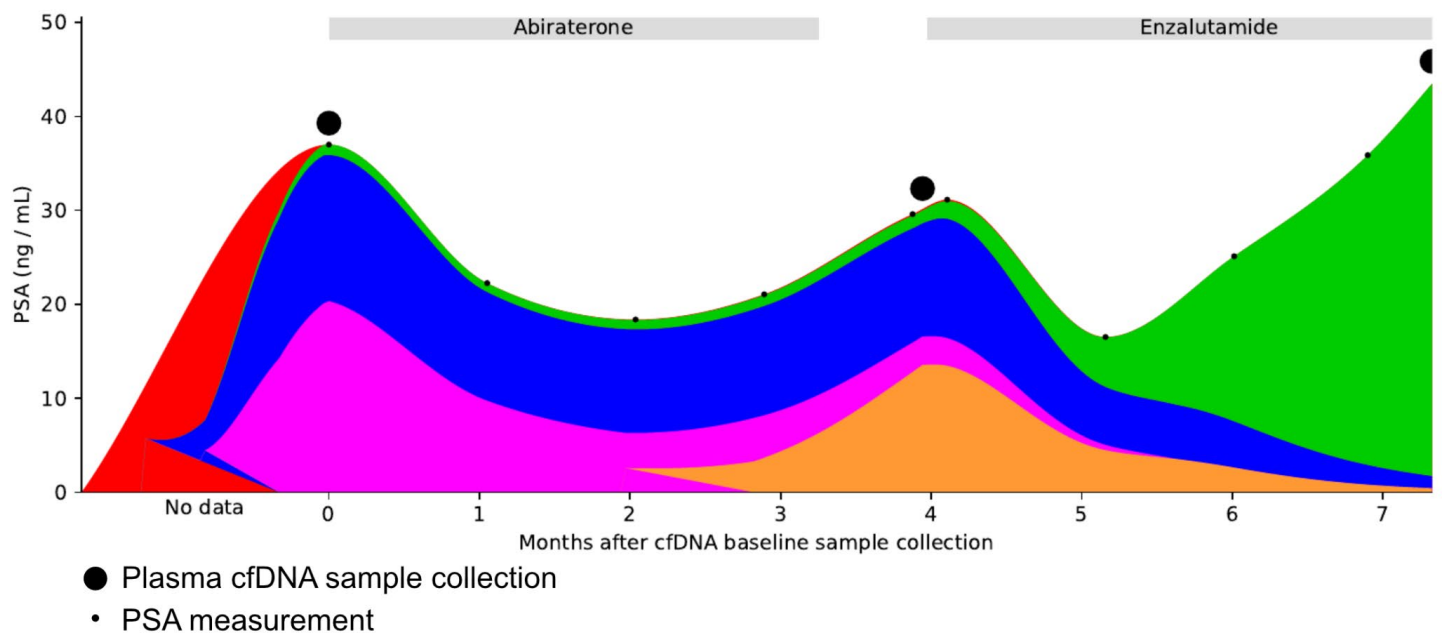
$R = R_e - R_o $	R_e = expected coverage logratio
$H = \frac{ H_e - H_o }{\sigma}$	R_o = observed coverage logratio
$R_e = \log_2\left(\frac{FC + 2 - 2F}{P}\right) + D$	H_e = expected average BAF
$H_e = \frac{FA + (1 - F)}{FC + (1 - F)P}$	H_o = observed average BAF
$C = \sum_{j=1..n} T_j(A_j + B_j)$	σ = BAF standard deviation
$A = \sum_{j=1..n} T_j A_j$	F = cancer fraction
	P = copy number in normal cells
	D = diploid level
	j = subclone
	A_j, B_j = allele copy numbers
	T_j = subclone fraction



Supplementary Figure 11. Subclone copy number model fits for patient DTB-097. Top panels show the copy number profiles of the three subclones. Heights of red and blue regions indicate A and B allele copy numbers. Bottom panels show the coverage logratios and HSAFs of the two samples. Horizontal segment lines indicate the expected segment levels given by the average subclone copy number weighed by subclone fractions (right side stacked bar plots). Segments with at least two distinct subclone copy number statuses are colored green, otherwise they are colored by the same copy number colors as used in our truncl copy number profiles.

Visualization of tumor burden and subclone fractions over time

Monotone cubic interpolation using Steffen’s method was applied to PSA measurements obtained at 4-week intervals to visualize tumor burden over time ⁷. This interpolation method guarantees that the interpolation curve passes through all data points, and never switches derivative sign (i.e. overshoots) between data points. Population CCFs were estimated at plasma cfDNA collection timepoints, and interpolated between timepoints by fitting an exponential function $f(t) = ae^{bt}$ to the boundary conditions $f(t_1) = F_1$ and $f(t_2) = F_2$, where F_k is population CCF at timepoint k . This method of interpolation was chosen to simulate the exponential nature of cell population growth. If a new subclone emerged or an old subclone disappeared between two cfDNA timepoints, we visualized the subclone as having crossed the detectability threshold in the exact middle of the two timepoints. If N nested subclones emerged during time interval T between two cfDNA timepoints, the time interval was divided evenly such that each nested subclone emerged after T / N time. The combination of dense PSA and sparse subclone data from plasma cfDNA allowed us to simultaneously visualize treatment-driven changes in tumor burden (i.e. treatment responses), and subclonal dynamics. To indicate phylogenetic relationships between subclones already present in the first available cfDNA timepoint, a qualitative rendering of clonal expansion starting from the MRCA of all cancer cells was drawn on the left-hand side of the figure, without a quantitative time axis. No time series data was available from the pre-mCRPC stage of disease.



Supplementary Figure 12. Visualization of PSA-based tumor burden and subclone fractions over time for patient AE-030.

Mutation-based timing of whole-genome duplication

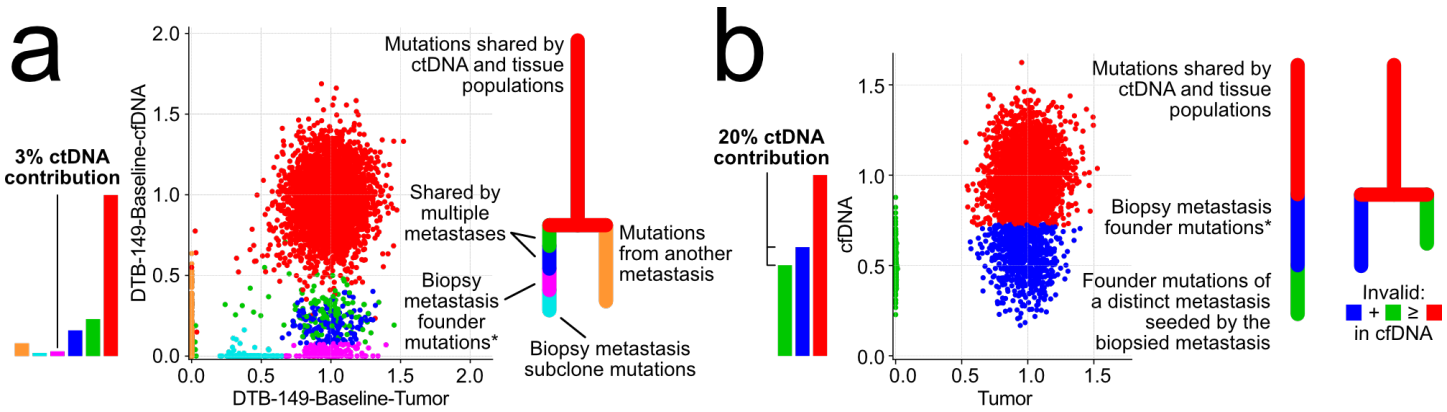
Mutations acquired before and after a whole-genome doubling (WGD) event harbor distinct mutant allele copy numbers (MACNs). This can be used to determine approximately when during a tumor's evolution WGD occurred (assuming a constant rate of mutagenesis). We focused on genomic regions with an identical integer allele-specific copy number in all samples, and denote $n+k$ as an allele-specific copy number state (major+minor allele copy number). For truncal events, the relative timing of WGD was calculated as the fraction of all truncal mutations with $\text{MACN}=1$ in genomic regions with $2+0$, $2+2$, or $4+2$ allele-specific copy number states. We excluded mutations in allele-specific copy number states $3+2$ and $2+1$ in our calculation, since in these genomic regions there exists MACNs that cannot be used to infer the mutation's timing relative to WGD due to multiple plausible etiologies. For example, $\text{MACN}=1$ in a $2+1$ allele-specific copy configuration could have originated via initial mutagenesis $\rightarrow$ WGD $\rightarrow$ mutant-copy loss, or WGD $\rightarrow$ mutagenesis $\rightarrow$ wildtype copy loss—therefore it is unclear whether mutations in $2+1$ regions occurred before or after WGD. When timing subclonal WGDs, the mutations were required to be truncal only in the subclones harboring the WGD. For estimating the timing of two subsequent WGDs resulting in an octaploid genome, we classified mutations in $4+4$ and $4+0$ regions to have occurred before the first WGD if $\text{MACN}=4$, between the WGDs if $\text{MACN}=2$, and after the second WGD if $\text{MACN}=1$. All calculations employ the 'infinite sites' assumption which states that the probability of the same mutation arising twice (in different alleles) is negligible.

Estimation of metastatic tissue contribution to ctDNA

To estimate the percentage contribution of a biopsied metastasis to total plasma ctDNA, we faced the initial problem that the cancer cell populations captured via metastatic tissue biopsy may not represent the entire lesion, and that closely related descendent populations may be present in multiple metastases (i.e. due to interlesion seeding) ⁸. To address these complicating variables, we surmised that each distinct metastasis is expected to harbor at least some unique truncal mutations by virtue of the founder effect. Therefore, we searched each tissue biopsy for a set of mutations that were truncal to the tissue biopsy but subclonal in time-matched cfDNA, and quantified the CCFs of these mutation clusters in the cfDNA sample (**Figure 3a; Supplementary Figure 13a**). If multiple tissue-truncal mutation clusters were identified, the cluster with the lowest ctDNA CCF was selected to quantify the contribution of the metastasis to bulk ctDNA. We assumed that the clusters with higher ctDNA CCFs represented mutations shared by closely related cancer cell populations spanning multiple metastases (i.e. entire branches of the phylogenetic tree). If all tissue-truncal mutations were also truncal in cfDNA, the biopsied tissue clone was inferred to be the sole or dominant source of ctDNA.

If the biopsied metastasis had seeded other metastases, those metastases would carry the same founder mutations present in the biopsied metastasis, and could confound the analysis by increasing the perceived plasma ctDNA contribution of the biopsied metastasis. However, the contribution of these descendant populations can be estimated and subtracted if their plasma CCF contribution is sufficiently high to allow their independent detection, and they can be phylogenetically proven as descendants of the biopsied metastasis (**Supplementary Figure 13b**).

In our cohort, tumor contributions (upper bounds) and detection thresholds were too low to detect if the biopsied metastasis had seeded new metastases.



*The founder mutation cluster may also include mutations that were acquired in the metastasis, and became truncal to it in a later clonal expansion

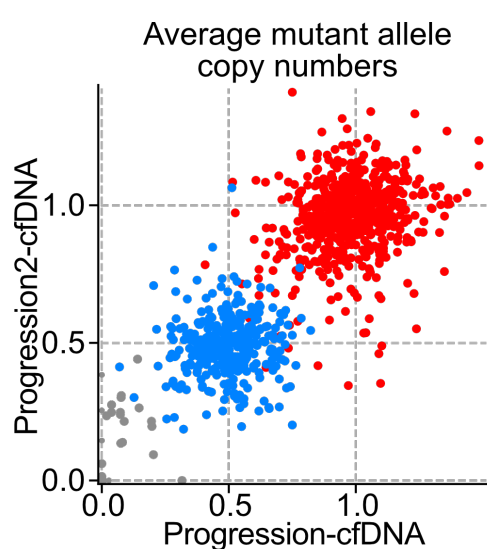
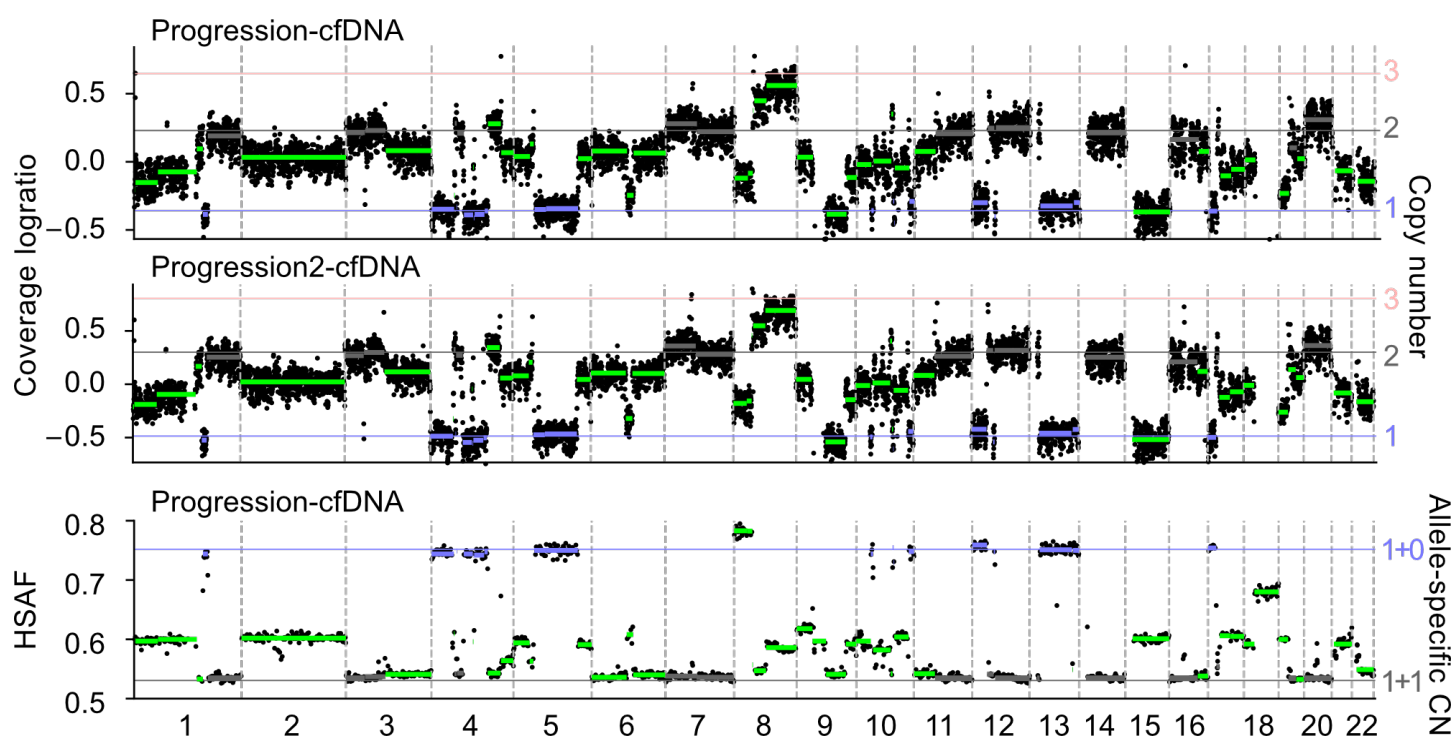
Supplementary Figure 13. Estimation of metastatic tissue contribution to ctDNA. a. Estimation of plasma ctDNA contribution by the biopsied metastasis of patient DTB-149. **b.** Hypothetical example of detecting a metastasis that was seeded by the biopsied metastasis.

Detection of subclonal whole genome duplication events

Whole-genome duplication (WGD) is a biologically relevant and common somatic feature of tumor genomes but its detection from bulk sequencing can be challenging. Many contemporary approaches for inferring WGD assume that samples never harbor multiple populations with differing WGD statuses. While prior landmark sequencing studies have proven this assumption to be applicable to the majority of tumor genomes, this assumption may preclude detection of other biologically plausible WGD configurations—including WGD occurring in one or more subclonal populations and/or successive WGDs occurring within a single branch of the phylogenetic tree.

Below we demonstrate an example of how we detected a subclonal WGD event in the plasma cfDNA samples of patient AE-018, through a process of eliminating simpler hypotheses. We then show how we calculated sample cancer fractions and population CCFs in this complex circumstance.

AE-018 Hypothesis 1: Diploid clones in all samples

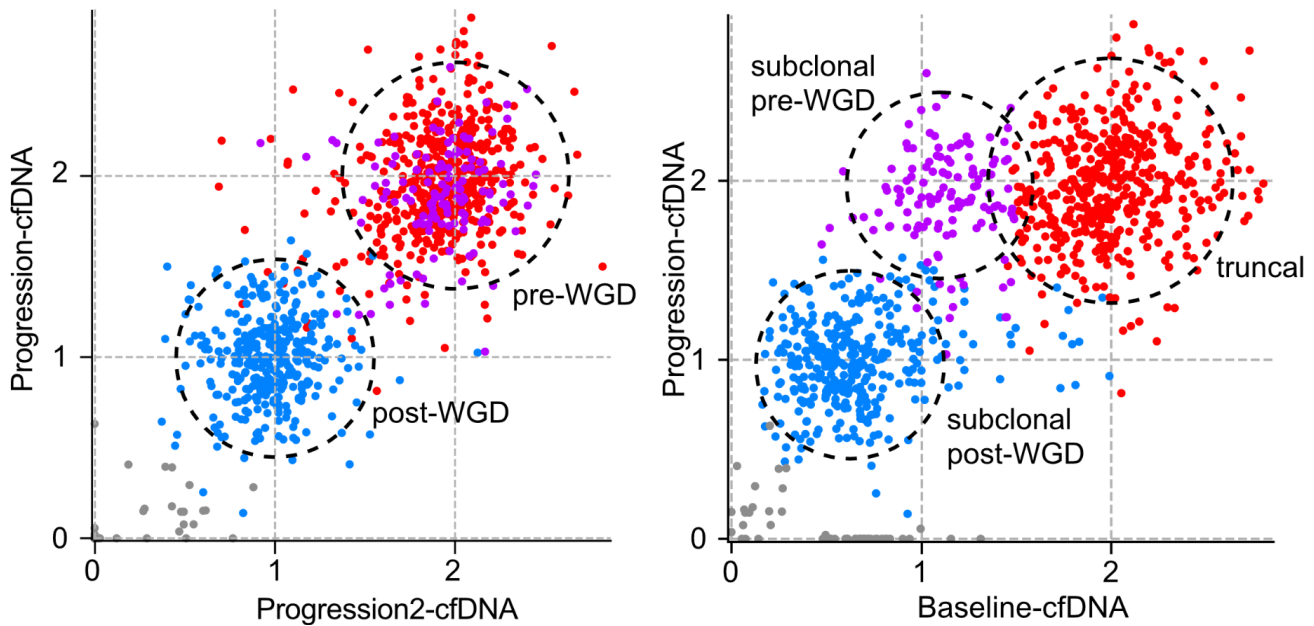
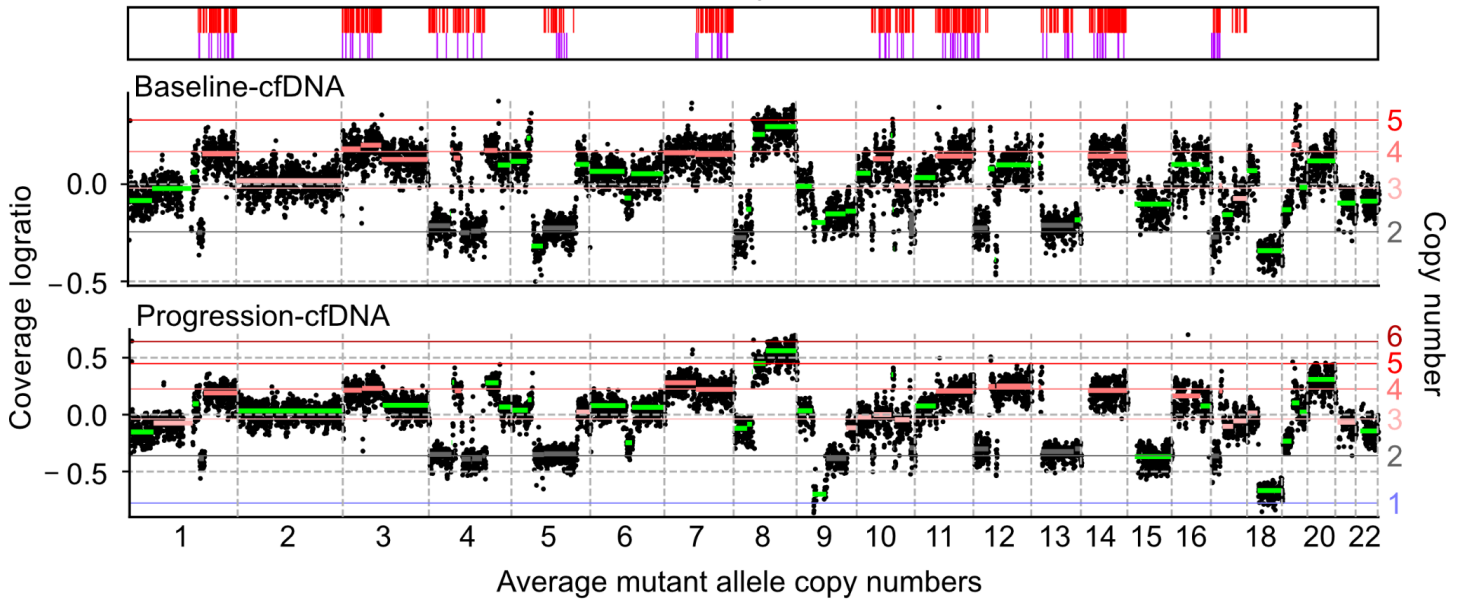


In hypothesis 1, a diploid CN model was fitted to both cfDNA samples. The hypothesis was rejected because with this model we can see an apparent subclone with exactly 50% CCF in both cfDNA samples. Chance observation of this precise clonal configuration (50% CCF) in two separate samples is unlikely, and therefore the cluster is more parsimoniously explained by a WGD. Assuming that the populations captured by these samples share a common truncal WGD, the 100% CCF cluster would represent truncal pre-WGD mutations and the 50% CCF cluster would represent truncal post-WGD alterations.

This visualization also shows that the 1st and 2nd progression cfDNA samples (the latter collected only 29 days apart after ineffective enzalutamide treatment) had no detectable differences in terms of subclonal or genomic composition.

AE-018 Hypothesis 2: One WGD event truncal to all samples

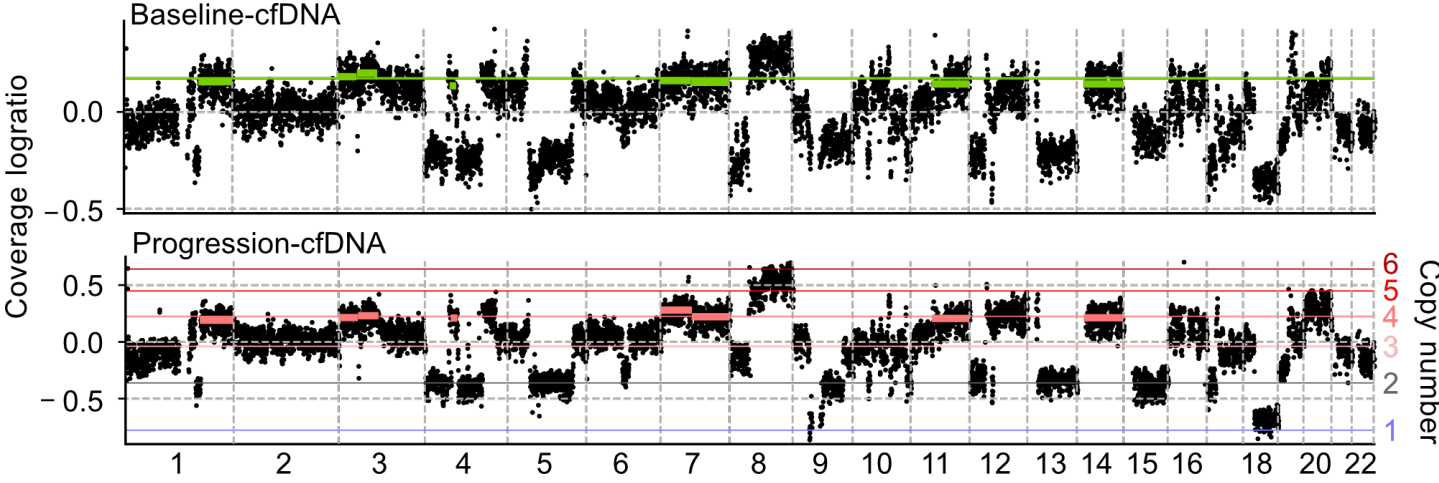
Mutation positions



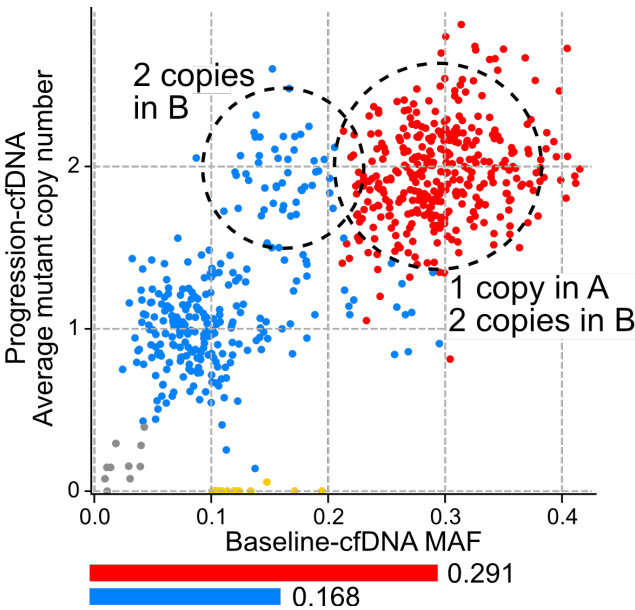
In hypothesis 2, we fitted a truncal WGD model to all samples. Mutation clustering revealed a set of purple mutations that were pre-WGD in Progression-cfDNA and Progression2-cfDNA samples, but separate from the red cluster in Baseline-cfDNA. If the WGD event was truncal to all populations, all pre-WGD mutations should have $MACN \geq 2$ in all samples unless mutant allele copies were lost due to a post-WGD CNA. However, the purple cluster was not explained by post-WGD CNAs, because red and purple mutations are found in the same set of regions, and there are no distinguishable CNAs affecting those regions in Baseline-cfDNA. Thus, hypothesis 2 is rejected.

AE-018 Hypothesis 3: One diploid population and one WGD population

After rejecting the hypotheses of diploid populations and a shared truncal WGD, we conclude that there must be both a diploid population **A** and a WGD population **B**. Our analysis of hypothesis #2 suggested that sample Baseline-cfDNA contains a mixture of both populations, while the progression cfDNA samples only contain the WGD population **B**. Under this hypothesis, the subclonal reconstruction becomes more complex: since one population has undergone a WGD event, no regions share the same allele-specific copy number in both populations. Thus in Baseline-cfDNA, all genomic regions have a subclonally heterogeneous CN state, and we cannot fit a truncal copy number model or calculate MACN values using our standard procedure. Instead, we focus on our analysis on the genomic regions highlighted with thick green and red lines below:



In the Progression-cfDNA sample, these originally diploid regions have a copy number of 4 (see copy number model above). In Baseline-cfDNA, these regions still line up with each other, indicating that none of them have undergone copy number alterations other than the WGD event in population **B**. Therefore in Baseline-cfDNA these regions have a copy number of 2 in the diploid population **A**, and a copy number of 4 in the WGD population **B**.



Since we cannot directly calculate Baseline-cfDNA MACN values, we instead visualize Baseline-cfDNA MAFs together with Progression-cfDNA MACNs. We cluster the mutations and color them based on whether they are **unique to the WGD population B**, **unique to the diploid population A**, or **truncal mutations shared by both populations** (see figure on the left).

We observe that we can express the MAF of any somatic autosomal mutation in Baseline-cfDNA as

$$\text{MAF} = \frac{F(F_A M_A + F_B M_B)}{F(F_A C_A + F_B C_B) + (1 - F) \cdot 2}$$

where F is the cancer fraction, and F_A and F_B are the CCFs of populations A and B in the sample.

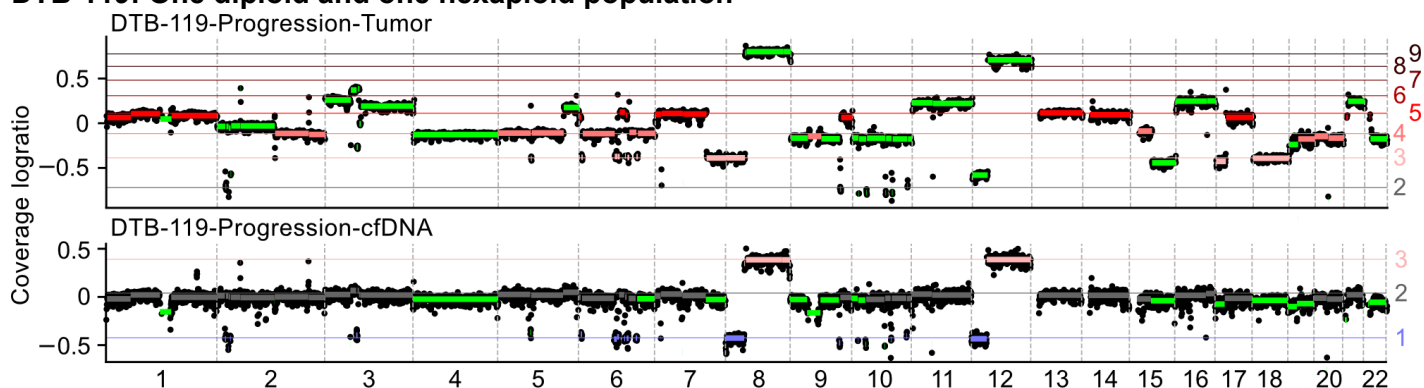
We then measure the average MAF of two mutation clusters (see circles in figure) in Baseline-cfDNA, and write out the

MAF equation for both:

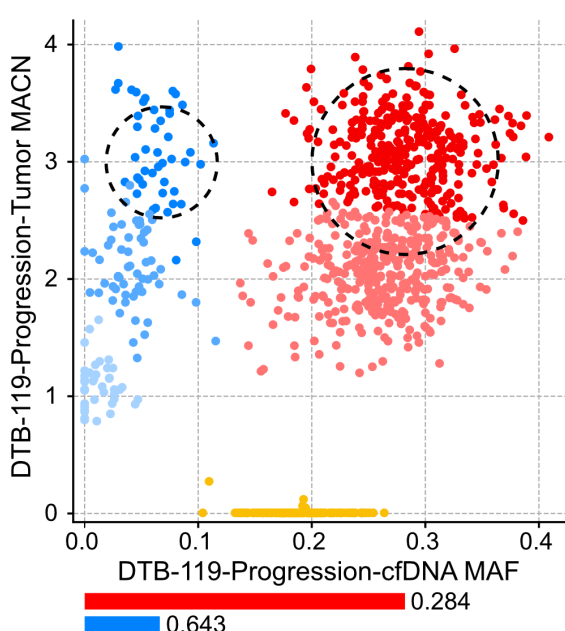
$$\text{MAF}_{\text{blue}} = \frac{F(F_A \cdot 0 + F_B \cdot 2)}{F(F_A \cdot 2 + F_B \cdot 4) + (1 - F) \cdot 2} \quad \text{MAF}_{\text{red}} = \frac{F(F_A \cdot 1 + F_B \cdot 2)}{F(F_A \cdot 2 + F_B \cdot 4) + (1 - F) \cdot 2}$$

Since we additionally know that $F_A + F_B = 1$, we have three equations with three unknowns, and can solve the sample's cancer fraction $F = 0.50$, and the population CCFs $F_A = 0.59$ and $F_B = 0.41$.

DTB-119: One diploid and one hexaploid population



The original truncal copy number profiles for patient DTB-119 suggested that the tumor sample was composed of a hexaploid population, and the cfDNA sample was composed of a diploid population. Subsequent mutation clustering revealed a pattern that led us to conclude that the cfDNA sample represented a mixture of a major diploid clone and a minor hexaploid clone.

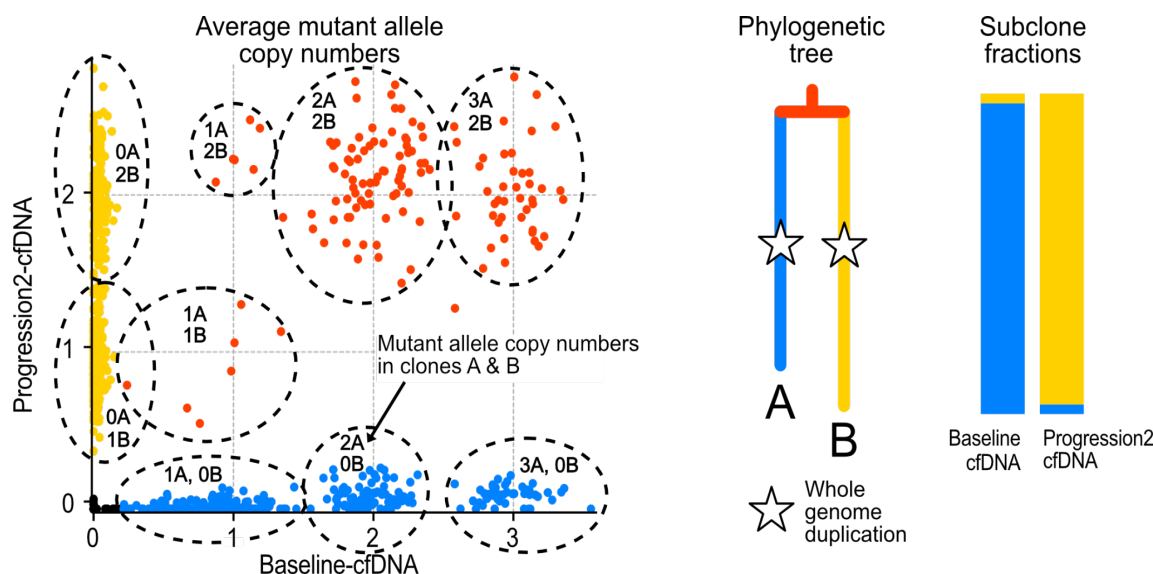


To demonstrate this, we select genomic regions with CN = 5 in the hexaploid sample and CN = 2 in the diploid population (e.g. chromosomes 7, 13 and 14). We then plot tumor MACN values and cfDNA MAF values for all mutations from these regions. The plot revealed **three clusters of blue mutations unique to the hexaploid population**, and two clusters of **red mutations shared by both populations**. Mutations with tumor MACN values < 3 showed consistently reduced MAF values in the cfDNA sample, consistent with subclonal presence of the hexaploid population in the cfDNA sample. We selected the two mutation clusters indicated with dotted circles (see figure on the left), and using the formulas presented on the previous page, solved the cfDNA sample to have a cancer fraction of 0.51, a hexaploid population CCF of 0.09, and a diploid population CCF of 0.91.

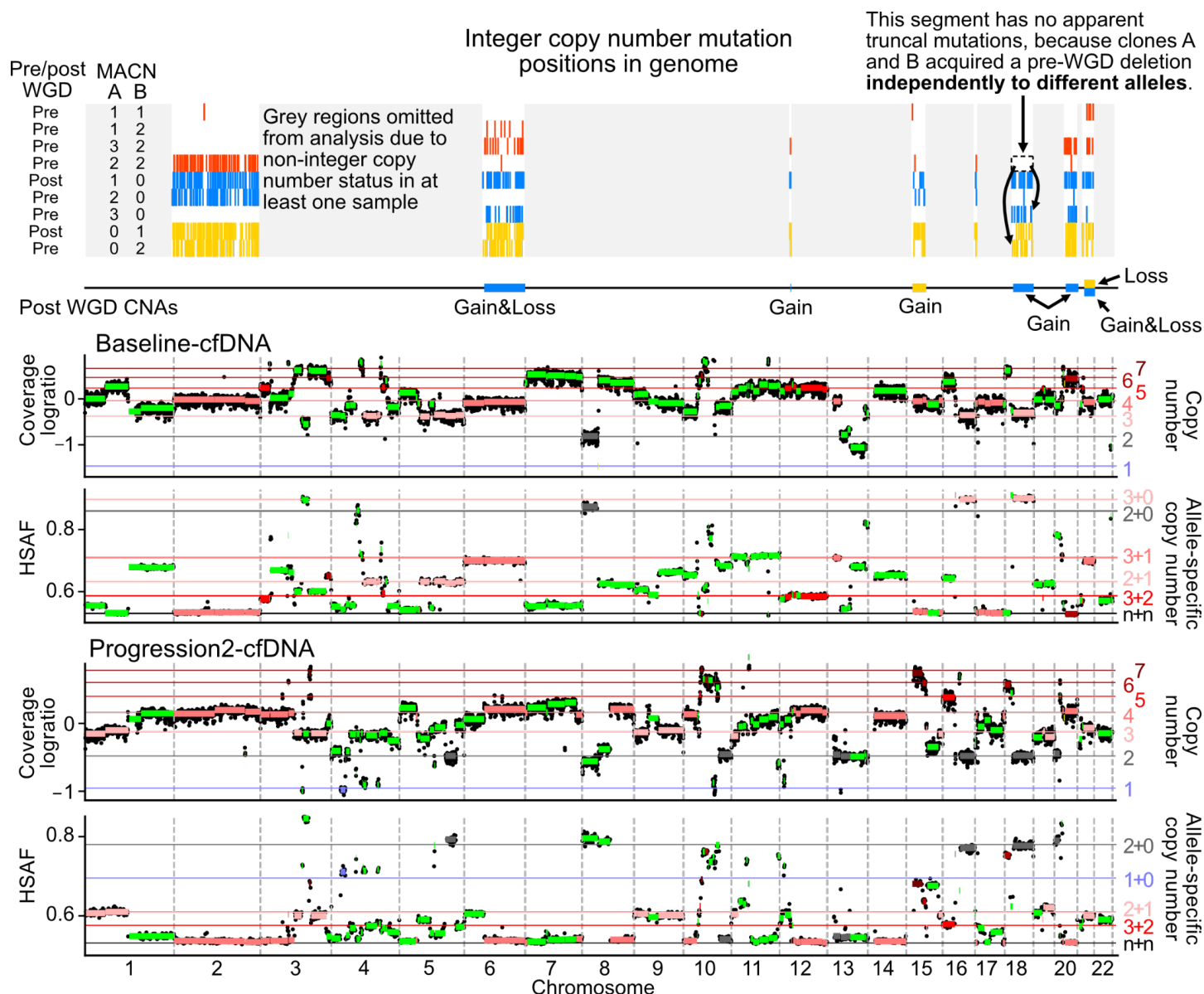
See section “Patient truncal copy number models and subclonal reconstructions” of this Supplementary Information document for full visualizations of the truncal models.

Evidence for parallel evolution of whole genome duplication

Below we provide detailed evidence for the independent occurrence of two whole genome duplication (WGD) events in separate sub-branches of the phylogenetic tree of patient AE-030.



Both clones have unique mutations with >1 copies, which are not specific to regions with post-WGD gains. Therefore, both lineages acquired the WGD independently.



Identification of somatic mutations that expanded during treatment

To search for potential somatic drivers of acquired resistance to AR targeted therapies, we sought to identify somatic mutations that showed an increase in their cancer cell fraction (CCF) or mutant allele copy number (MACN) in a post-treatment progression sample relative to a sample collected before treatment initiation. A naive comparison of mutation allele fraction (MAF) values between the two samples is insufficient, because the total plasma ctDNA fraction may change drastically between samples, and the copy number of the genomic locus harboring the mutation may also undergo changes if new cancer cell populations take over.

It can be impossible to identify a definite CCF value for every somatic mutation in every sample, because there can be multiple plausible scenarios consistent with a mutation's observed MAF. As a trivial example, if a sample contains two cancer cell populations at a 2:1 ratio, it is impossible to know whether a mutation is present in the dominant subclone at MACN = 1, or present in the minor subclone at MACN = 2, without external information. This problem is amplified in genomic regions that show CN heterogeneity between subclonal populations, as this increases the range of possible scenarios.

To deal with these challenges, we introduce the concept of *average mutant allele copy number*, defined as the average number of mutant allele copies in all cancer cells represented in a given DNA sample. In a sample with multiple subclonal populations, the mutant allele fraction of a mutation can be mathematically expressed as

$$\text{MAF} = \frac{F(M_1F_1 + \dots + M_nF_n)}{F(F_1C_1 + \dots + F_nC_n) + (1 - F) \cdot P}$$

where F is the total cancer fraction of the sample, F_k is the CCF of subclone k , M_k is the mutant allele copy number of the mutation in subclone k , C_k is the total copy number of the mutated locus in subclone k , and P is the germline copy number of the mutant locus. We also know that $F_1 + \dots + F_n = 1$. Now since $M_1F_1 + \dots + M_nF_n$ is the average MACN and $F_1C_1 + \dots + F_nC_n$ is the average locus copy number in the cancer cells represented in the sample, we can forget about the subclones and express the average MACN simply as

$$\text{MACN}_{\text{mean}} = \text{MAF}[FC_{\text{mean}} + (1 - F) \cdot P].$$

This formula allows us to calculate the average MACN for any mutation in any sample with zero required knowledge about the subclonal composition of the sample. All terms on the right-hand side (MAF and C_{mean}) can be easily determined based on a sample's truncal copy number model and the coverage logratio at that locus.

By studying temporal changes in average MACN rather than changes in CCF, we sidestep issues that normally arise when attempting to assign mutations to individual subclones (and solving the MACN values of those mutations in those subclones). Any change in the average MACN of a mutation between two samples must represent either 1) a change in the mutation's CCF, and/or 2) a change in the mutation's MACN in the cancer cells. Importantly, differences in total cancer fraction or the copy number of the wildtype allele between samples have no effect on average MACN, and are therefore eliminated as confounding factors.

Since MAF measurements are based on random sampling of mutant and wildtype DNA fragments in a sample, they are subject to binomially distributed sampling noise. This error propagates to the calculated average MACN values. A statistical analysis is therefore necessary in order to determine whether a change in a mutation's average MACN value between two samples represents a genuine biological change, or mere measurement noise. We start with the null hypothesis that our mutation of interest has the same average MACN value in two samples A and B. Under this null hypothesis, our best estimate for the mutation's true average MACN value is the average of the measured average MACN values in samples A and B. After calculating this best estimate of the true average MACN, we use the formula on page 3 to convert this average MACN value to "true" MAF value for both samples. We then add binomially distributed noise to these MAF

values (based on the total sequenced fragment depth at the locus in each sample), and convert them back to MACN values. We then calculate the difference between the two simulated MACN values. This simulation is repeated 10,000 times, and we count in how many iterations the difference between the original MACN values was more extreme than the difference between the MACN values that were simulated under the null hypothesis. This process (known as a permutation test) provides us with a p-value for the probability of the observed data under the null hypothesis. We reject the null hypothesis if $p < 0.0005$.

Identification of rearrangements that expanded during treatment

To search for somatic chromosomal rearrangements that showed an increase in their allele fraction between consecutive samples (adjusted for cancer fraction differences between the samples), we calculated in each sample s the number of rearrangement supporting DNA fragments A_s and the total number of cancer-derived DNA fragments $T_s = N_s F_s$ where N_s is the sample's total read count and F_s is the sample's cancer fraction. We then used a standard Fisher's exact test to calculate a p-value for the null hypothesis $A_1 / T_1 = A_2 / T_2$. This method controls for between-sample differences in cancer fraction and sequencing depth. It circumvents the need for quantifying the precise allele fraction of a rearrangement, which is challenging because rearrangements involve two chromosomal loci, each with their own wildtype sequence, copy number, and sequencing depth. The null hypothesis was rejected if $p < 0.05$ after multiple hypothesis test correction using the Benjamini-Hochberg method.

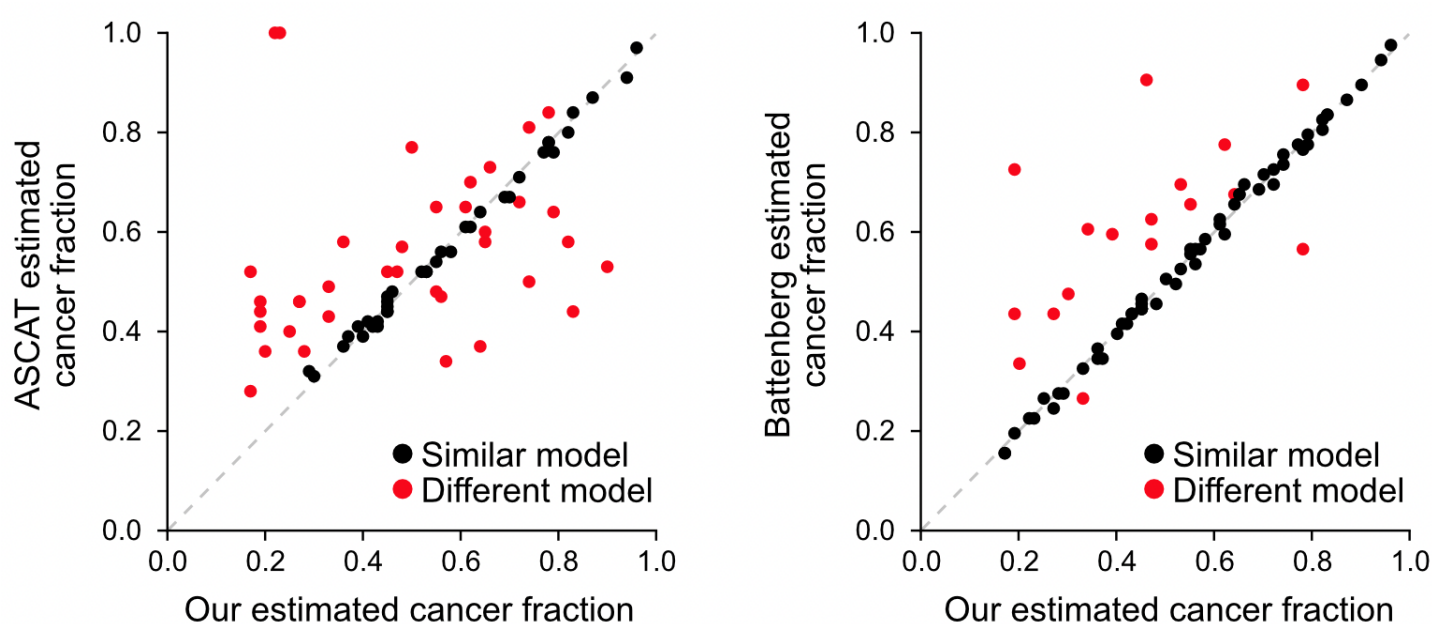
Selection of patient samples with DNA repair defects

Homologous recombination repair defects (HRRd), mismatch-repair defects (MMRd), and *CDK12* defects were originally determined from prior targeted exon sequencing data (enabling selection of samples with these subtypes for subsequent WGS) ^{9–11}. Specifically, we selected samples harboring evidence for putatively deleterious defects in *BRCA2* (HRRd), *MSH2/MSH6* (MMRd), or *CDK12*. Mutation pathogenicity was assessed on the basis of predicted protein-level consequences: truncating variants—constituting frameshift insertions/deletions, splice site mutations (mutations within ± 2 bp from an exon/intron junction), and nonsense (stopgain) mutations—as well as deep deletions (predicated on gene logratio and lack of HSAF deviation) were presumed deleterious. Germline missense mutations classified as 'pathogenic' or 'likely pathogenic' in the ClinVar database were also considered deleterious. Missense variants disrupting the *CDK12* kinase domain are linked to protein inactivation and were additionally classified as deleterious if predicted to be at least 'probably damaging' by Polyphen2. Samples with putatively deleterious defects in these genes that were included in our WGS cohort had strong genome-wide evidence for associated phenotypes (including established COSMIC v2.0 mutational signatures for HRRd (signature 3) and MMRd (signatures 6, 15, and 21), as well as the *CDK12*-associated tandem duplicator phenotype ¹²).

Validation

Validation of truncal copy number models

We validated our curated copy number models using ASCAT 2.5.2 ¹³ using the gamma = 1.0 setting designed for sequencing data. ASCAT identified a cancer fraction identical to ours for 50% of samples, a different model for 47% of samples (**Supplementary Figure 14**), and failed to produce any model for 3% of samples (all cfDNA samples with extensive polyclonality). Additional validation was carried out using Battenberg 2.2.9 ¹ in hg38 native tumor-normal mode, and with max_ploidy = 6.8. Battenberg discovered a copy number model identical to ours for 78% of samples, and a different model for 22% of samples. Side-by-side comparisons of the proposed models overlaid on the underlying raw coverage logratio and HSAF data suggested that our models provided a better fit in the discordant cases (see figures in section “Battenberg copy number models” of this supplement). Since Battenberg outputs an average sample ploidy instead of a diploid coverage logratio, we set the diploid level for the Battenberg visualizations such that the resulting average sample ploidy matched the one reported by Battenberg. This comparison is shown for validation purposes, and is not intended as a comparison between unsupervised algorithms, since a few of the models reported by our algorithm were manually adjusted based on information obtained from same-patient samples (particularly with respect to whole genome duplication).



Supplementary Figure 14. Comparison of our curated copy number models and models generated automatically by ASCAT and Battenberg algorithms. Each dot represents a sample. Samples where the two approaches yielded a different cancer fraction (absolute difference > 4%) are shown as red dots.

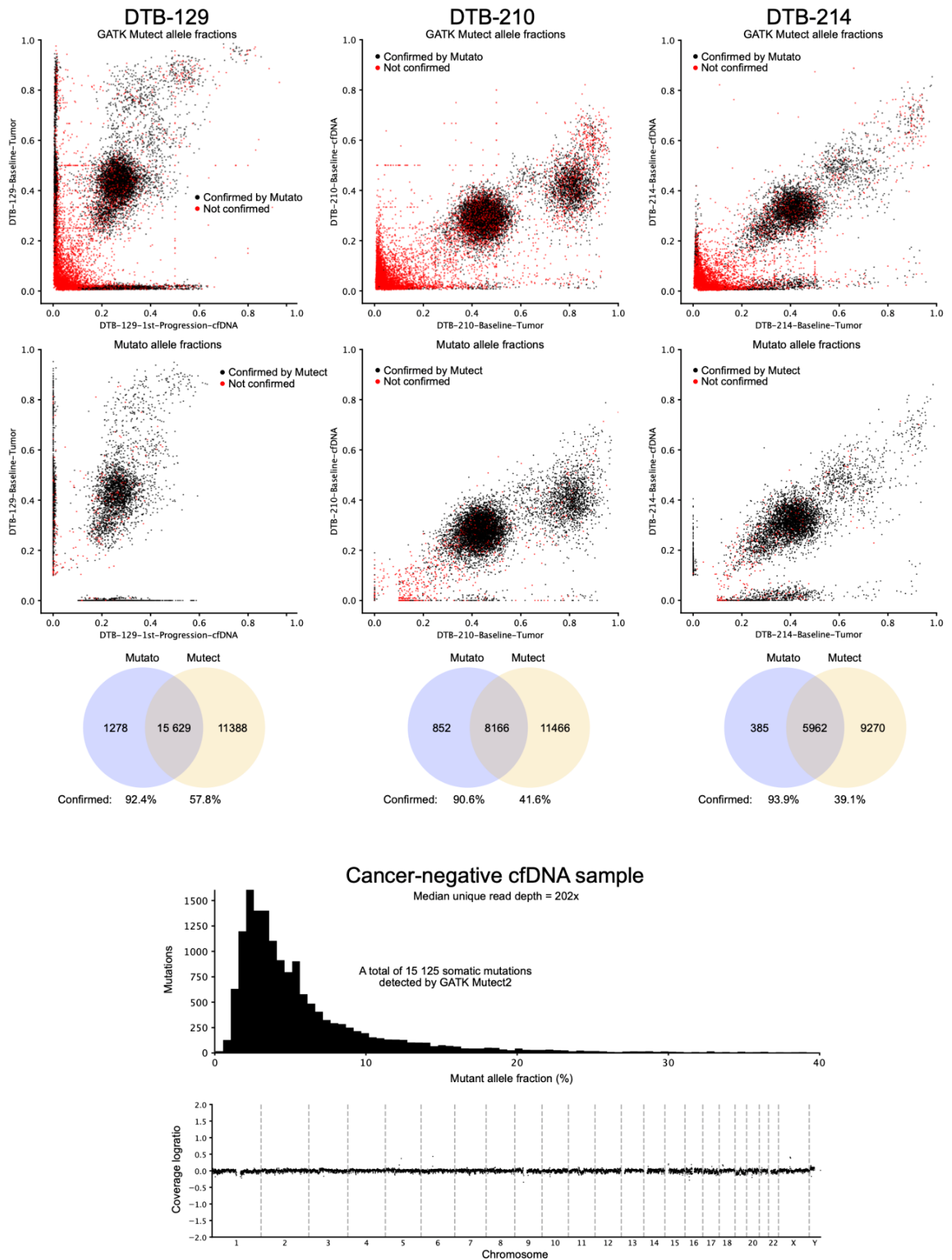
Validation of somatic mutation calls using GATK Mutect2

To estimate the false discovery rate of our somatic mutation calling methodology, we sequenced a ctDNA-negative plasma cfDNA sample from a 82-year old prostate cancer patient to a median depth of 202x, and a matched leukocyte DNA sample to a median depth of 27x. We analyzed these samples together with the rest of the cohort, using identical methodology. A total of 36 somatic mutations were detected in this control cfDNA sample, suggesting a false discovery rate of 1.0%, given that an average of 3779 mutations were detected in our cohort's samples. The ctDNA-negative nature of this plasma sample was previously determined based on targeted sequencing and validated via subsequent WGS, providing an upper bound for the sample's ctDNA fraction at $\leq 0.5\%$.

To validate our somatic mutation calls, we ran the published GATK Mutect2 mutation calling pipeline (version 4.2.3.0) ¹⁴ on the plasma cfDNA and metastatic tissue samples of 3 cancer patients in our cohort (4 plasma

cfDNA, 3 metastatic tissue, and 3 leukocyte DNA samples), as well as one ctDNA-negative plasma cfDNA sample. For maximum compatibility with GATK analysis, we realigned our adapter-trimmed sequencing data using BWA version 0.7.17¹⁵ with default parameters, and performed base quality recalibration using GATK BaseRecalibrator and GATK ApplyBQSR. Duplicate marking was carried out using SAMBLASTER¹⁶. After completing the reanalyzed BAM files, we ran GATK Mutect2 on a set of 36 leukocyte DNA samples and 5 cancer-negative plasma cfDNA samples to generate a panel of normals using GATK CreateSomaticPanelOfNormals. We then ran GATK Mutect2 on the samples of each patient, designating the leukocyte DNA samples as matched germline normals, and incorporating the panel of normals in each patient's analysis. We restricted the GATK Mutect2 mutation calls to the recommended genomic regions listed in the `broad_wgs_calling_regions.hg38.interval_list` file provided by the Broad Institute. A read orientation bias model was generated for each sample using the `--f1r2-tar-gz` parameter of Mutect2 and GATK LearnReadOrientationModel with default parameters. GATK FilterMutectCalls with default parameters was used to post-process and filter the identified mutation candidates, incorporating the read orientation bias model using the `--ob-priors` argument, as recommended in GATK documentation. Only mutations with the value PASS in the FILTER column of the resulting VCF files were counted as detected mutations.

A head-to-head comparison of the somatic mutation calls made by the two analysis pipelines revealed that GATK Mutect2 confirmed 92.2% of the mutations detected by our analysis. The GATK Mutect2 pipeline identified on average 2.06-fold more mutations than our pipeline, primarily at low allele fractions (**Supplementary Figure 15**). In the cancer-negative plasma cfDNA sample, GATK Mutect2 identified 15,125 somatic mutations, compared to 36 somatic mutations identified by our analysis. The false positive mutations in the ctDNA-negative sample had low allele fractions, similar to the discordant mutations detected by GATK Mutect2 in the cancer samples. GATK Mutect2 can be tuned for higher specificity (at the expense of sensitivity) by adjusting the `-f-score-beta` parameter, but for the purposes of validation we used the default value.



Supplementary Figure 15. Validation of somatic mutation calls using the GATK Mutect pipeline. Scatter plots show allele fractions of somatic mutations detected by GATK Mutect (top row) and Mutato (second row) in the tumor and plasma cfDNA samples of three patients. Mutations (dots) are colored to indicate whether the mutation was also independently detected by the other pipeline. Venn diagrams indicate the overlap between mutation calls. Histogram at the bottom shows allele fractions of mutations reported by GATK Mutect2 in a known ctDNA-negative cfDNA sample, as confirmed by the flat coverage logratio profile at the bottom.

Validation of structural variants using GRIDSS2

To validate the somatic structural variants (SVs) called by Breakfast, we ran the established SV-calling software suite GRIDSS2 (version 2.10.0) (<https://github.com/PapenfussLab/gridss>) on the plasma cfDNA and metastatic tissue samples of 9 patients in our cohort (9 plasma cfDNA, 2 metastatic tissue, and 8 leukocyte DNA samples), as well as one ctDNA-negative plasma sample¹⁷. To mitigate against unexpected tool behavior and maximize compatibility with the GRIDSS2 pipeline, we realigned the adapter-trimmed sequencing data to the hg38 reference genome with BWA-mem (version 0.7.15-r1140) using default parameters, and performed duplicate marking with Picard (version 2.18.0; <https://broadinstitute.github.io/picard/>)¹⁵. GRIDSS2 was run using default parameters on sample-patient tumor-normal sample pairs (i.e., joint calling mode as per recommended practices). GRIDSS2 output files were then post-processed using the somatic filtering tool GRIPSS (version 2.0) using default parameters and the default 'Panel of Normals' blacklist. Only SVs labeled as PASS in the FILTER column of the resulting post-GRIPSS VCF files were retained. The only custom filter applied was that we required all somatic SVs to have at least 15 supporting split-reads and be >10bp length (this latter criteria was also applied to Breakfast for validation).

DNA sequence homology between rearrangement breakpoints can make it impossible to localize the chromosomal breakpoints at a single base accuracy (i.e. multiple potential breakpoint locations can result in the observed junction sequence). Differences in how SV calling algorithms break ties between equally likely breakpoint locations mean that identical true-positive SVs may be reported with slightly different breakpoint coordinates. Therefore to evaluate concordance between Breakfast and GRIDSS2, we performed a fuzzy intersect between same-sample SVs identified by both approaches. Specifically, two SVs were considered matching if both their breakpoint start and end coordinates were within ± 6 bp of each other (i.e. $\text{abs}(\text{SVb_start} - \text{SVg_start}) \leq 6$) and $\text{abs}(\text{SVb_end} - \text{SVg_end}) \leq 6$; where SVb and SVg denote individual SVs as called by Breakfast and GRIDSS2). The reciprocal comparison was also made (i.e. comparing start to end and vice versa) to accommodate potential differences in genome annotation conventions between the two tools.

A head-to-head comparison of the somatic SV calls made by Breakfast and GRIDSS2 (using default parameters) revealed that GRIDSS2 recapitulated 63.4% (2046 of 3228) of the SVs originally identified by our analysis—with similar levels of redetection across all samples (median 62.1%), as well as all classes of SVs as determined by Breakfast (i.e. deletions, inversions, duplications, and translocations). GRIDSS2 identified on average 3.4-fold (range: 1.7-6.8) more SVs than our pipeline (among tumor-positive samples), and identified 349 SVs in the ctDNA-negative control sample (whereas Breakfast identified 18). There were no major differences in software concordance between tissue and cfDNA (median 51.4% vs. 62.1%, $p=0.20$). Of the somatic rearrangements detected by GRIDSS2, 50.1% were not unique to a single patient (i.e. GRIDSS2 had detected them in two or more patients). GRIDSS2-assigned QUAL scores were much lower for the non-unique rearrangements (median 1301 vs 1902, $p = 5.85\text{e-}41$, MWU). Since it is extremely unlikely that tumors from two patients would carry an identical somatic rearrangement, we conclude that at least half of the GRIDSS2 detected rearrangements in this cohort were false positives.

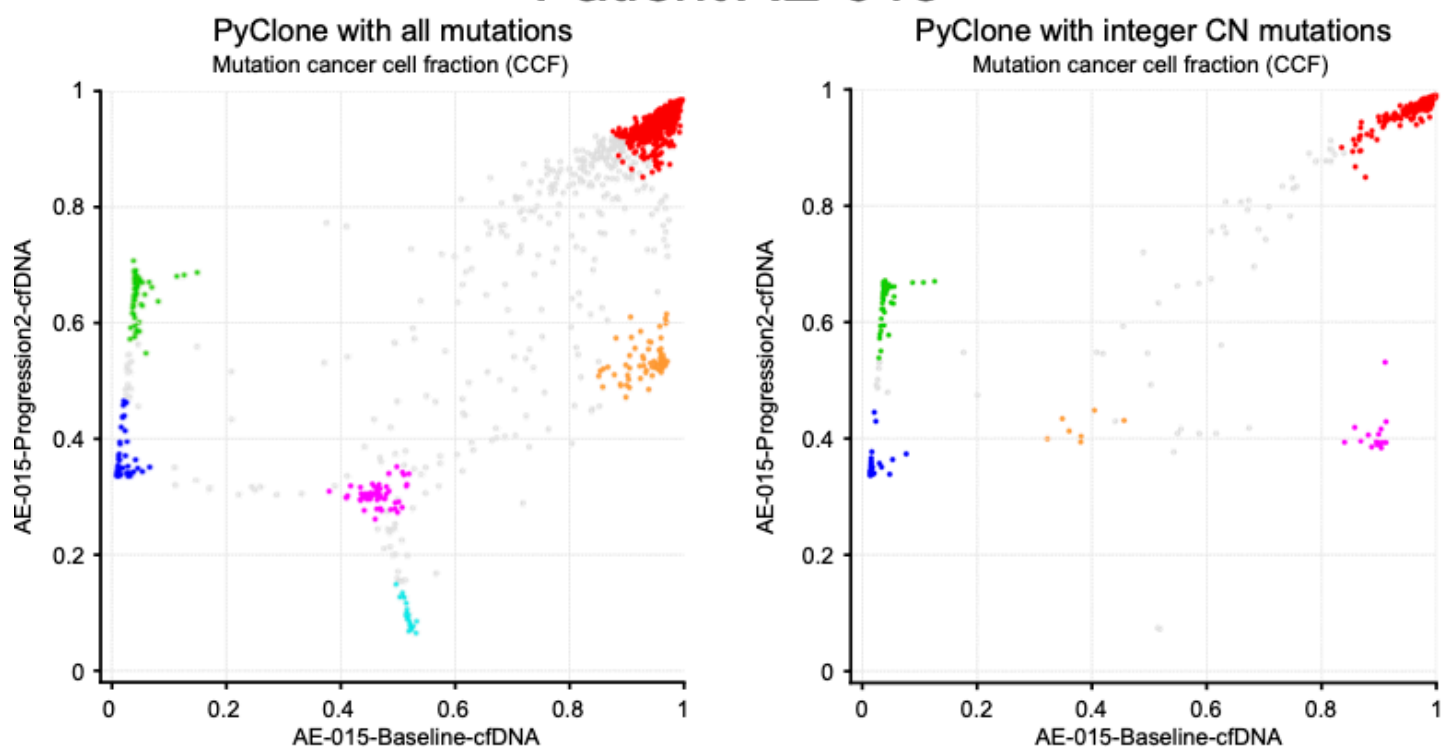
Validation of subclonal reconstruction using PyClone and PhyloWGS

We validated our mutation clustering step using PyClone version 0.13.1, employing its “run_analysis_pipeline” script with default parameters¹⁸. A cancer cell fraction was provided for each input sample, as per recommendations. The PyClone input file format allows the associated paternal and maternal allele copy number to be provided for each somatic mutation, but has no accommodation for mutations located in genomic regions that harbor subclonal CN heterogeneity (i.e. a sample harboring multiple subclones with a different CN state). The official PyClone documentation and publication do not provide guidance on how to handle subclonal CN heterogeneity. First, we provided PyClone with an input table of all somatic mutations in a patient's samples, including mutations in subclonal CN regions, and assigned each mutation an allele-specific CN state that was the closest fit the observed coverage logratio and HSAF (even if the closest fit was

imperfect). This resulted in PyClone identifying artefactual mutation clusters composed solely of mutations localized to subclonal CN regions (**Supplementary Figure 16**). PyClone calculates a CCF estimate for every somatic mutation, but adjusts the CCF estimates based on the identified clusters so that they fall within the range 0.0 - 1.0, explaining the unusual distribution of mutation CCFs. For patient AE-098 (with high mutation burden due to a somatic mismatch repair defect), PyClone analysis was terminated after executing at 100% CPU utilization for 30 days.

Next, we provided PyClone only with somatic mutations in genomic regions with no subclonal CN heterogeneity, the same way we did in our own analyses. The resulting cluster assignments were largely consistent with our clustering, when clusters with < 10 mutations were excluded from the analysis (see Supplement section “PyClone and PhyloWGS subclonal reconstruction validation data”). PyClone also identified hundreds of single-mutation clusters in each patient, representing individual outlier mutations.

Patient AE-015



Supplementary Figure 16. Scatter plot showing mutation cancer cell fraction (CCF) values inferred by PyClone for somatic mutation detected in two plasma cfDNA samples from patient AE-015. Mutations are colored according to PyClone cluster assignment. The left panel shows the result when all somatic mutations were included in the analysis. The right panel shows the result when only mutations located in genomic regions with an integer copy number (indicating absence of subclonal copy number heterogeneity) were analyzed. Mutations shown in light gray were assigned to clusters with less than 10 mutations.

We carried out additional validation of our mutation clustering and phylogenetic tree construction steps using PhyloWGS version 2018-11-05. We converted our somatic mutation calls and CN segments into formats supported by the `create_phylowgs_inputs.py` script, which then combined the multi-sample input data for each patient into two unified files (one for mutations and one for CN segments). We then used the PhyloWGS `multi_evolve.py` script to carry out the subclonal reconstruction for each patient, using default parameters and 4 concurrent MCMC chains (as recommended for accuracy in PhyloWGS documentation). Output files were summarized into JSON format using the PhyloWGS `write_results.py` script, and the best tree was determined based on normalized likelihood values utilizing the same formula used by the PhyloWGS `tree_viewer.js` implementation. Mutation cluster assignments were extracted from the best tree’s JSON file. Since PhyloWGS does not calculate a CCF or MACN value for individual somatic mutations, we visualized the PhyloWGS

mutation clustering results using mutation CCF values calculated by our methodology, and compared the cluster assignments against our own.

Both PyClone and PhyloWGS showed a tendency for modeling observed mutation clouds as multiple overlapping clusters. We believe this was partly caused by improper modeling of cluster CCF variances (i.e. how wide the mutation clusters are expected to be along each dimension). The documentation for both PyClone and PhyloWGS requests that somatic mutation allele fractions are provided in the input files as the number of mutation-supporting reads and total overlapping reads. But in paired-end sequencing the individual reads are not always independent observations, since paired mate reads from the same original DNA template display the same allele. Therefore if the variance of the mutation clusters is modeled using a binomial distribution $\text{Bin}(n, p)$ where n = total number of overlapping reads, the model will underestimate the true variance in CCF space. This artifact is particularly pronounced in plasma cfDNA samples, where DNA fragments are so short that paired end reads typically have significant mutual overlap. This increases the likelihood that model-based clustering will represent an observed mutation cluster as arising from multiple overlapping clusters.

Another potential confounding factor was the common occurrence of somatic mutations with CCF values falling on the diagonal between the truncal mutation cluster and the origin (CCF = 0). Most of these outlier mutations localized to genomic sites with challenging DNA sequences, and were thus explained as alignment artifacts and inaccurate allele fraction estimation. But it has also been shown in literature that all mutation clusters have an associated “tail” of mutations with increasing density towards the origin. These “tails” are believed to arise from neutral accumulation of mutations during the initial stages of clonal expansion¹⁹. In other words, when a cell acquires a proliferative advantage and creates a new clonal expansion, the new mutations that accumulate during the earliest subsequent cell divisions will be present in a significant fraction of cells in the new population. These mutations “tails” (whether genuine biological events or technical artifacts) can theoretically result in the detection of spurious clusters, although in practice these mutations are often sparse.

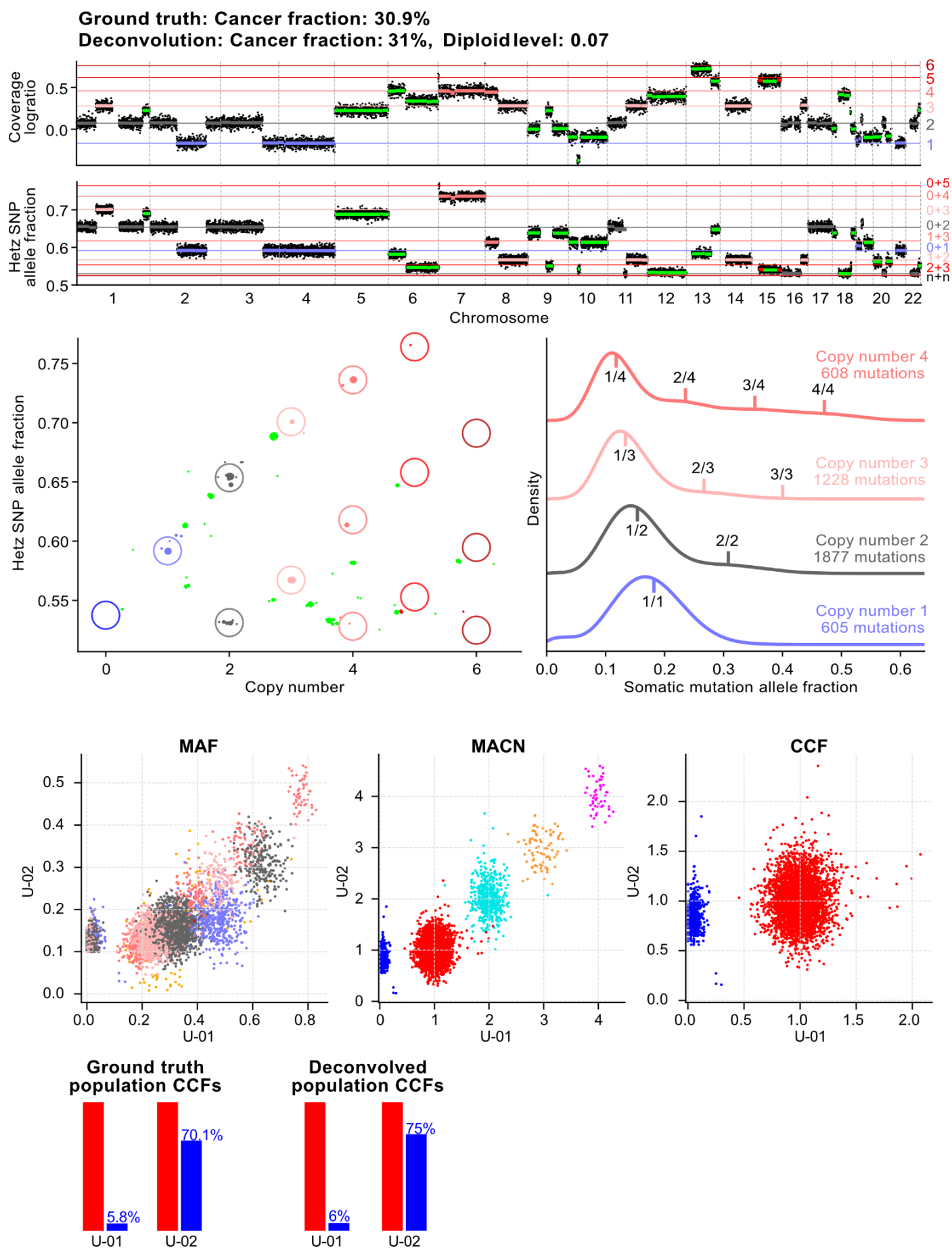
Validation of mutation CCF clusters using a generative simulation

To validate the accuracy of our mutation clustering step, we implemented a simulation framework that takes as input the proposed number of subclones, the CCFs of each subclone in each sample, the cancer fraction of each sample, the number of mutations in each subclone, the sequencing depth of those mutations, and their copy number state. The simulation framework then generates a CCF scatterplot by sampling each mutation’s MAF from a binomial distribution and transforming the MAF to a CCF value using the same approach we utilized for real data. This generative simulation accurately recapitulated the shape (variance) of the observed mutation clusters, demonstrating that our proposed subclonal models provide a parsimonious explanation of the observed mutation data (see Supplement section “PyClone and PhyloWGS subclonal reconstruction validation data”). To make the simulation recapitulate clusters partially truncated by mutation detection thresholds, we resampled any simulated mutations whose mutant read count and allele fraction fell below mutation detection thresholds in all patient samples.

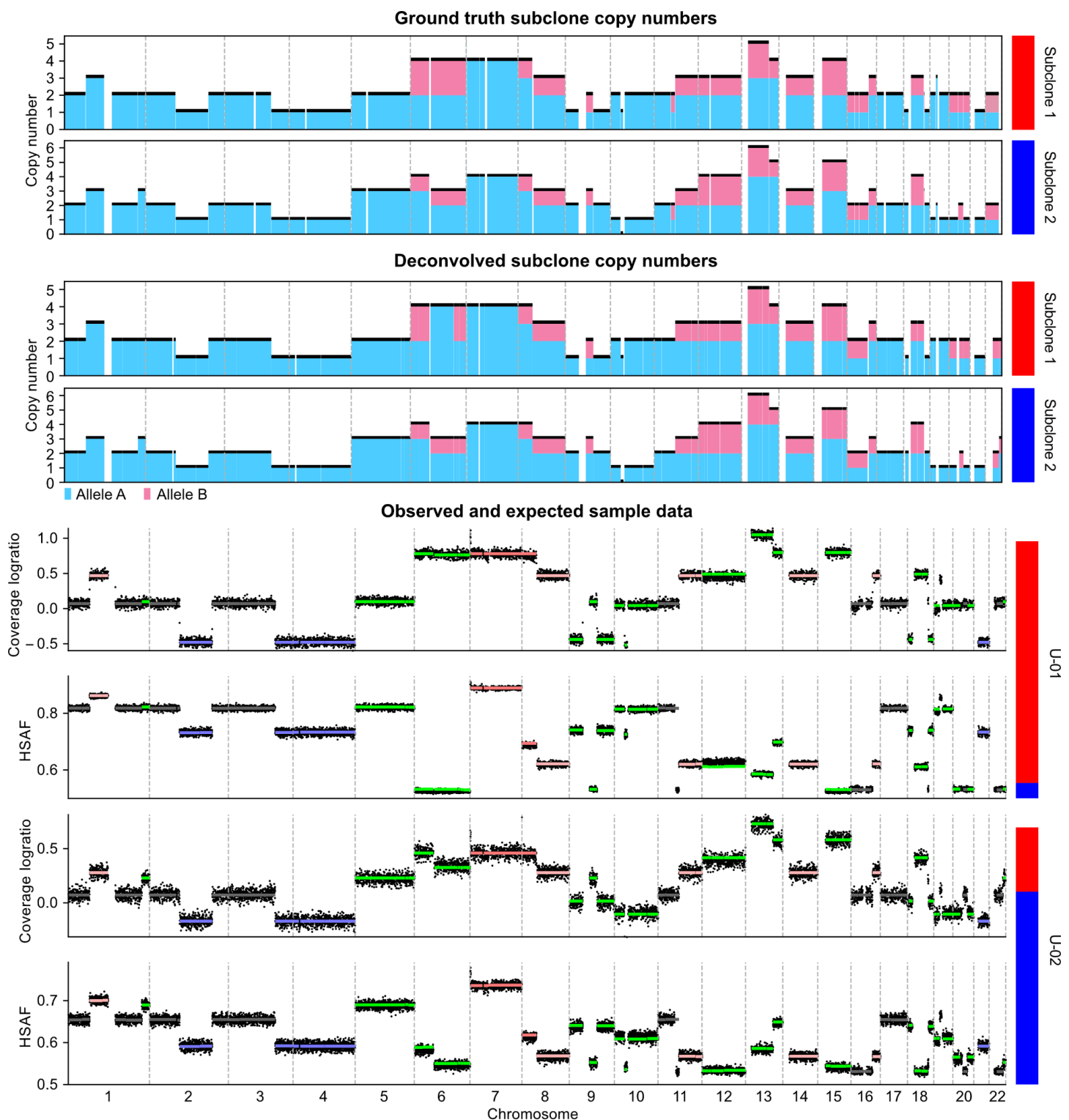
Validation of subclonal reconstruction using a blinded simulation

To validate our process of subclonal reconstruction in its entirety (including the truncal copy number model fitting, mutation clustering, and CN deconvolution steps), we implemented an algorithm that generates a family of phylogenetically related cancer cell populations by sequentially applying random somatic mutations and copy number alterations (including whole genome duplications in 50% of patients) to a simulated cancer genome. The algorithm tracks the copy numbers of the parental and maternal allele at each genomic locus, and the number of mutant allele copies for each somatic mutation, so that copy number alterations affect mutant allele copy numbers in a realistic fashion, and completely deleted alleles are never recovered through subsequent copy number gains (**code available:** <https://github.com/annalam/cfdna-wgs-manuscript-code>). We generated 22 patient cases with two samples each, for a total of 44 simulated samples, each carrying a

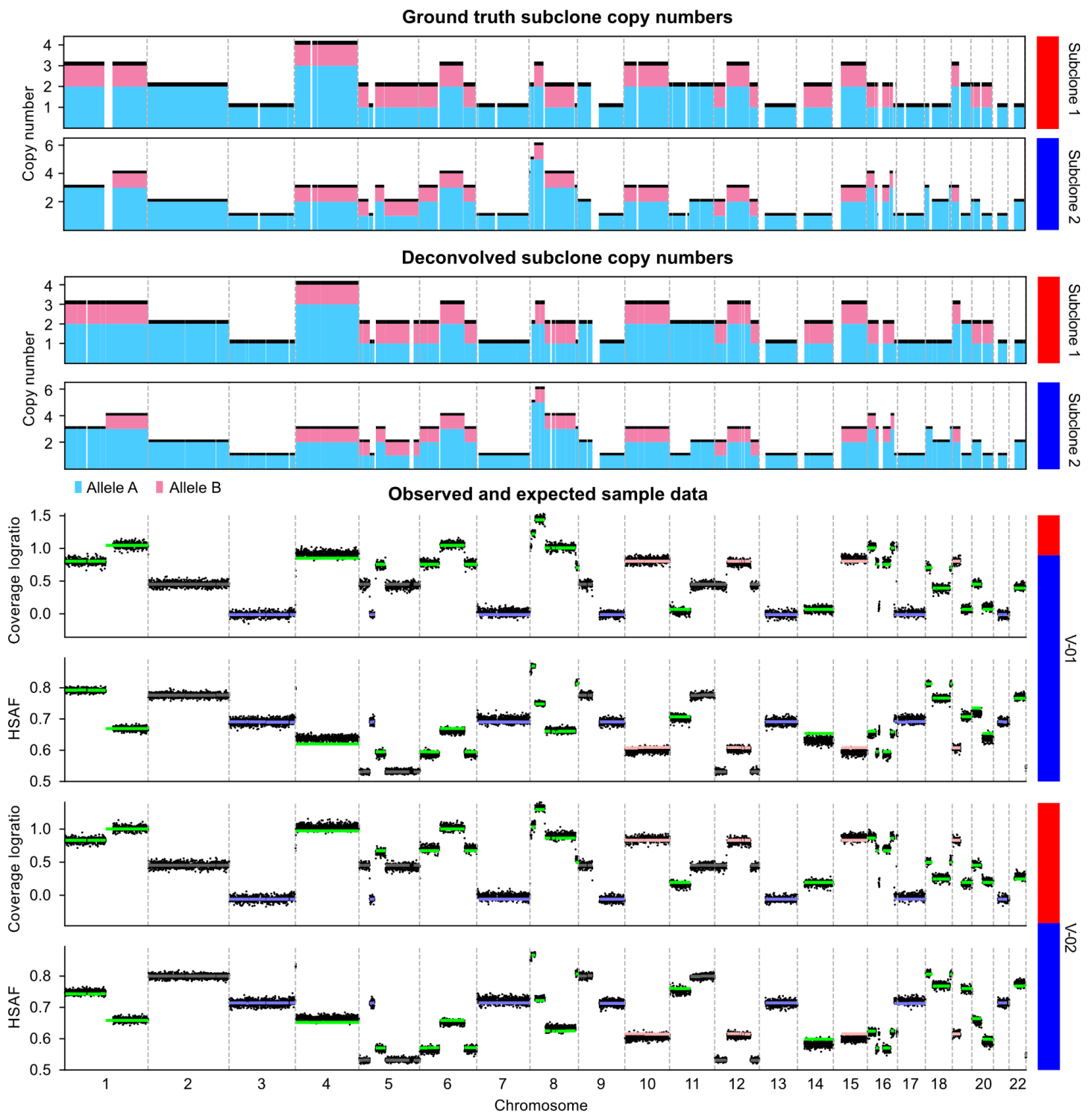
randomized cancer fraction (30 - 80%) and 2 - 3 cancer cell subpopulations. For each sample, the simulation produced realistic coverage logratios, HSAFs, and somatic mutation allele fractions based on the underlying ground truth (**Supplementary Figures 17-18**). We performed a blind subclonal reconstruction to infer the cancer fractions and subclone CCF fractions in each sample, as well as subclone-specific CN profiles (**Supplementary Figure 19**). The resulting cancer fraction estimates were all within 1.0 percent points of the ground truth value (average error 0.248 percent points; **Supplementary Table 5**). Estimates of the CCF of the dominant non-truncal subclone in each sample were also highly accurate, with an average error of 2.47 percent points. Average fraction of the genome with correct allele-specific copy numbers for the truncal and the dominant non-truncal populations was 99.4%.



Supplementary Figure 17. Representative example of a truncal model fit for simulated patient “U”. Truncal copy number model fit to simulated sample U-02 is shown in the top panels. Mutation-based subclonal reconstruction is shown in the bottom panels.

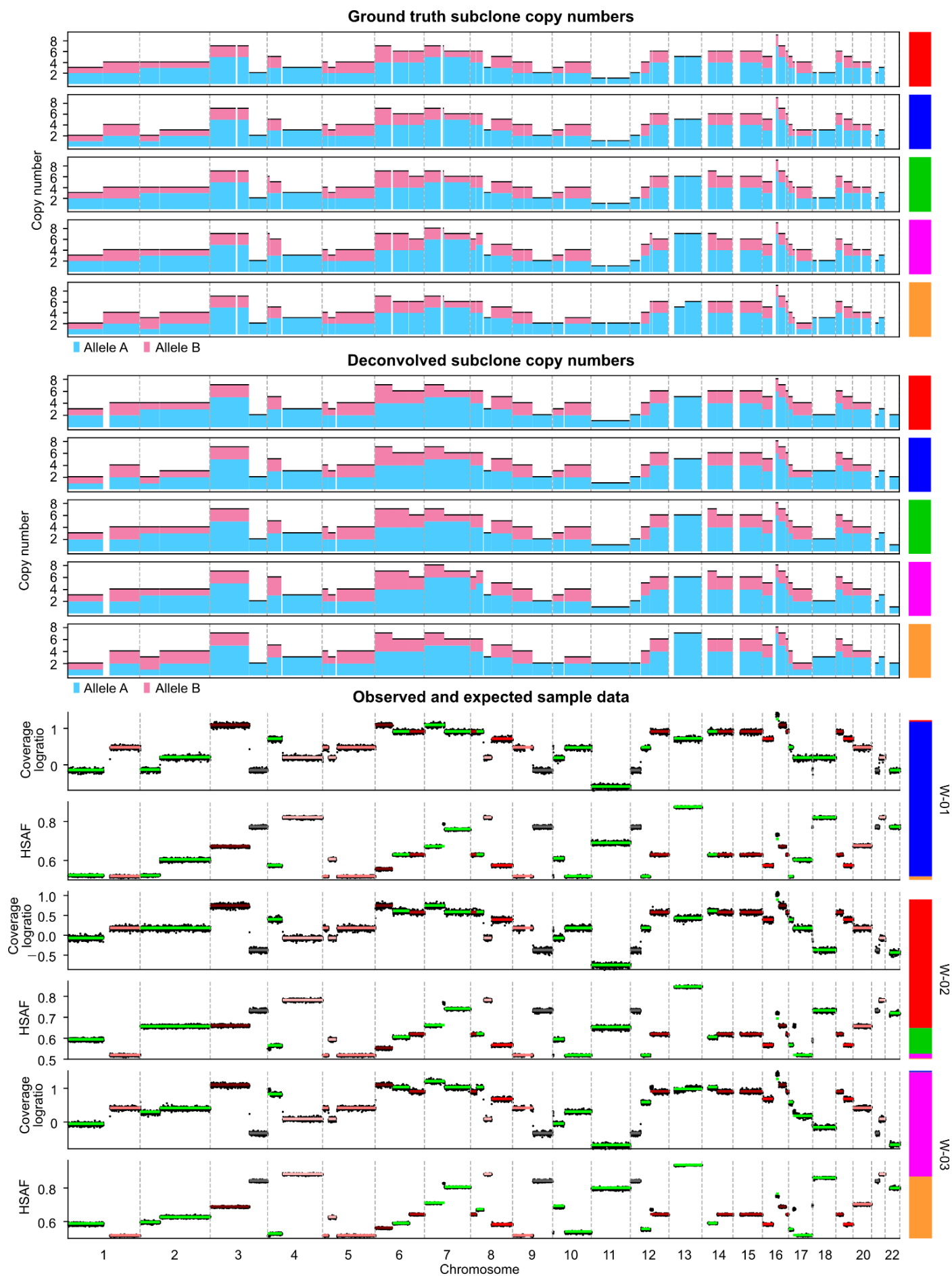


Supplementary Figure 18. Deconvolution of subclone copy number profiles for simulated patient “U”. The ground truth subclone copy number profiles are shown at the top, and below that the subclone copy number profiles inferred through a blinded analysis (ground truth was not known to the investigators). Observed coverage logratio and HSAF values (black dots) in the two samples U-01 and U-02 are shown at the bottom, overlaid with their mathematically expected values (horizontal segments) based on the inferred subclone copy number profiles. Subclone CCFs are indicated by the stacked bars on the right side.



Supplementary Figure 19. Deconvolution of subclone copy number profiles for simulated patient “V”.

As an added validation, we generated 4 complex patient cases, each with 3 samples and 5 cancer cell subpopulations in a randomly generated phylogenetic tree. Blinded analysis of these simulated patient cases identified a correct truncal copy number model for all 12 samples, and a correct phylogenetic tree for all 4 patients. Identified subclone fractions were highly accurate (average absolute error 1.3 percentage units). Copy number deconvolution analyses were also accurate, with an average 99.1% of the genome assigned with the correct copy number state in each cancer cell population (**Supplementary Table 5**). A representative example can be found in **Supplementary Figures 20 - 21**.

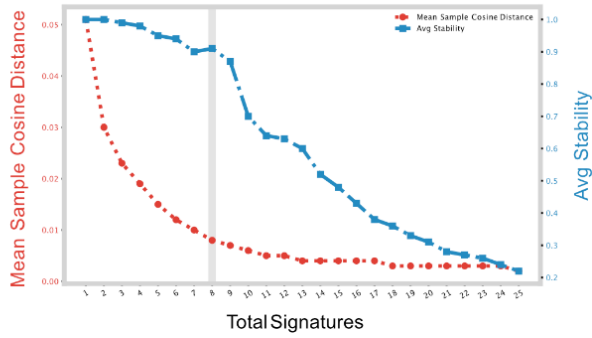


Supplementary Figure 21. Deconvolution of subclone copy number profiles for simulated patient “W”.

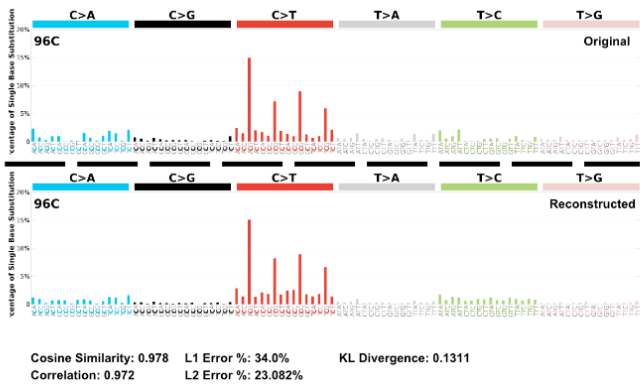
Validation of mutational signature decomposition on bulk data using SigProfileExtractor

To verify that the COSMIC v2 mutational signatures (which were originally inferred from pan-cancer tissue-based WGS) are appropriate for application in prostate-cancer ctDNA, we applied SigProfileExtractor across all bulk cfDNA sequencing samples (analyzed together) to generate *de novo* base substitution trinucleotide context mutation signatures, using default tool parameters (i.e., `cosmic_version=2`, `context_type='96'`, `reference_genome="GRCh38"`, `opportunity_genome='GRCh38'`, `minimum_signatures=1`, `maximum_signatures=25`, `nmf_replicates=100`)^{20,21}. SigProfileExtractor inferred a total of 8 *de novo* signatures (**Supplementary Figure 22**). Dominant *de novo* signatures include a signature characterized by frequent C>T transitions appearing overall highly similar to the COSMIC 1 aging signature (cosine similarity: 0.84). From bulk decomposition of all samples, SigProfileExtractor also identified *de novo* signatures with high cosine similarities to established COSMIC v2 signatures associated with mismatch-repair (MMRd) and homologous recombination repair deficiency (HRRd). Expectedly, the *de novo* mutation signatures associated with MMRd and HRRd were most strongly represented in samples with putatively causal deleterious defects in *MSH2/6* and *BRCA2*, respectively (**Supplementary Figure 22**). Prior publications (in other cancers but also in mCRPC tissue²²) have implied that the majority of *de novo* signatures inferred from single cancer-type bulk NGS data almost exclusively recapitulate established signatures extracted from larger pan-cancer efforts^{20,21}. Our analysis using SigProfileExtractor in mCRPC ctDNA and metastatic tissue broadly reaffirms this observation and credentials the use of COSMIC signatures for the characterization of our samples.

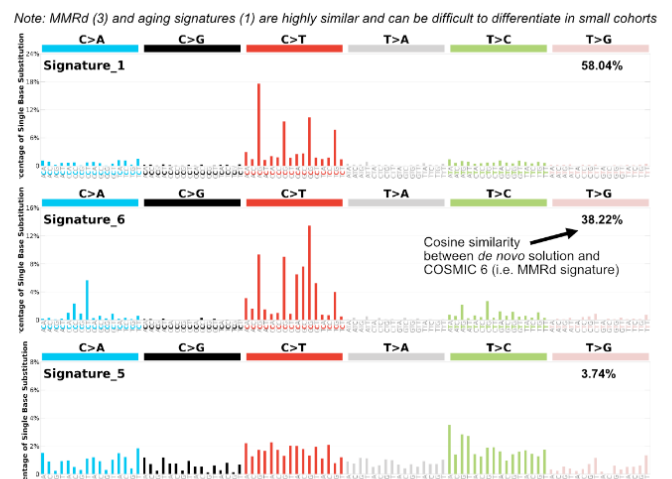
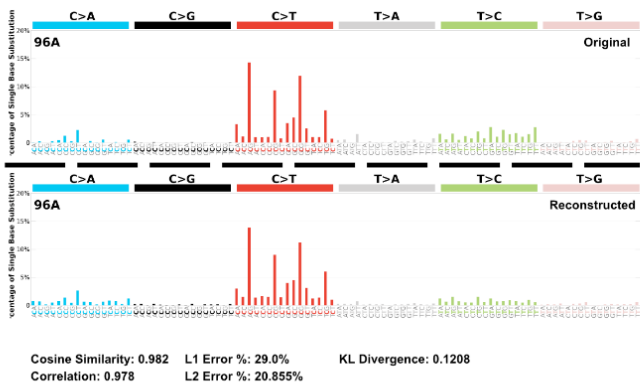
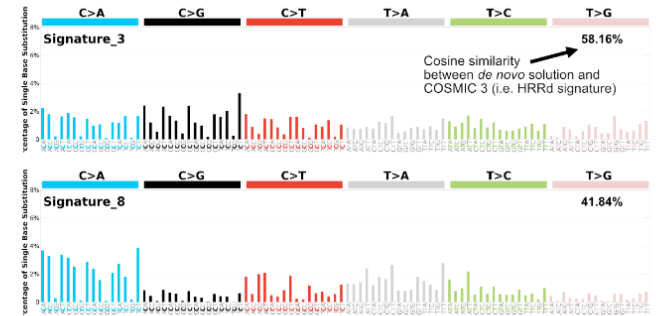
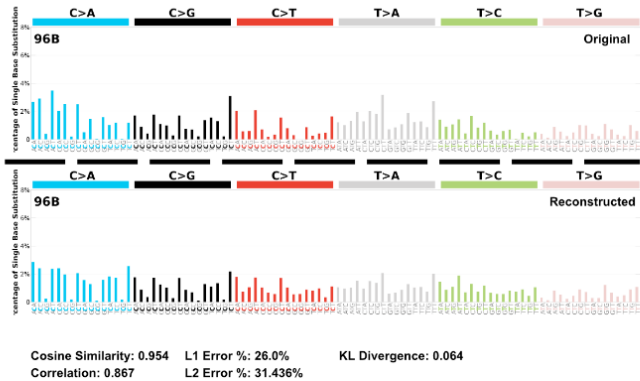
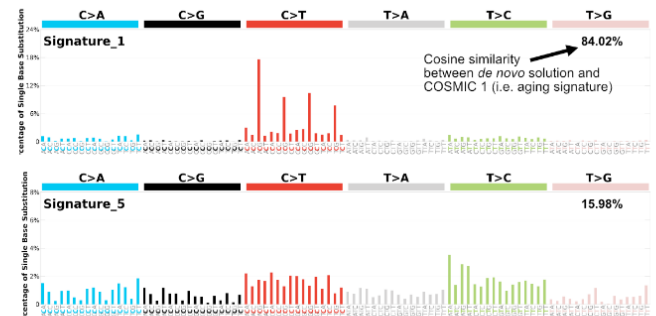
Signature selection (n=76 WGS samples)



De novo SBS mutation signatures (from 76 cfDNA and metastatic tissue mCRPC samples)



Reconstructed COSMIC v2 signatures



Supplementary Figure 22. Derivation of the optimal number of *de novo* signatures via SigProfileExtractor applied to bulk sequencing data from 76 whole-genome ctDNA and metastatic tissue mCRPC samples. The

three most dominant *de novo* single-base substitution (SBS) mutation signatures inferred from our cohort. All *de novo* signatures (trinucleotide context plots shown left) broadly recapitulate established COSMIC signatures with known activity in prostate cancer (reconstructions shown right).

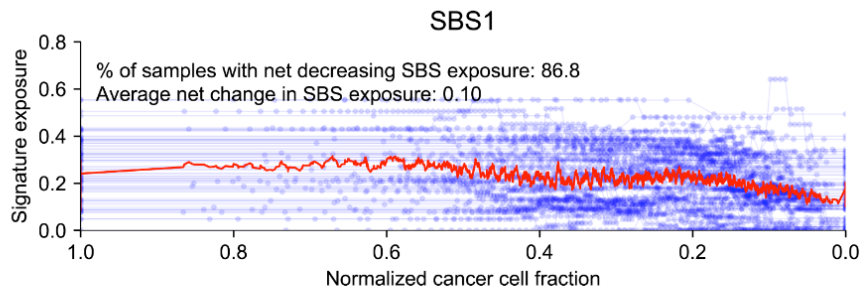
Validation of evolutionary changes in mutational signature activity using TrackSigFreq

To provide orthogonal validation of changing mutational signature activity over evolutionary pseudotime, we applied TrackSigFreq on each individual sample in our cohort (using default tool parameters and the COSMIC v2 signature catalog)²³. TrackSigFreq approximately orders mutations in descending order of their cancer clonal fraction (i.e. VAF corrected for copy-number state and tumor purity), using CCF as a surrogate for evolutionary time. TrackSigFreq does not require prior mutation CCF clustering, excluding this as a potential source of technical error. We supplied TrackSigFreq with tumor purity estimates and integerized total copy number from our automated truncal copy number model fitting (see Supplementary Information section “Copy number model fitting and cancer fraction estimation”), and SNVs identified from Mutato (excluding SNVs in regions harboring subclonal copy number heterogeneity). All SNVs were first converted from hg38 to hg19 using CrossMap v0.2.7 (TrackSigFreq uses hg19 by default), and sample VCF files were capped at 8000 SNVs (i.e. non-randomly subsampled) to improve runtime²⁴. To mitigate potential signature overfitting and to accommodate characterization of signatures linked to molecular subtypes represented in our cohort (e.g. mismatch-repair deficiency), we ran TrackSigFreq using an expanded list of all 30 single-based substitution (SBS) signatures (rather than TrackSigFreq’s default of restricted fitting to cancer-specific signatures inferred via detectActiveSignatures).

87% of cfDNA or metastatic tissue samples showed a net temporal decrease in COSMIC SBS signature 1 (i.e. aging signature). For all samples with deleterious *BRCA2* or mismatch-repair defects, TrackSigFreq revealed temporarily increasing signature 3 and 6 trajectories (respectively) coupled with temporally decreasing aging signatures (**Supplementary Figure 23-24**). We also generated TrackSigFreq signature trajectories using a fixed set of 7 signatures (COSMIC v2: 1, 3, 5, 6, 9, 15, 21) applied consistently across all samples but observed no major difference in trajectories (apart from higher average signature weights as expected). Collectively these results broadly validate the temporal variation in aging and DNA repair-associated signature weights as determined from initial signature reconstruction (using DeconstructSigs) of discrete clonal compartments (from subclonal reconstruction)—**Figure 3e**.

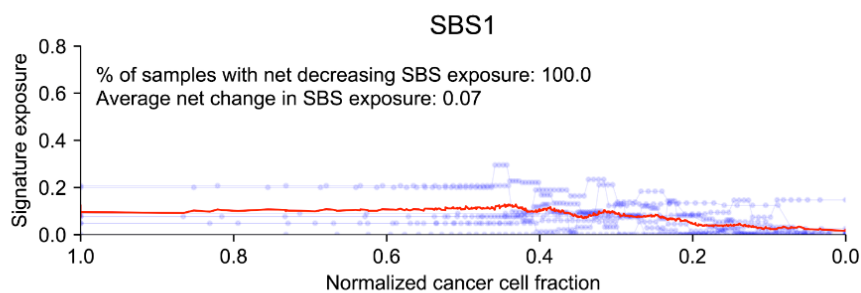
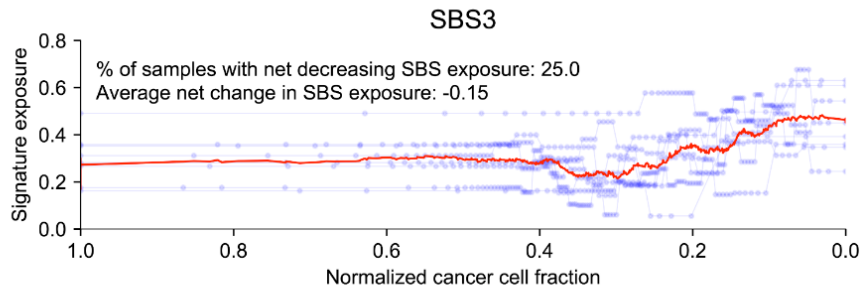
a

All samples

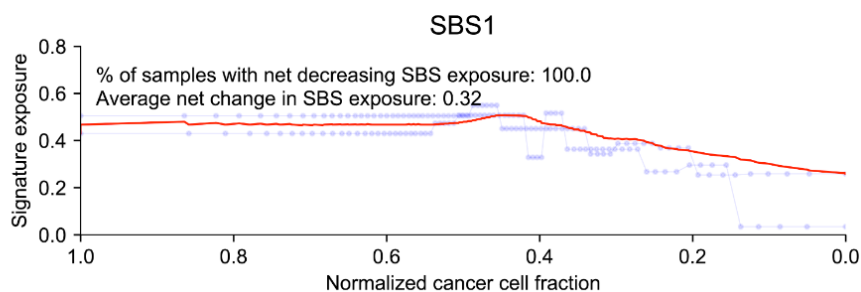
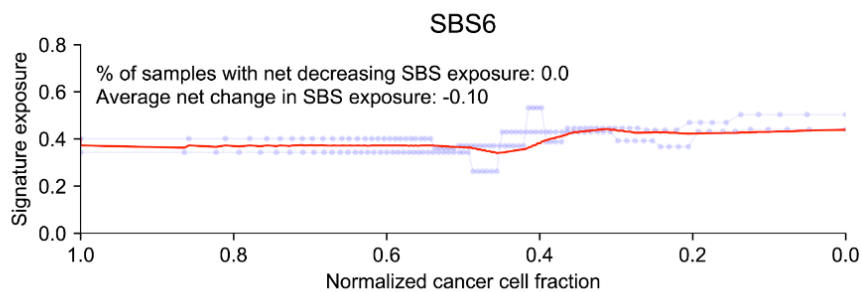


b

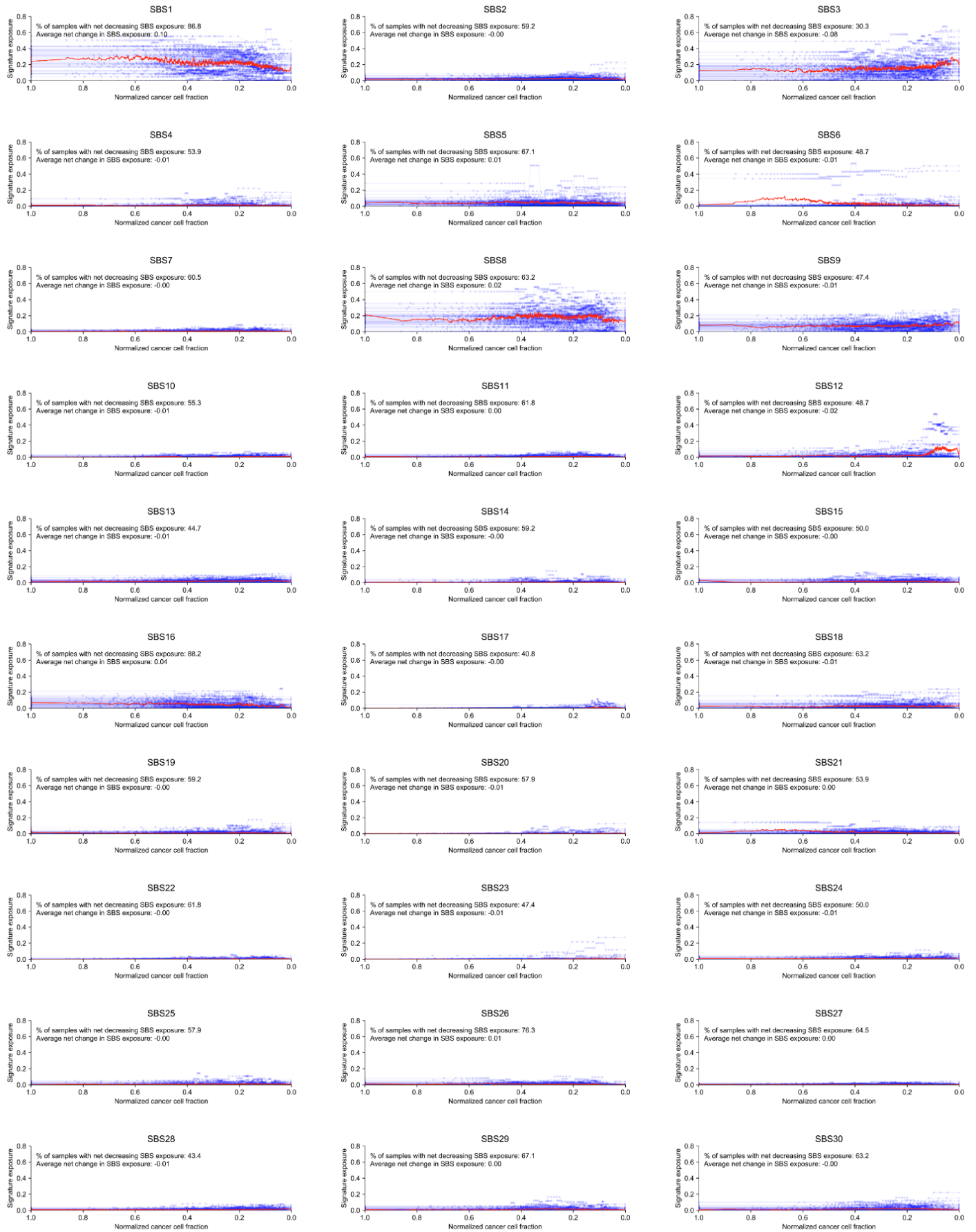
Only *BRCA2*
deficient samples



Only *MSH2/MSH6*
deficient samples



Supplementary Figure 23. Aggregated TrackSigFreq signature weight trajectories in **(a)** all 76 cfDNA and metastatic tissue WGS samples in our cohort and **(b)** the subset of samples with known DNA repair deficiencies. Only select mutational signatures are displayed (the mitotic ‘aging’ signature as well as signatures associated with DNA repair deficiencies in samples with putatively causal defects). Blue lines show individual sample trajectories (where dots indicate signature exposure changepoints). To visualize all per-sample trajectories on the same horizontal axis, individual per-sample cancer cell fraction (CCF) values were minmax normalized to fall between [0,1]. Red lines show the rolling average signature trajectory.



Supplementary Figure 24. Aggregated TrackSigFreq signature weight trajectories for all 76 cfDNA and metastatic tissue WGS samples in our cohort shown for all 30 COSMIC v2 signatures.

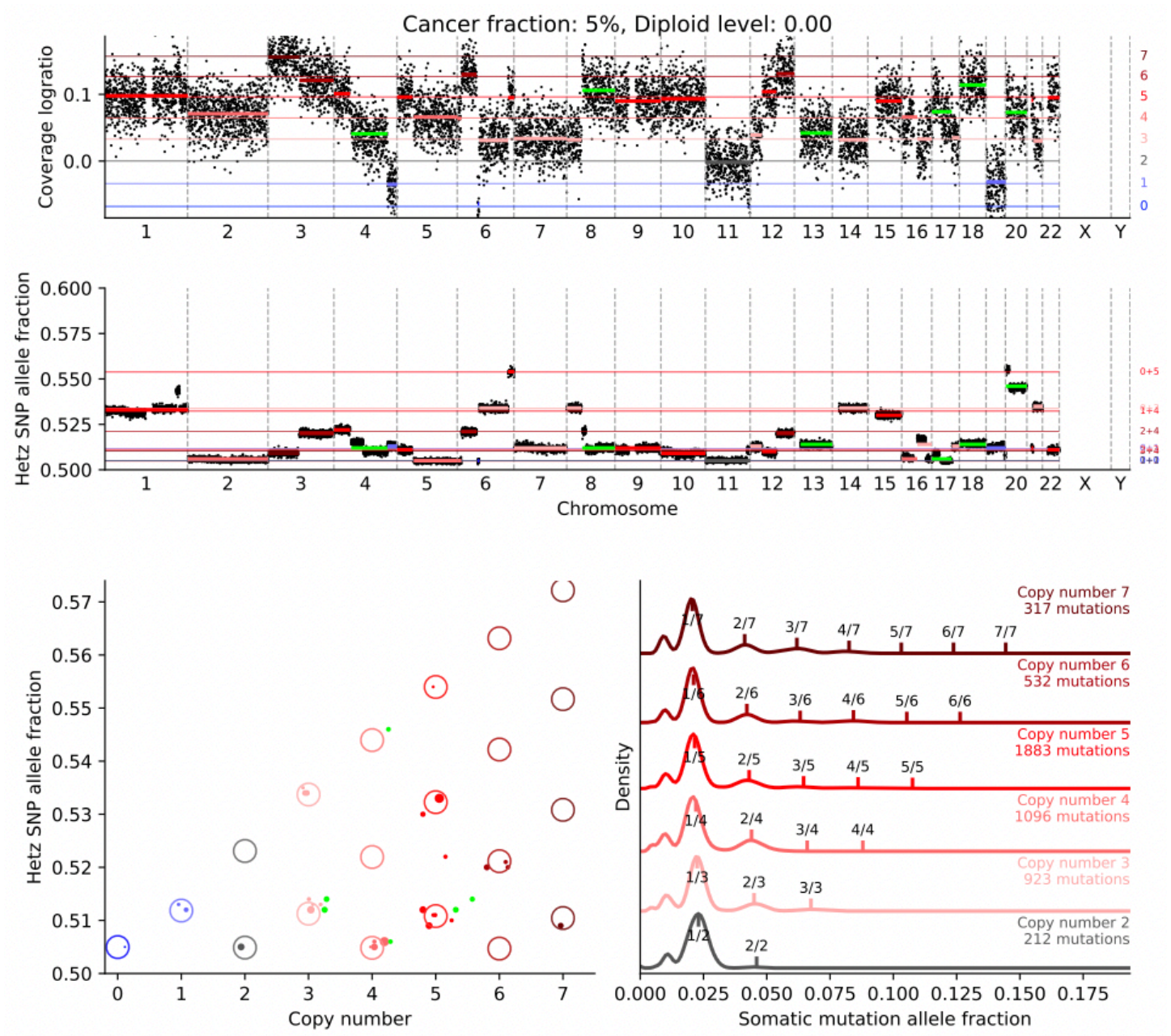
We also directly tested the variance in fitted mutation signature weights between same-patient subclones. To do this, we reapplied DeconstructSigs to each separate per-patient subclonal population (i.e. including each colored segment of a phylogenetic tree in Figure 2b) inferred from our multisample subclonal reconstruction²⁵. For each DeconstructSigs COSMIC v2 signature, we first calculated the variance in signature fraction across same-patient subclones and then calculated a weighted average of variances over all patients (i.e. a pooled variance). Weights were equal to the number of same-patient subclones. The median pooled variance across all COSMIC v2 signatures was $\sigma^2 = 9.9 \times 10^{-4}$ (the mean signature fraction is $\bar{x} = 2.9 \times 10^{-2}$ across all subclones). Low inter-subclone variance ($\sigma^2 = 5.0 \times 10^{-3}$) was also observed even after exclusively retaining signatures with nonzero fraction in at least one same-patient subclone (whose prior inclusion would *decrease* overall variance). Finally, inter-subclone variance appeared similarly low for COSMIC signatures 1 (aging: $\sigma^2 = 2.4 \times 10^{-2}$), 3 (homologous recombination repair deficiency: $\sigma^2 = 3.1 \times 10^{-2}$), 6 and 21 (mismatch-repair associated: $\sigma^2 = 8.6 \times 10^{-3}$ and $\sigma^2 = 4.2 \times 10^{-3}$, respectively), which constitute the main focus of the data presented in Figure 2.

Validation of sample identities via SNP profile comparison

To validate sample identities, we determined high confidence genotypes for known human SNPs (GNOMAD database version 3) in each sample. A SNP genotype was considered high confidence if a sample had 30x or higher sequencing depth at the locus, and had an allele fraction between 30 - 70% (heterozygous) or within the ranges 0 - 1% or 99 - 100% (homozygous). Two samples were considered to have a matching SNP profile if at least 99% of high confidence genotypes were in agreement between the samples. SNPs that did not have a high confidence genotype in both compared samples were omitted from analysis.

Simulation of subclonal reconstruction in 5% ctDNA sample

To investigate the physical limits of genomic and subclonal reconstruction in plasma samples with low ctDNA fractions, we used our simulation framework to generate a plasma sample with 4.6% ctDNA, 5000x median fragment depth, truncal whole genome duplication, and a 42.5% subclone. We then ran our truncal copy number model fitting algorithm on the sample. The algorithm was able to identify the WGD, and the truncal copy number profile (**Supplementary Figure 25**). Since we have not whole genome sequenced samples to this depth yet, we cannot know the precise amount of coverage logratio noise in 5000x depth data (as it is composed of both sampling noise that decreases as depth increases, and depth-independent sources of noise such as nucleosome footprint and GC bias), but we conservatively estimated the noise level as 70% of the noise level seen for 200x depth cfDNA WGS data. On the other hand, for heterozygous SNP allele fractions in WGS data we have observed no deviations from pure binomial sampling noise, and believe that this simulation accurately models their expected noise level.



Supplementary Figure 25. Truncal copy number model fitting for a simulated plasma sample with a ground truth of 4.6% ctDNA, truncal whole genome duplication, and a 42.5% subclone.

Patient truncal copy number models and subclonal reconstructions

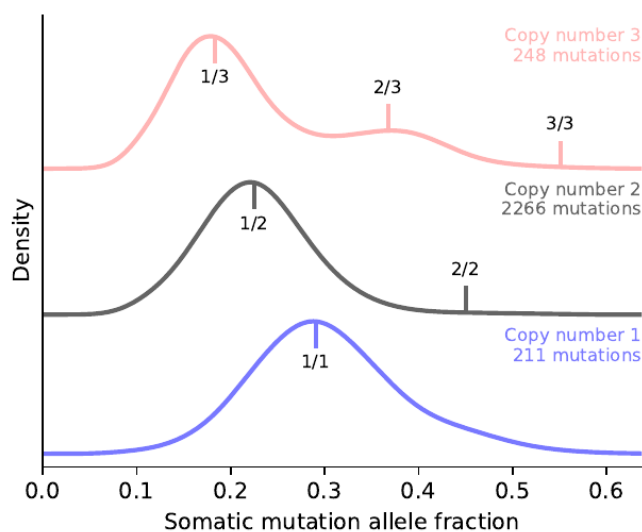
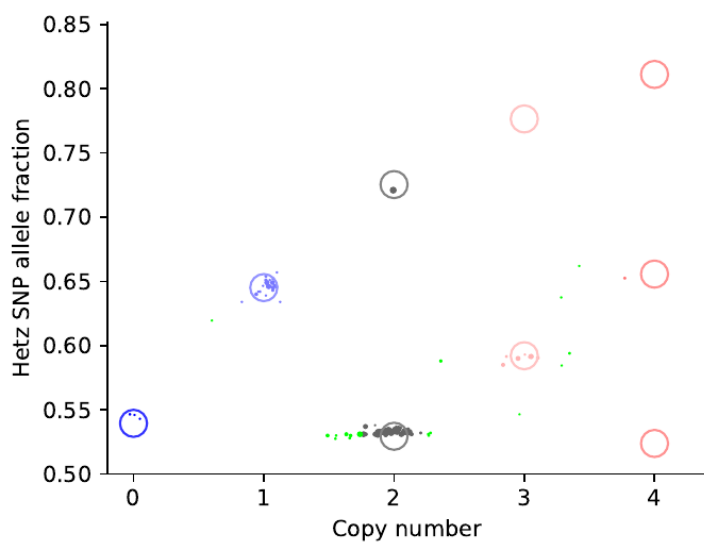
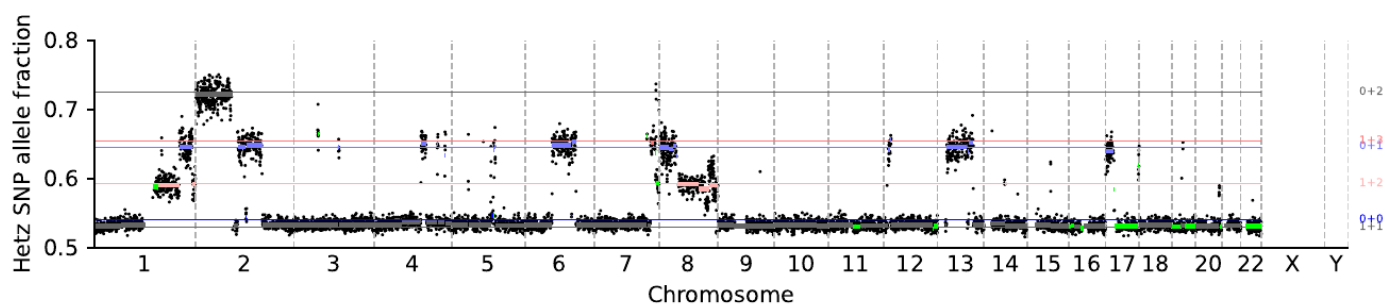
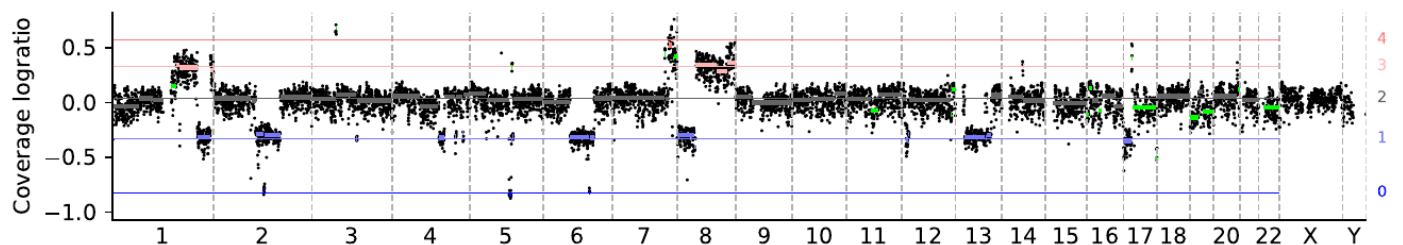
In this section, we provide visualizations of inferred truncal copy number models and subclonal reconstructions for all patients and samples in the cohort. The visualizations are shown in the following order for each patient:

- 1) Truncal copy number models for each cfDNA and tumor sample, each taking up a full page. At the top of each page is the sample name, cancer fraction, and diploid level (i.e. the coverage logratio level corresponding to a copy number of two). Below that are two panels showing the coverage logratio and heterozygous SNP allele fraction (HSAF) data, with colored horizontal lines indicating the mathematically expected coverage logratio and HSAF values corresponding to different allele-specific copy number states. The bottom left panel shows a *constellation diagram* where each dot is a genomic segment (size indicates segment length), and the colored circles indicate expected copy number and HSAF values corresponding to different copy number states. Dots on a trajectory between two circles indicate genomic regions with a subclonally heterogeneous copy number state. The bottom right panel shows kernel density estimates of somatic mutation allele fractions in genomic regions with different integer copy number states (mutations in genomic regions with subclonal CN heterogeneity are not shown).
- 2) Somatic mutation clustering in MACN and CCF space, to identify cancer subclones. The top row shows two-sample scatter plots of somatic mutations (dots) in MAF, MACN and CCF space. Only somatic mutations located in genomic regions with an integer copy number state in all samples are shown. MAFs are colored according to their total copy number (orange: samples have differing copy numbers). MACNs and CCFs are colored according to their cluster assignment. The next set of panels show mutation CCF histograms for the individual mutation clusters. Dotted vertical lines indicate the lowest CCF value that a MACN = 1 mutation can have and still be independently detected in the sample. This threshold depends on a genomic region's copy number, so multiple vertical lines representing different copy numbers are shown. The bottom panel shows all valid phylogenetic trees for the observed subclone CCF data. Most patients had only one valid phylogenetic tree. Horizontal length of the phylogenetic tree branches indicates the estimated number of somatic mutations acquired in that branch.
- 3) Genomic positions of the mutations belonging to each identified cluster. For each sample, a coverage logratio and HSAF plot is shown, and above that multiple colored tracks corresponding to the identified mutation clusters. Mutation positions are indicated with colored ticks within each track. Genomic regions with a subclonally heterogeneous CN state were excluded from analysis and are shown in light gray.
- 4) Deconvolution of subclone copy number profiles. The top panels show the subclone-specific copy number profiles (with color-coded bars on the right to indicate the subclone in question). Bottom panels show observed coverage logratio and HSAF values (black dots) in the patient samples, overlaid with their mathematically expected values (horizontal segments) based on the inferred subclone copy number profiles. Subclone CCFs are indicated by the stacked bars on the right side.
- 5) Clonal expansion timeline showing changes in patient tumor burden and subclone CCFs over time. PSA measurements (indicated with small black dots) were used as a proxy for tumor burden, with monotonic interpolation between timepoints. Subclone CCFs were quantified at plasma cfDNA sample collection timepoints (large black dots), with exponential interpolation between timepoints. Drug treatments received by the patient are indicated as horizontal bar segments at the top of the plot.

Patient AE-015

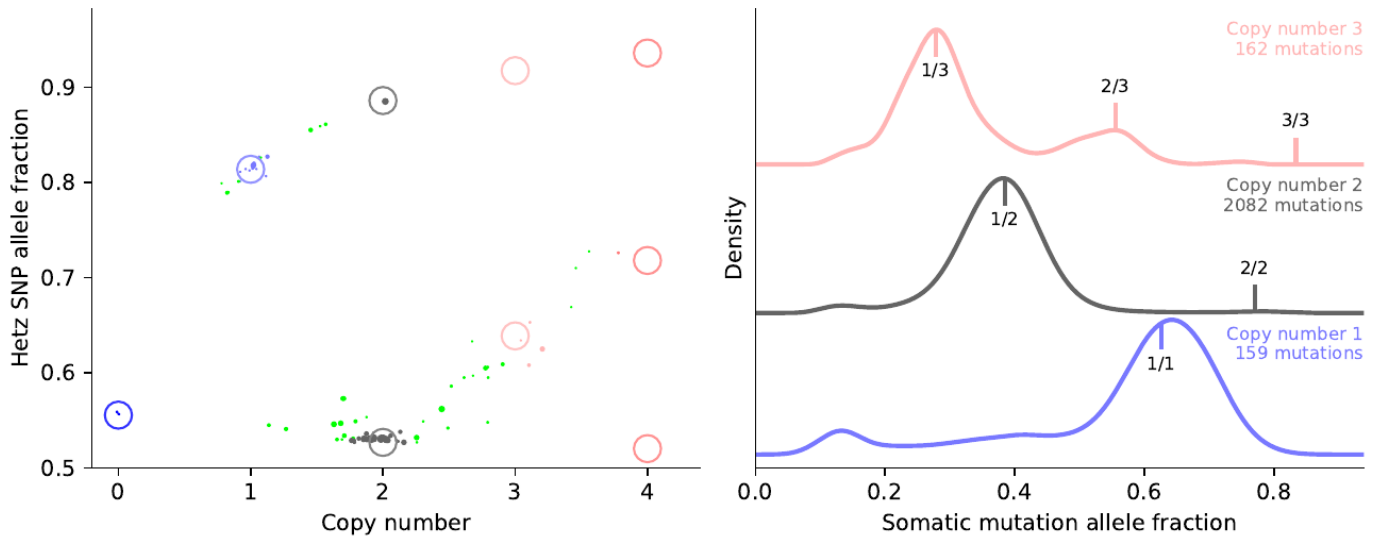
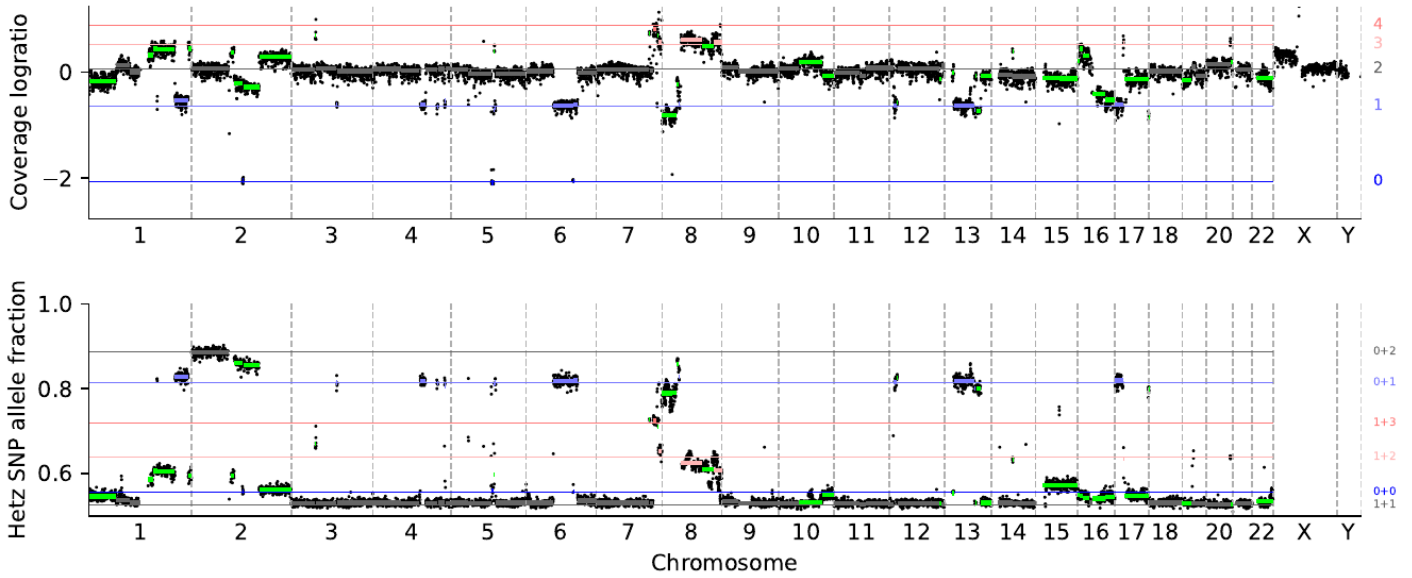
AE-015-Baseline-cfDNA

Cancer fraction: 45%, Diploid level: 0.04

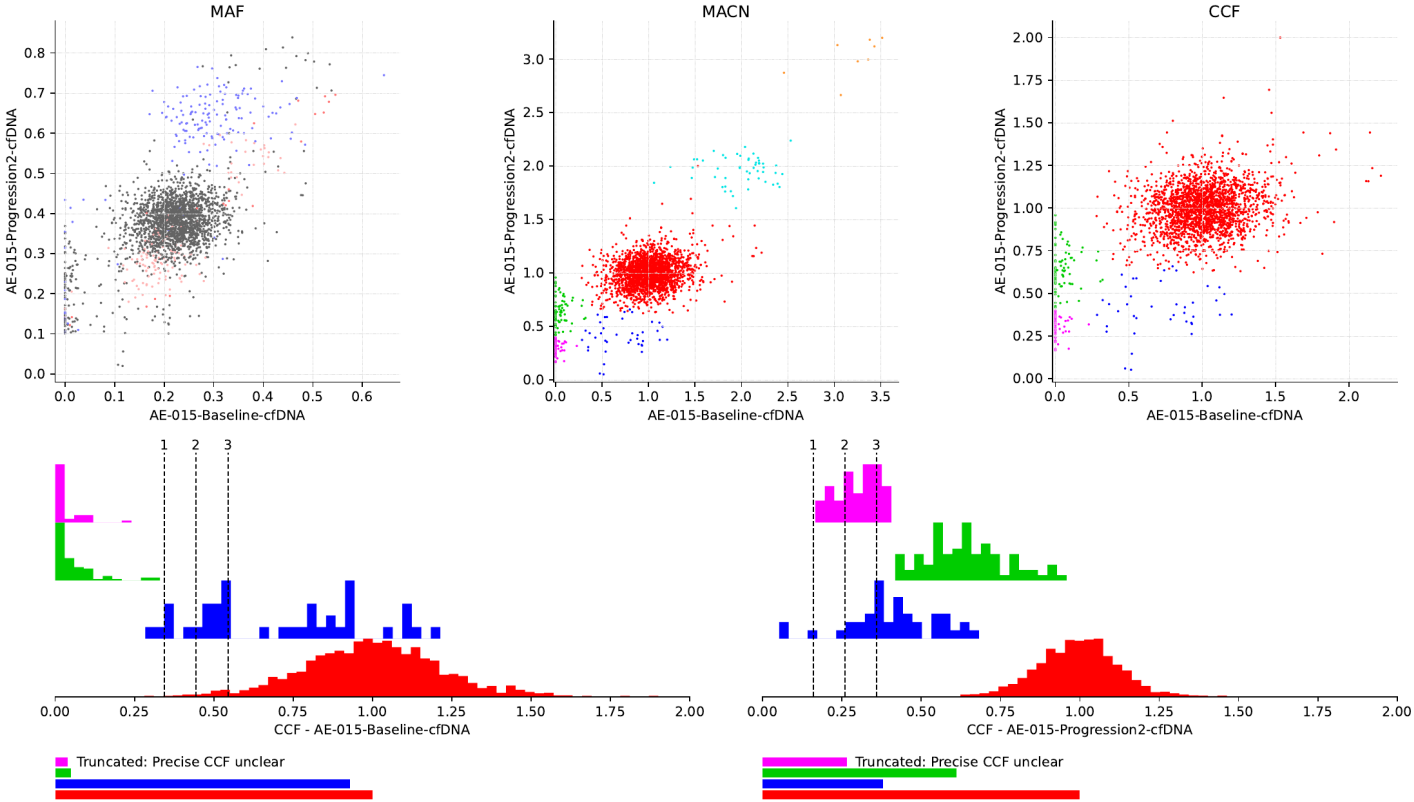


AE-015-Progression2-cfDNA

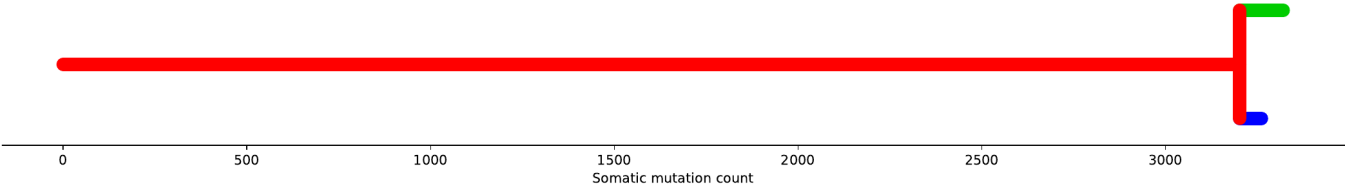
Cancer fraction: 77%, Diploid level: 0.06



AE-015 mutation clustering

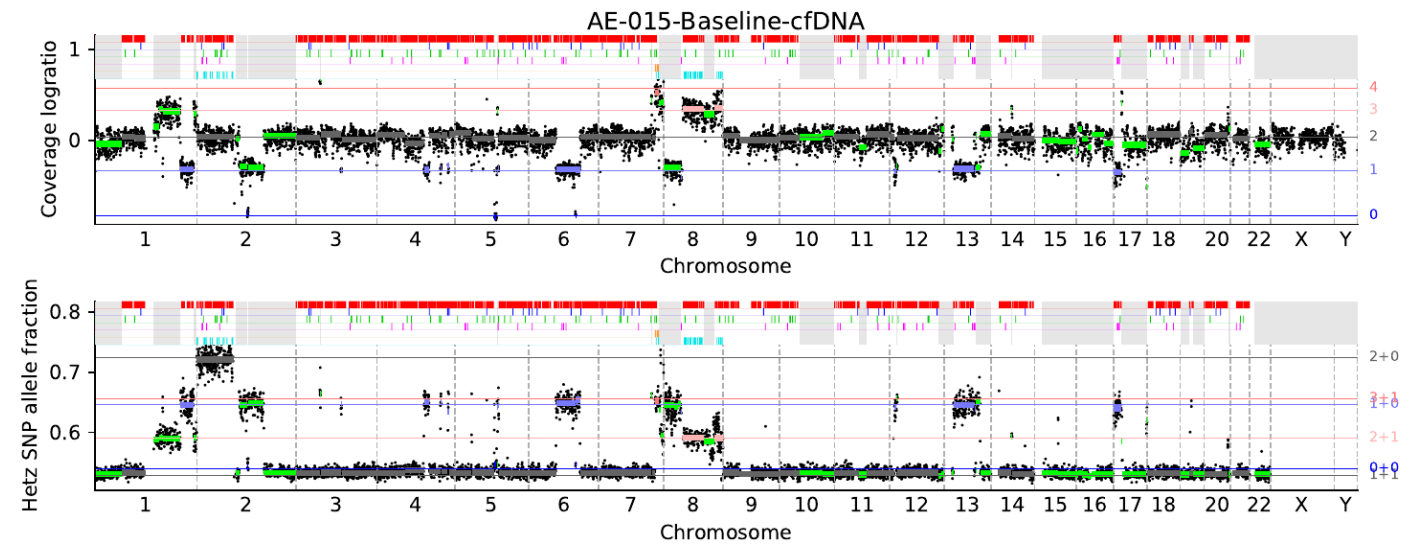


Phylogenetic tree

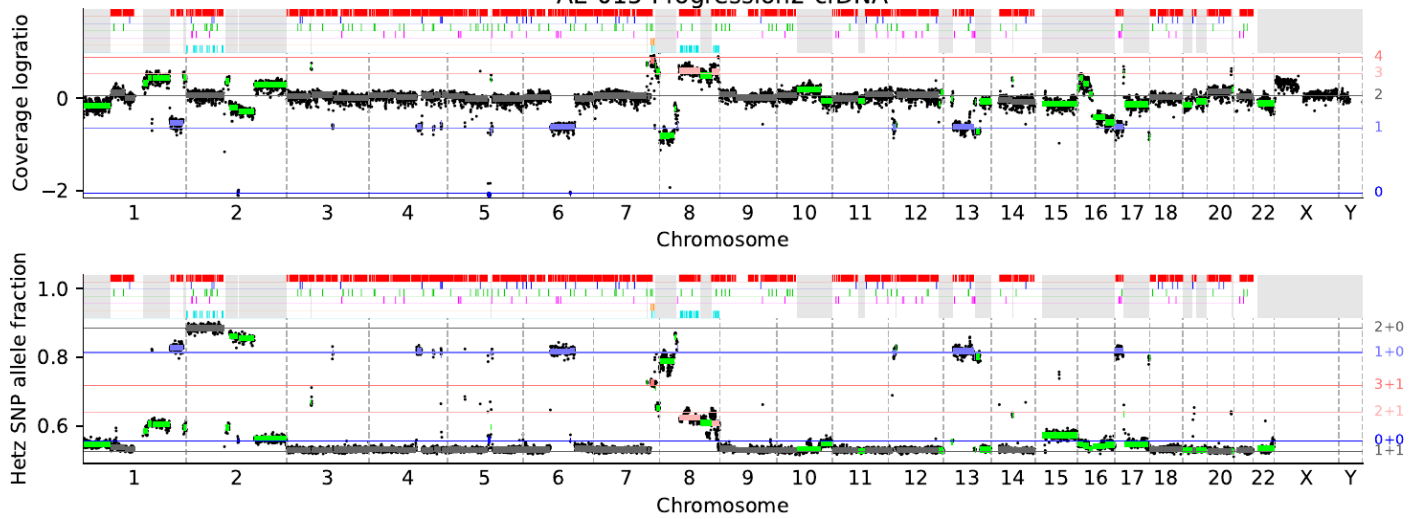


AE-015 mutation cluster positions

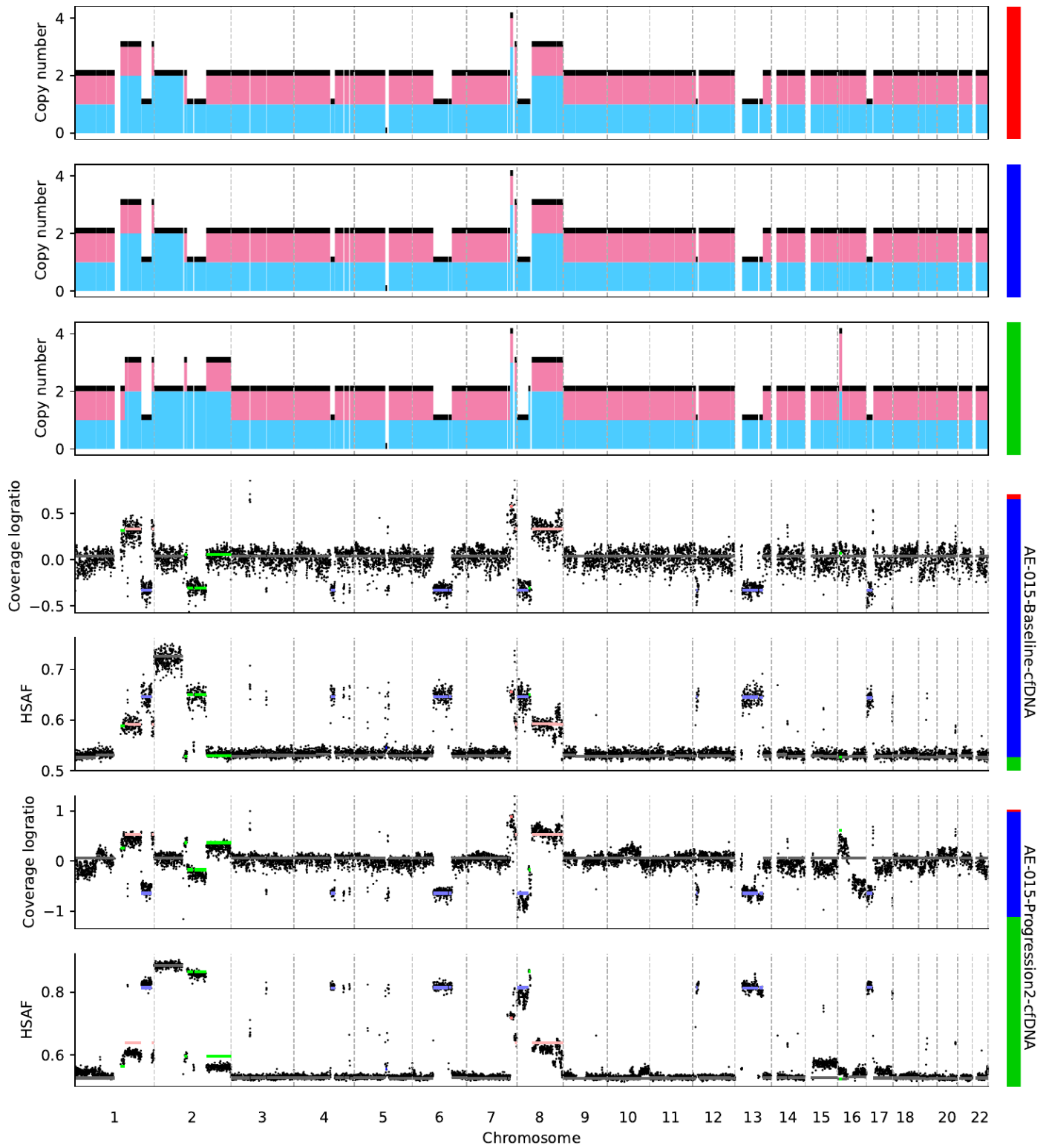
AE-015-Baseline-cfDNA



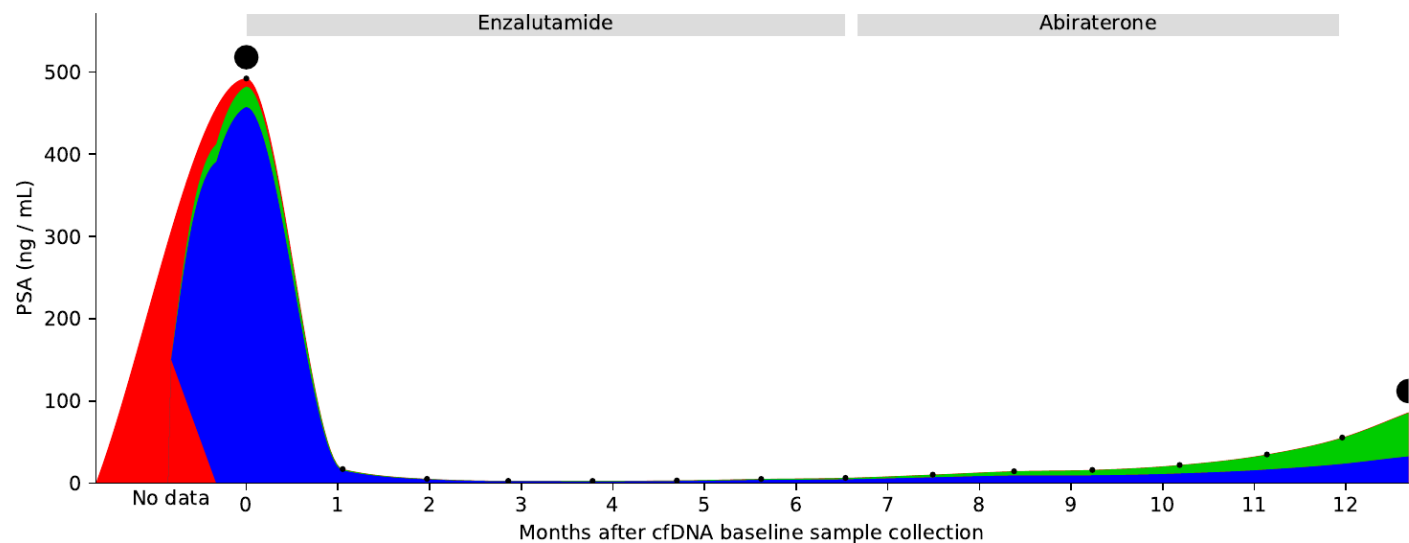
AE-015-Progression2-cfDNA



AE-015 copy number deconvolution



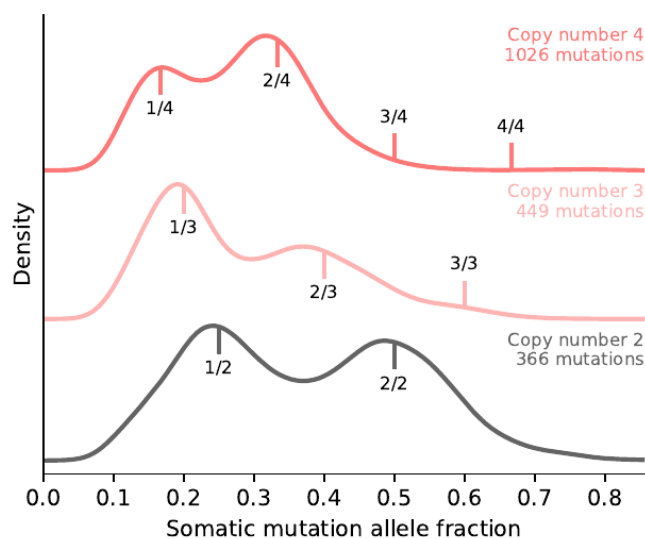
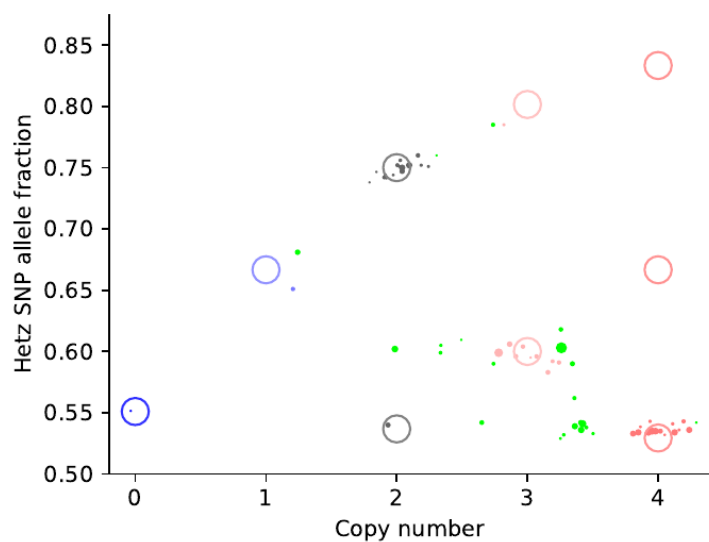
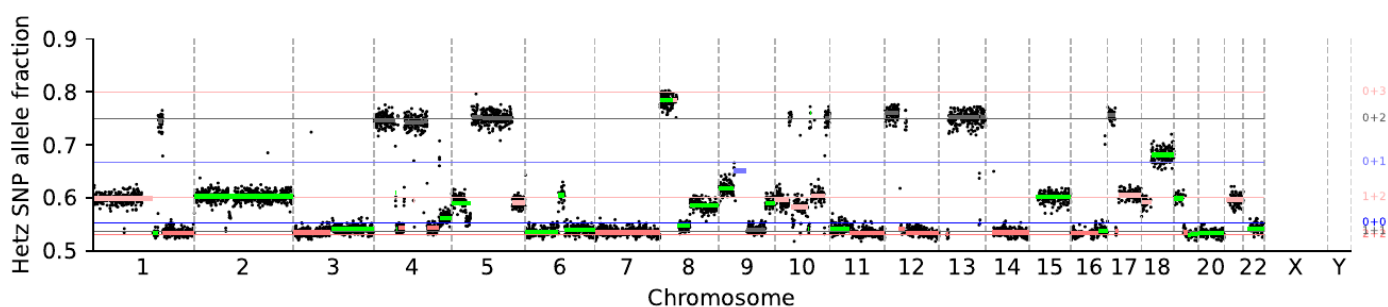
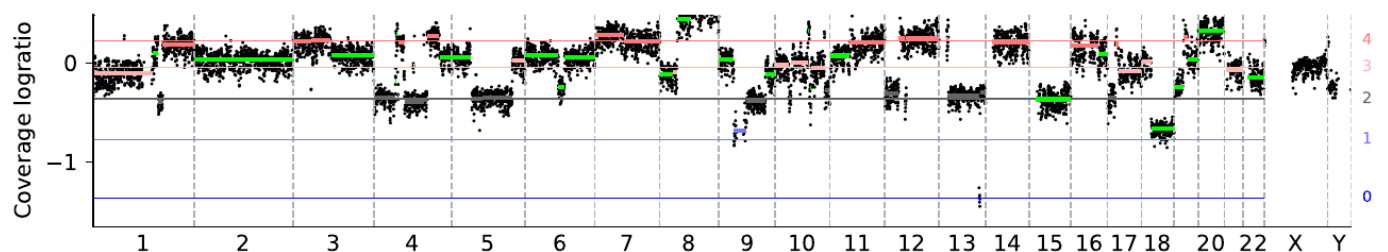
AE-015 clonal expansion timeline



Patient AE-018

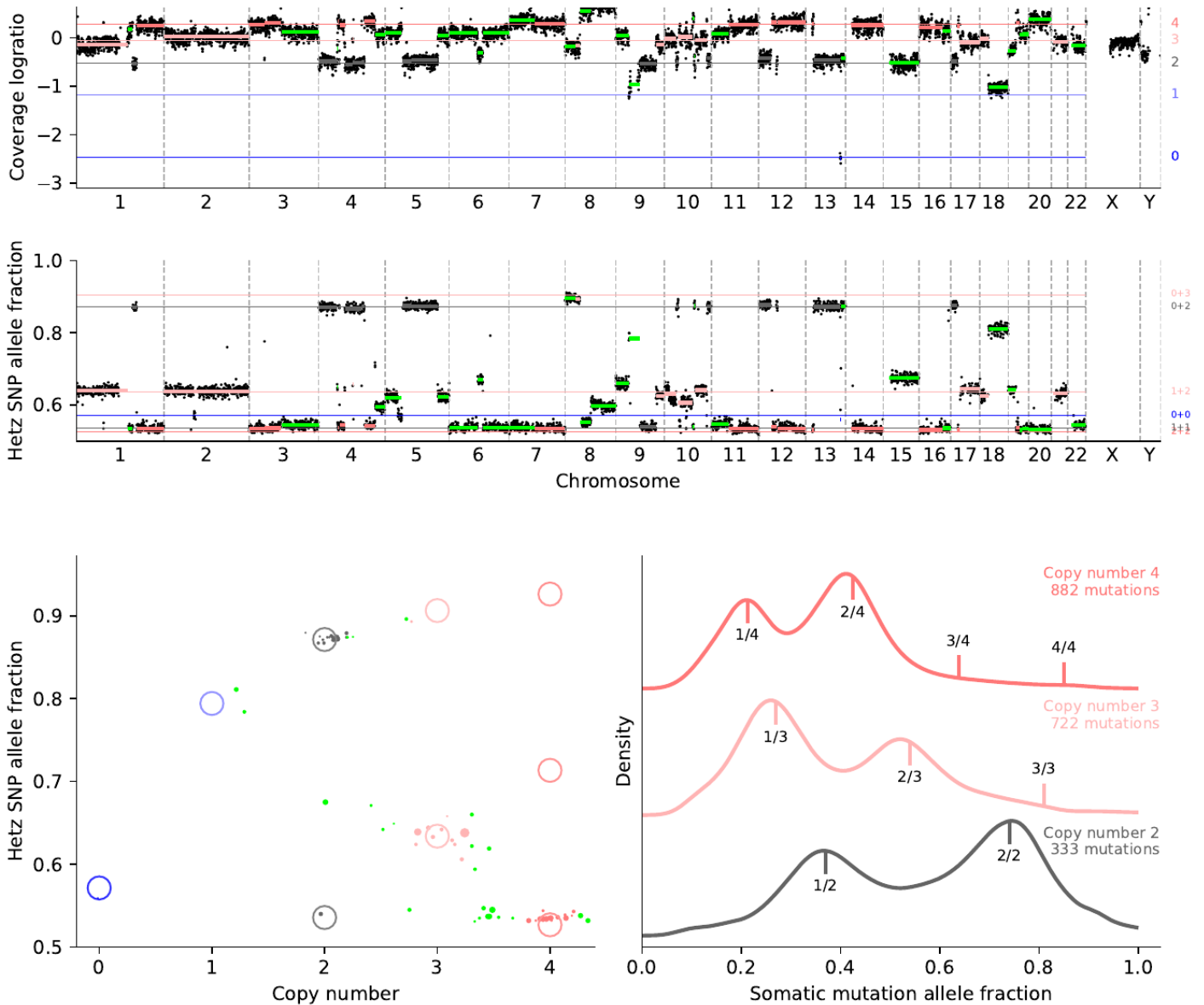
AE-018-Progression-cfDNA

Cancer fraction: 50%, Diploid level: -0.36



AE-018-Progression2-cfDNA

Cancer fraction: 74%, Diploid level: -0.52



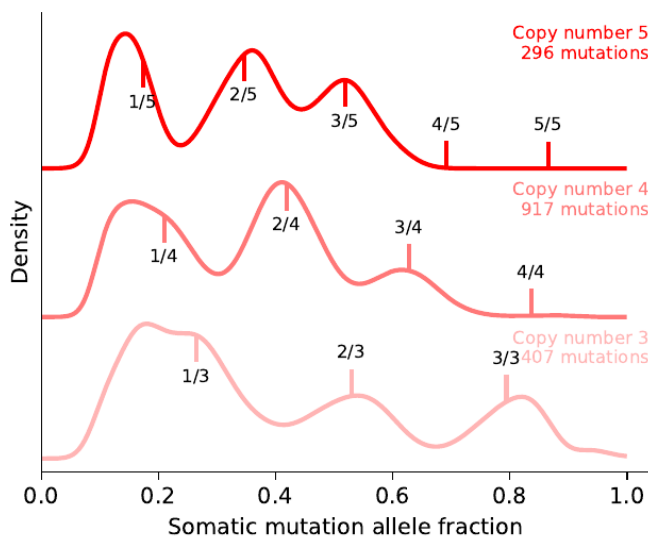
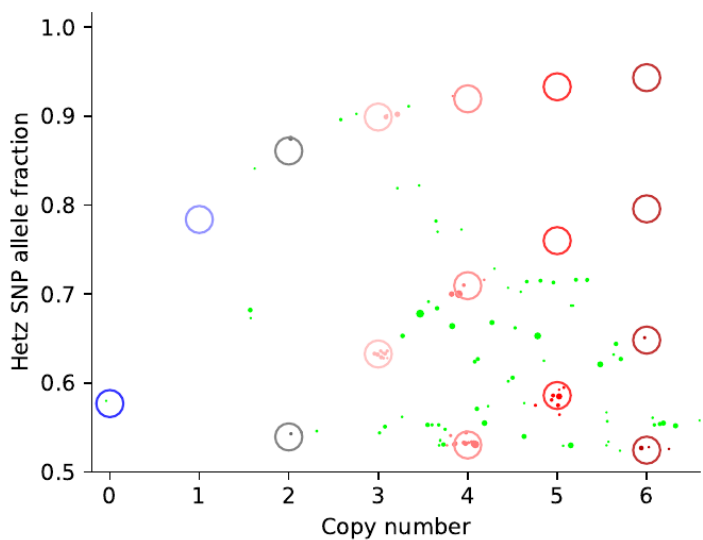
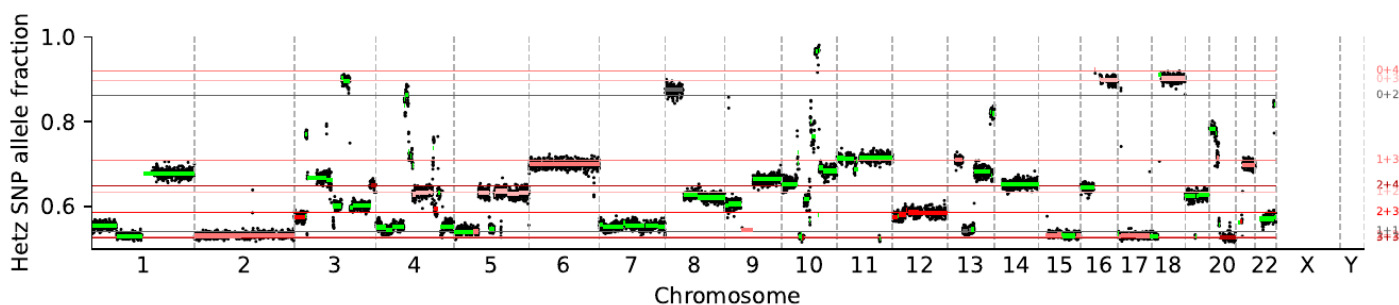
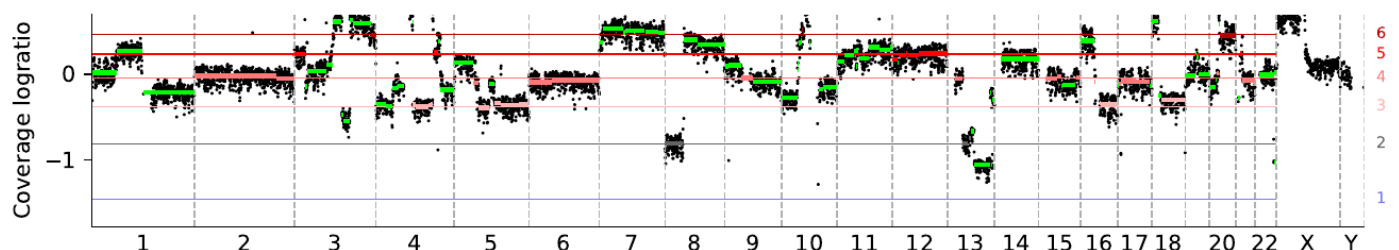
AE-018 mutation clustering

Patient AE-018 had samples that were admixtures of cancer cell subpopulations with different whole genome duplication statuses (i.e. subclonal heterogeneity with respect to presence of whole genome duplication). Please see Supplementary Information section "Detection of subclonal whole genome duplication events" for a detailed subclonal reconstruction.

Patient AE-030

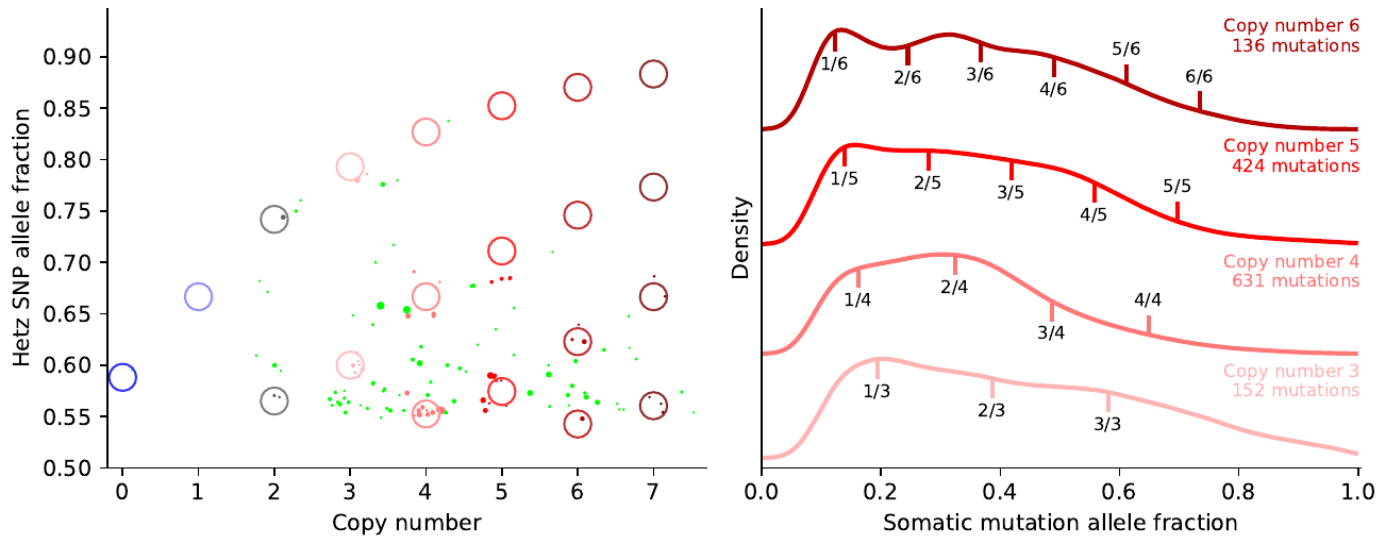
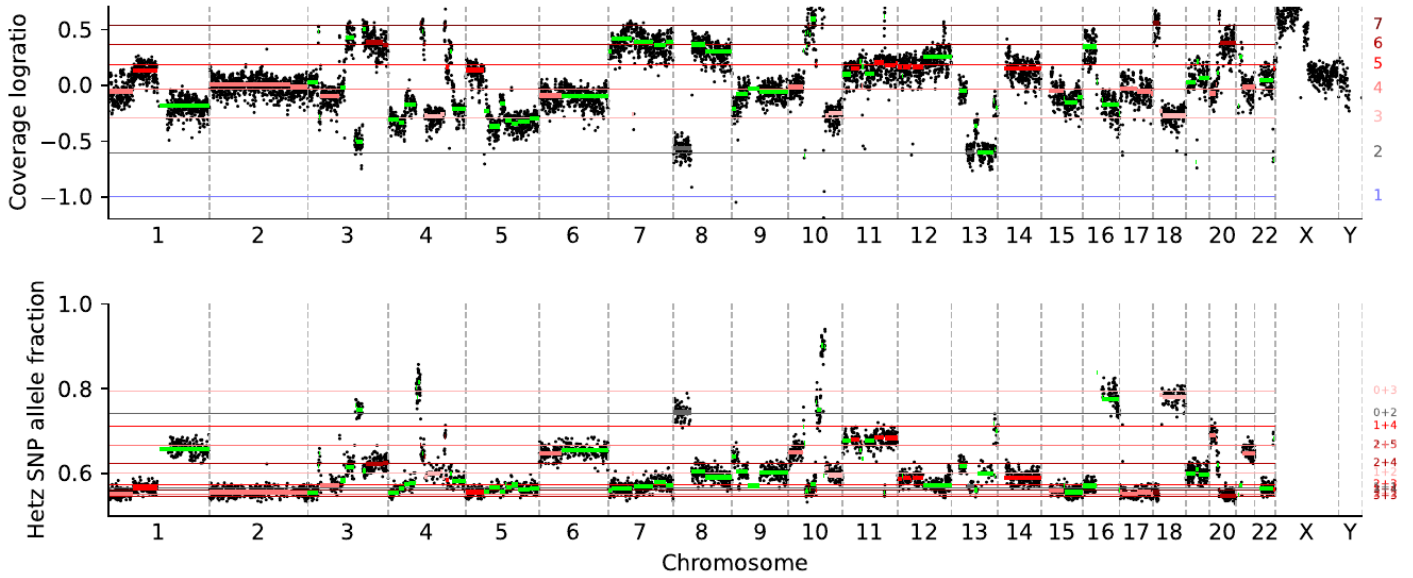
AE-030-Baseline-cfDNA

Cancer fraction: 72%, Diploid level: -0.82



AE-030-Progression-cfDNA

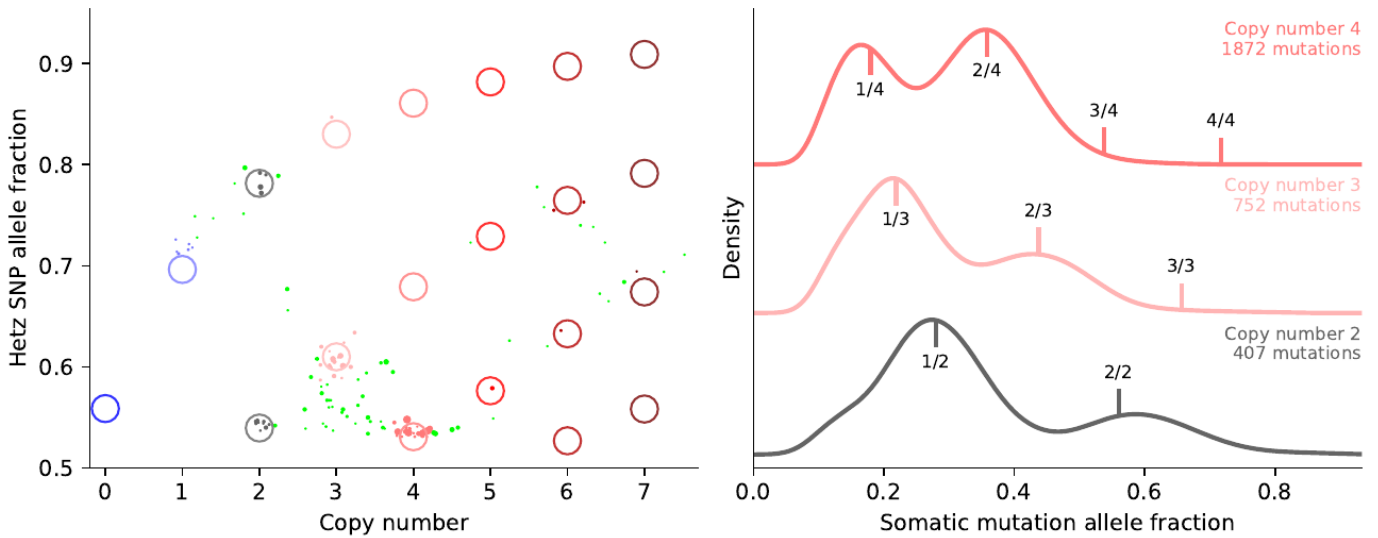
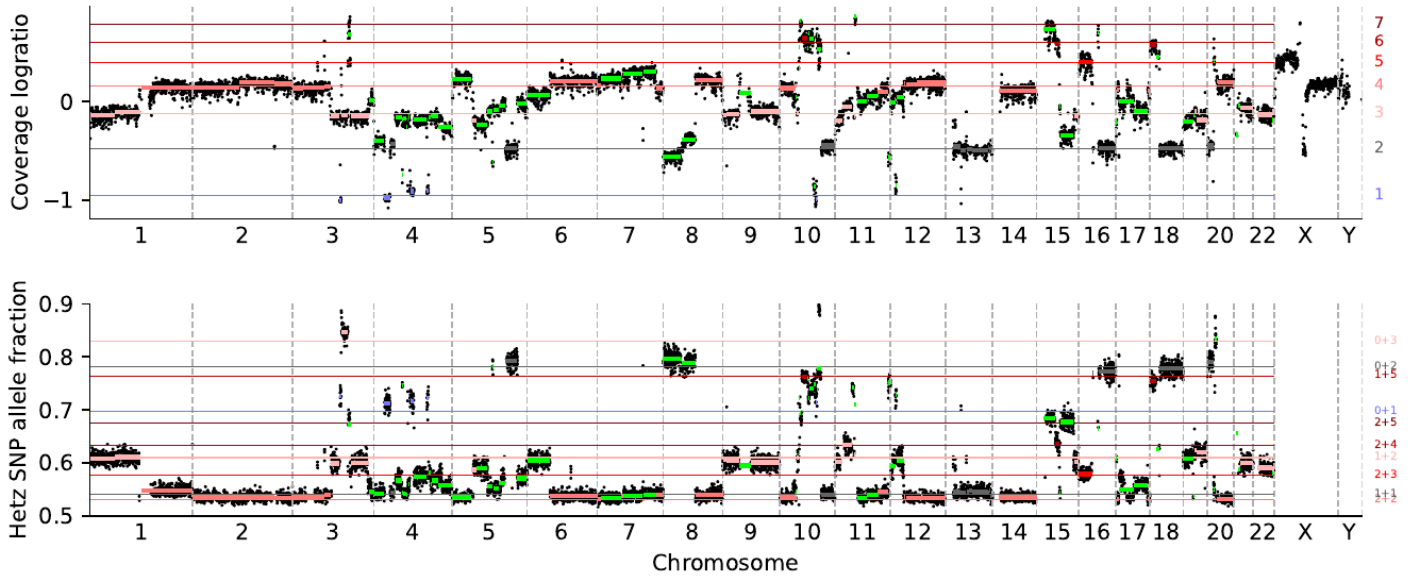
Cancer fraction: 48%, Diploid level: -0.60



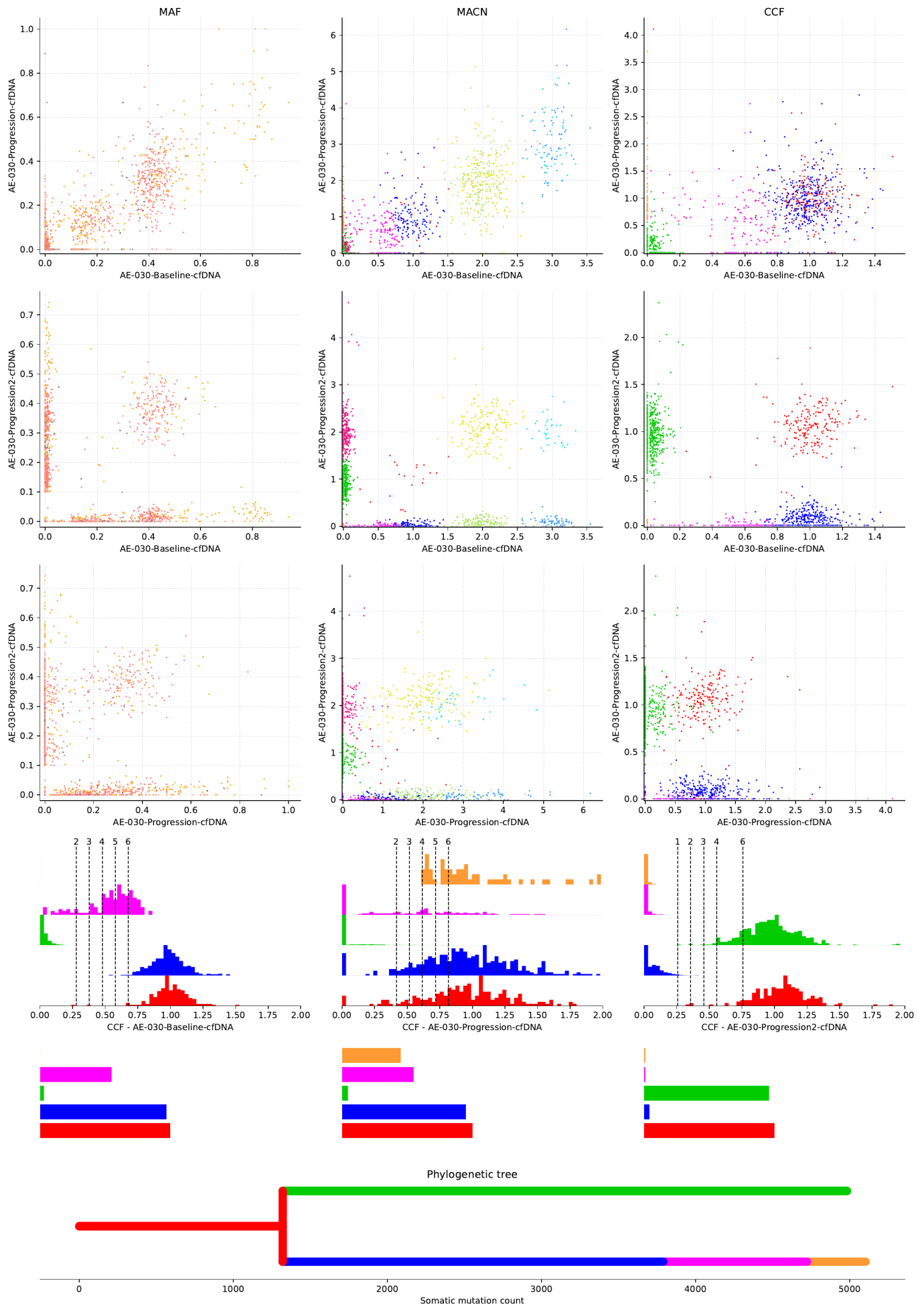
PLEASE NOTE: Low median sequencing depth (44x) of this sample has resulted in higher-than-usual noise levels.

AE-030-Progression2-cfDNA

Cancer fraction: 56%, Diploid level: -0.48

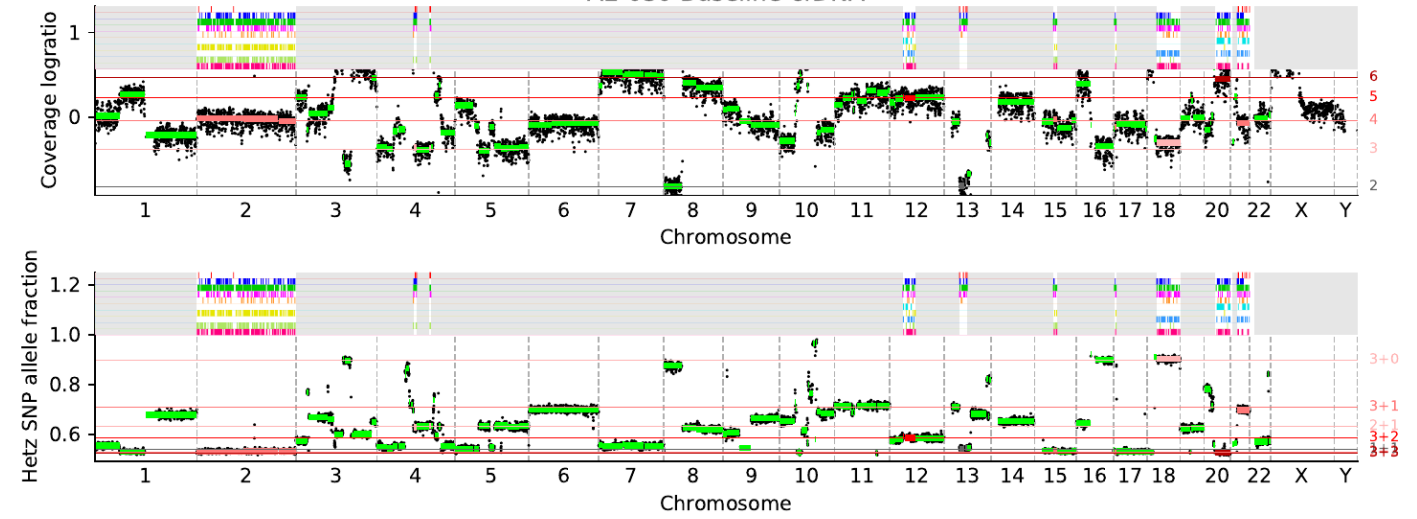


AE-030 mutation clustering

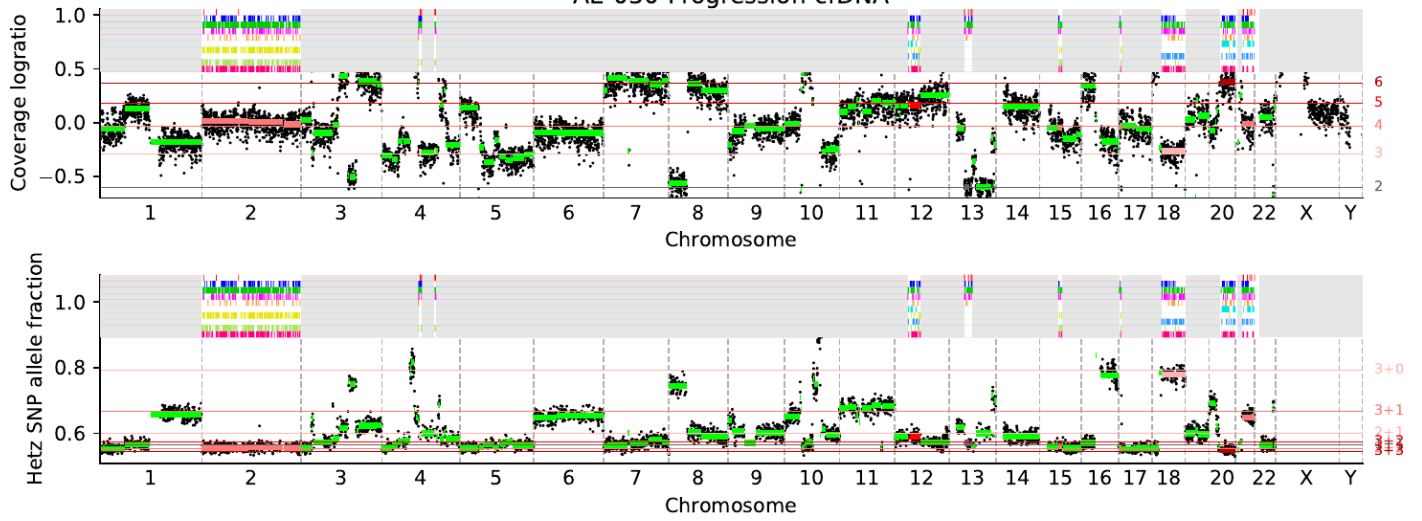


AE-030 mutation cluster positions

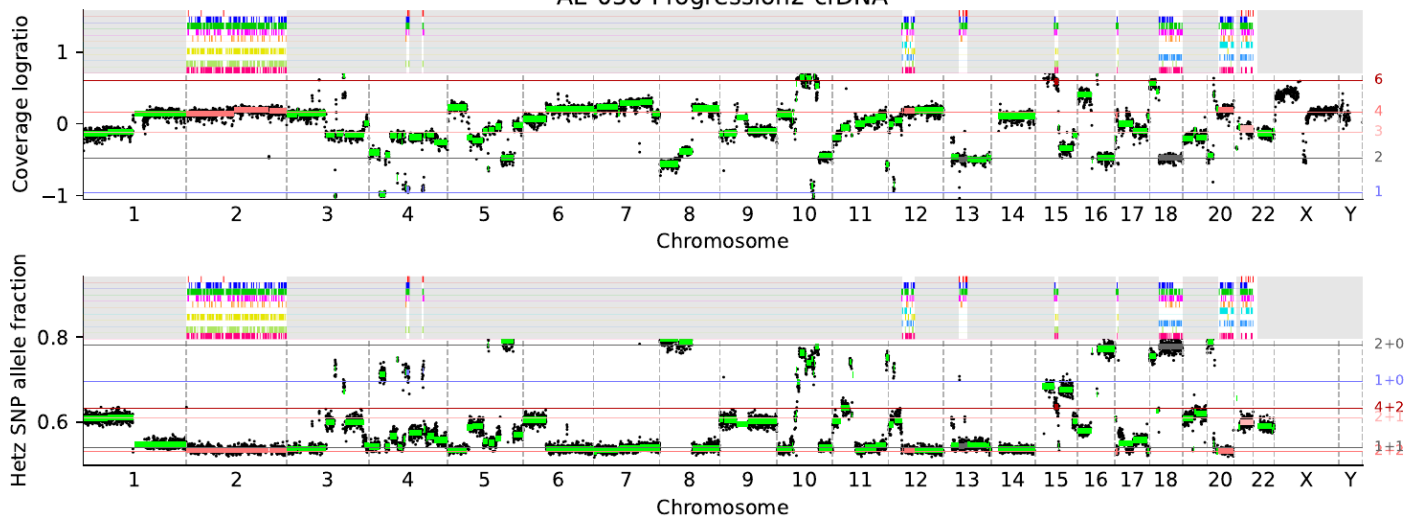
AE-030-Baseline-cfDNA



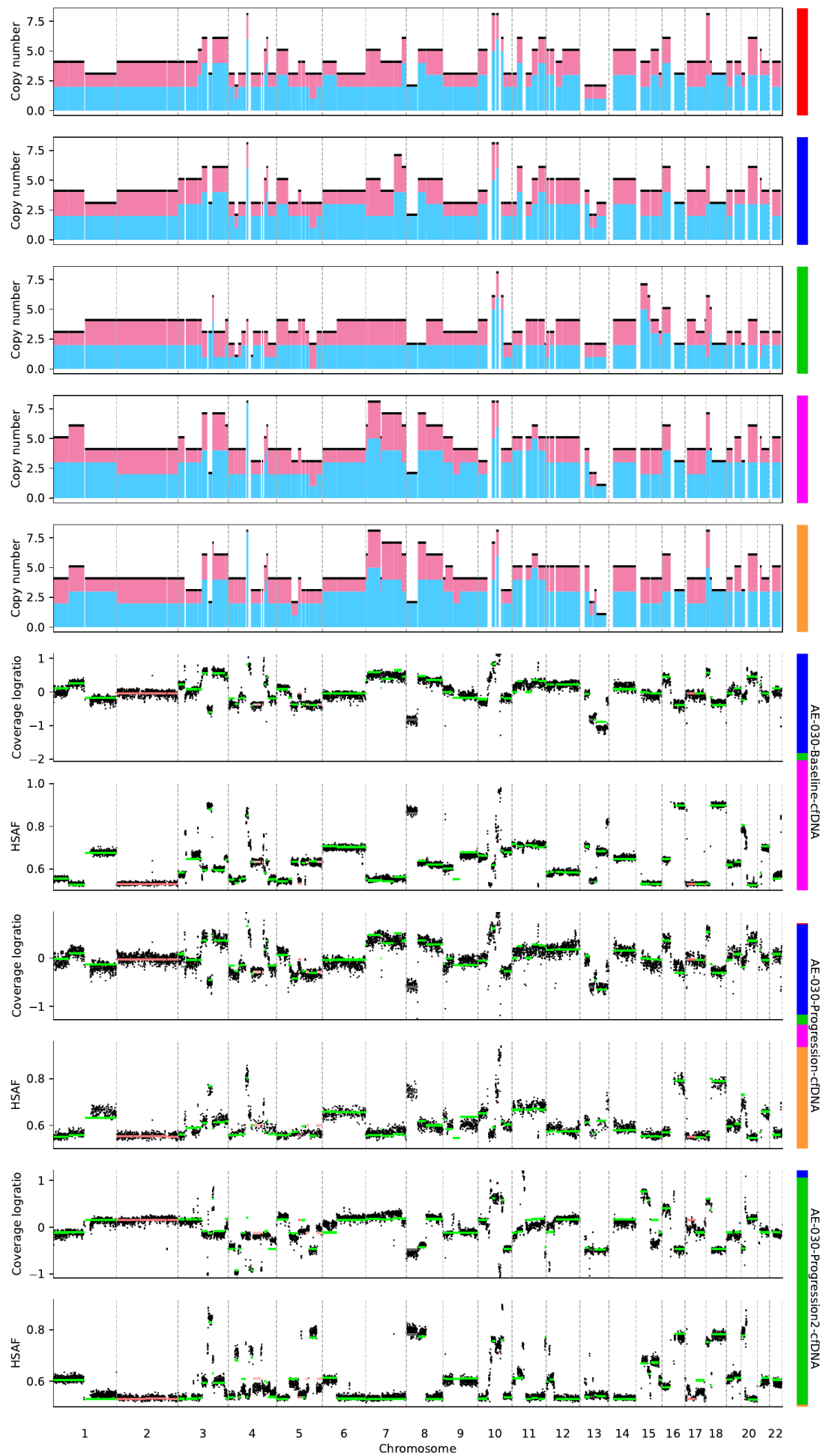
AE-030-Progression-cfDNA



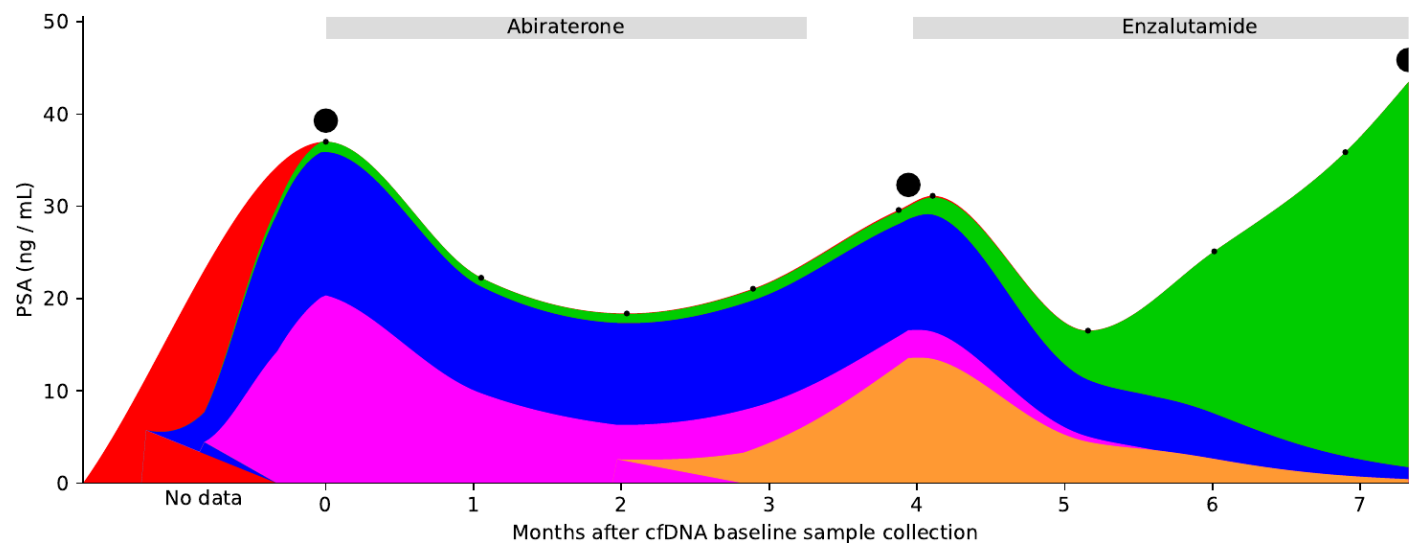
AE-030-Progression2-cfDNA



AE-030 copy number deconvolution



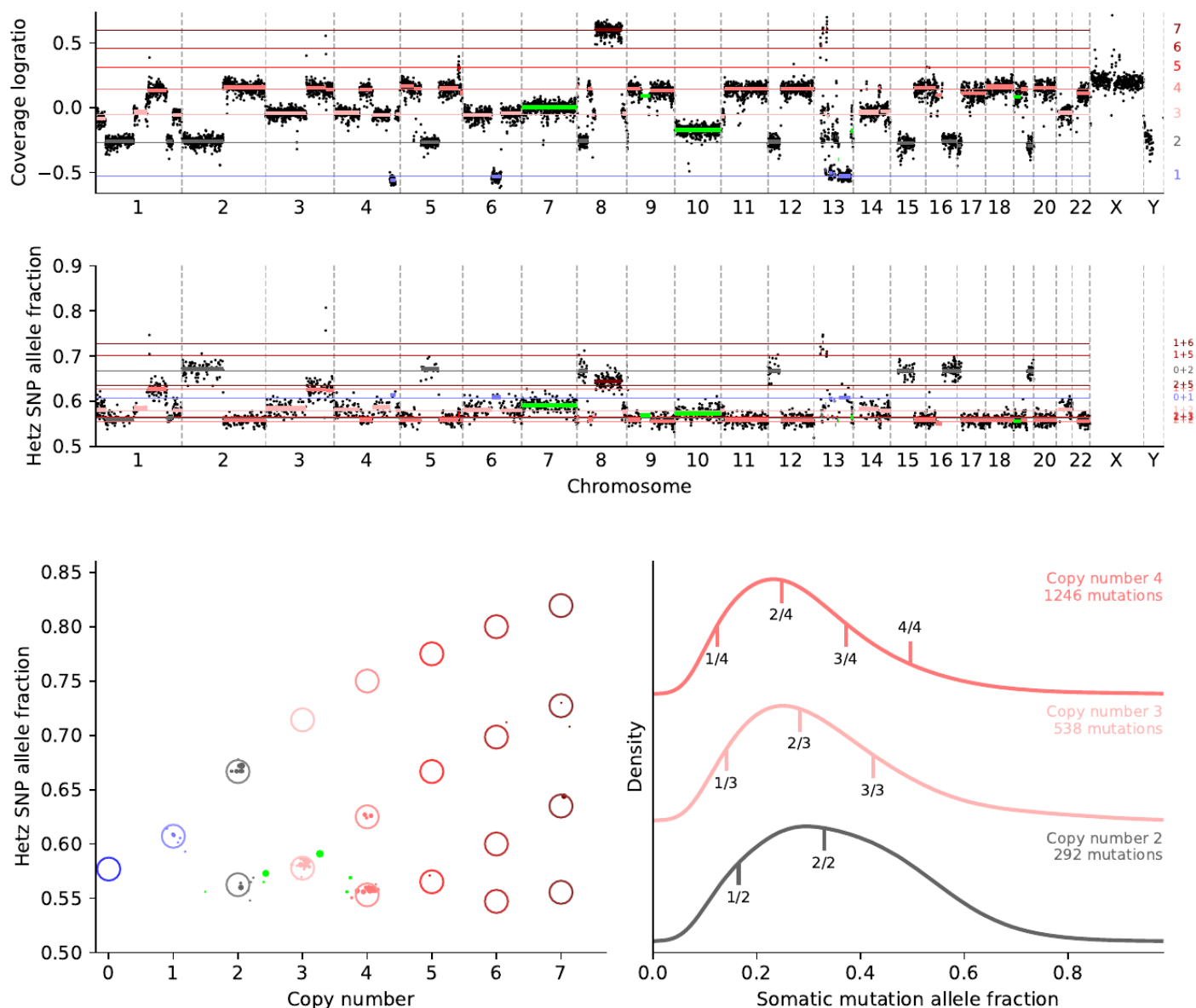
AE-030 clonal expansion timeline



Patient AE-035

AE-035-Baseline-cfDNA

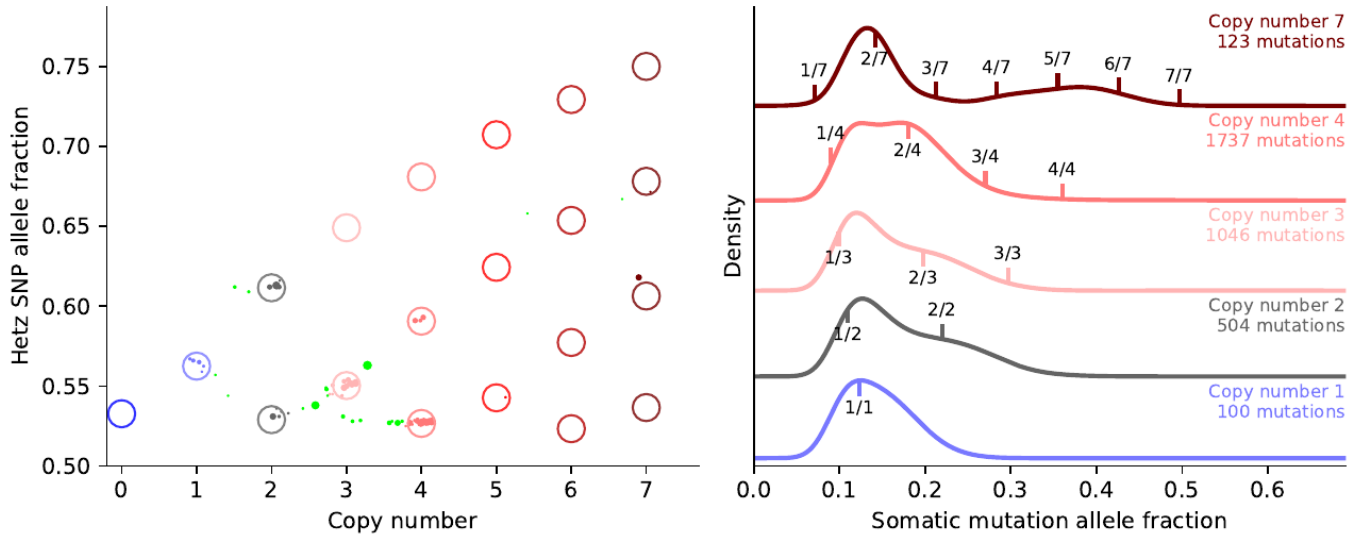
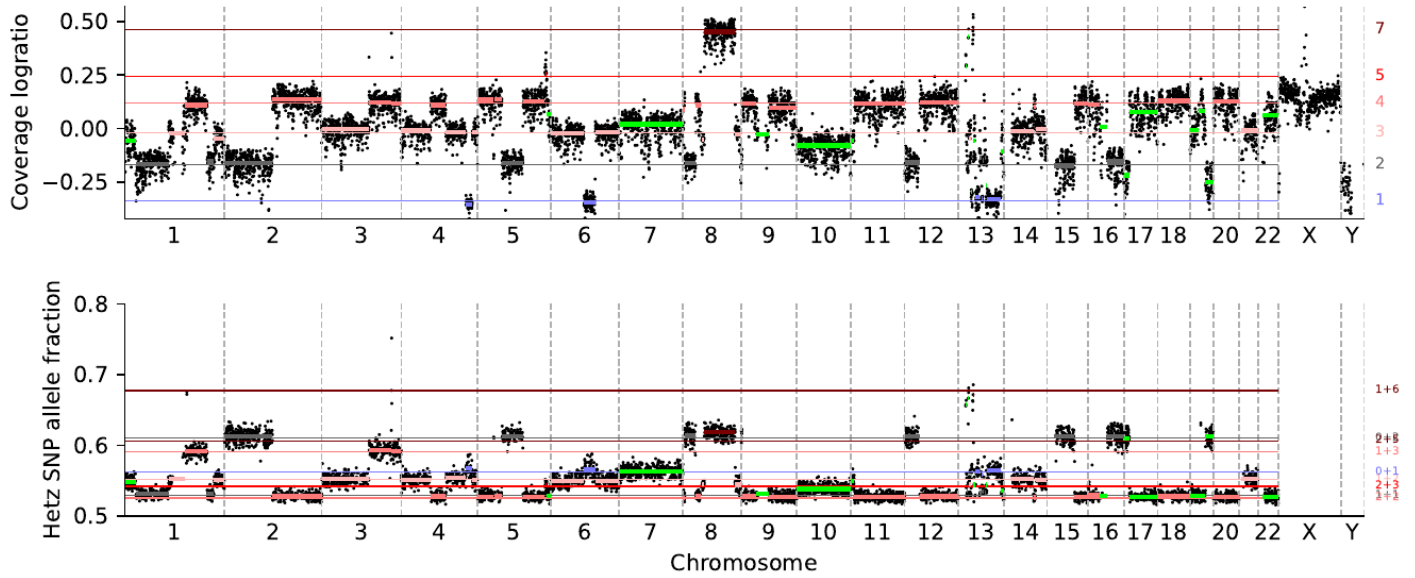
Cancer fraction: 33%, Diploid level: -0.27



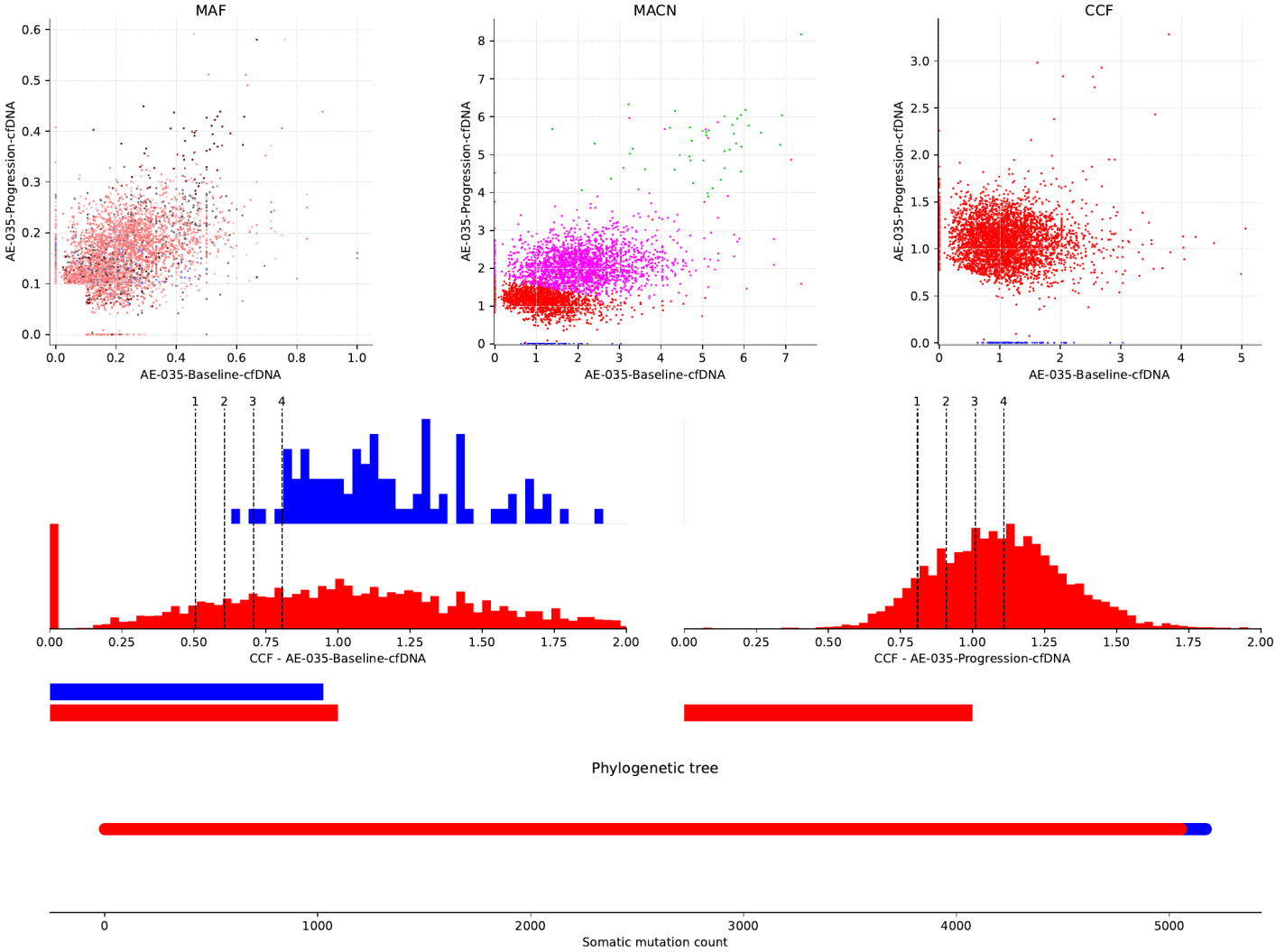
PLEASE NOTE: Low median sequencing depth (35x) of this sample has resulted in higher-than-usual noise levels.

AE-035-Progression-cfDNA

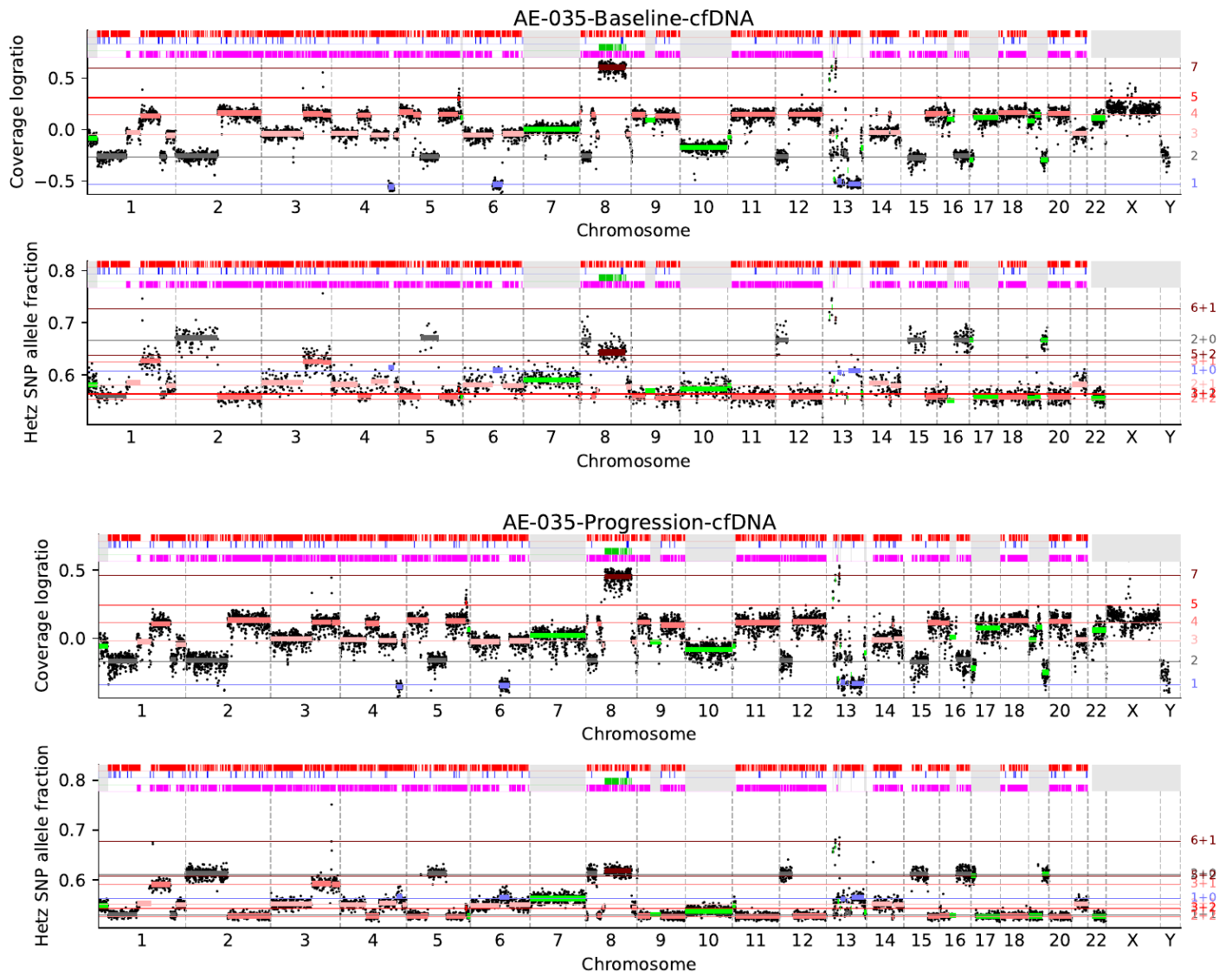
Cancer fraction: 22%, Diploid level: -0.17



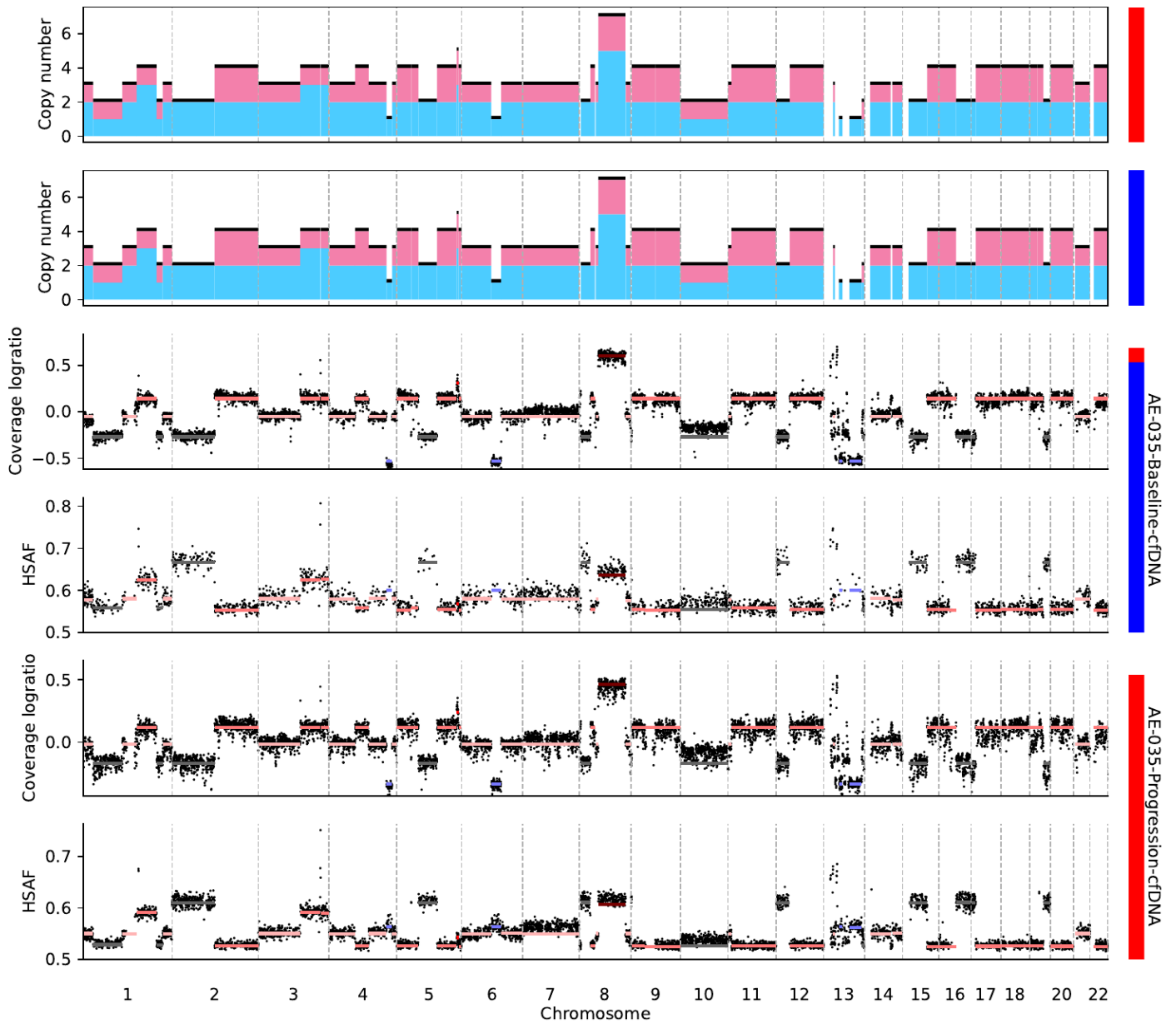
AE-035 mutation clustering



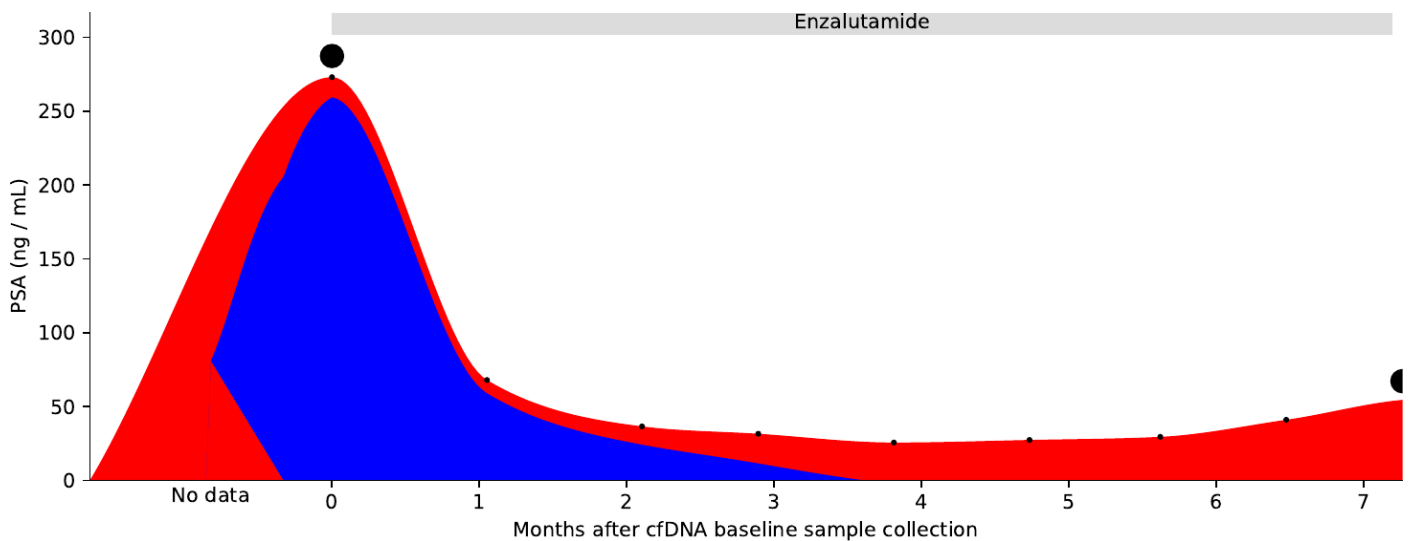
AE-035 mutation cluster positions



AE-035 copy number deconvolution



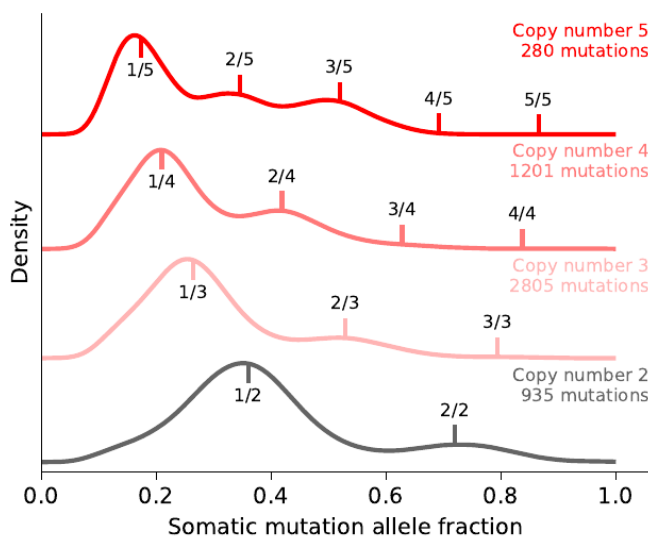
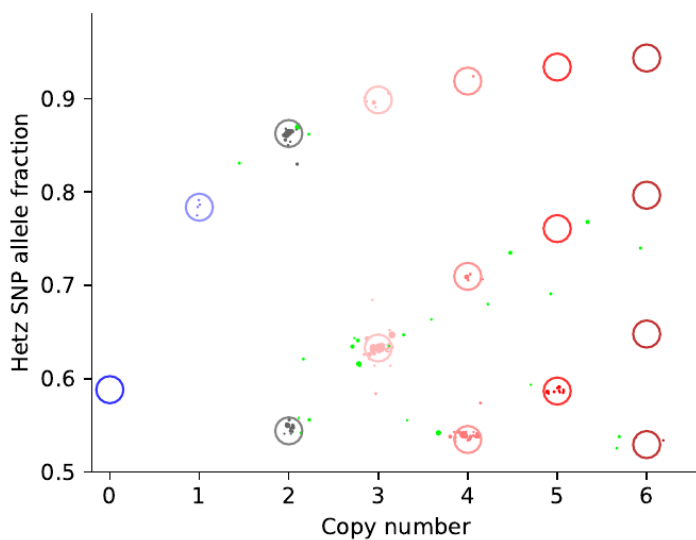
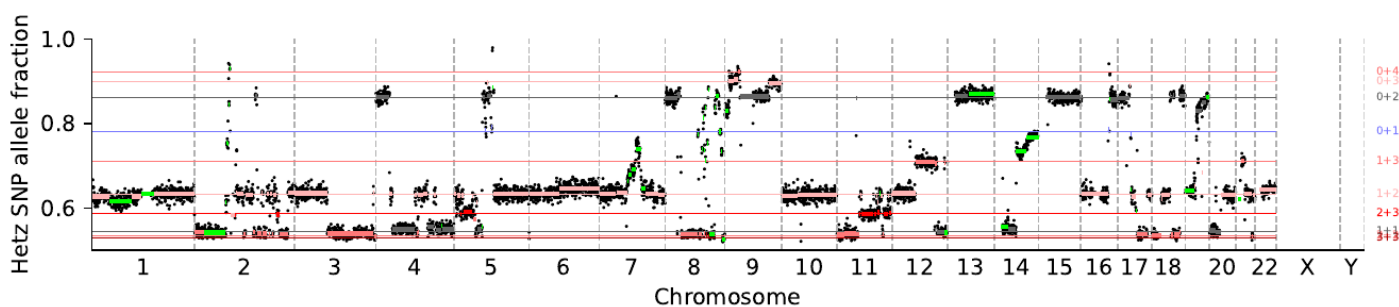
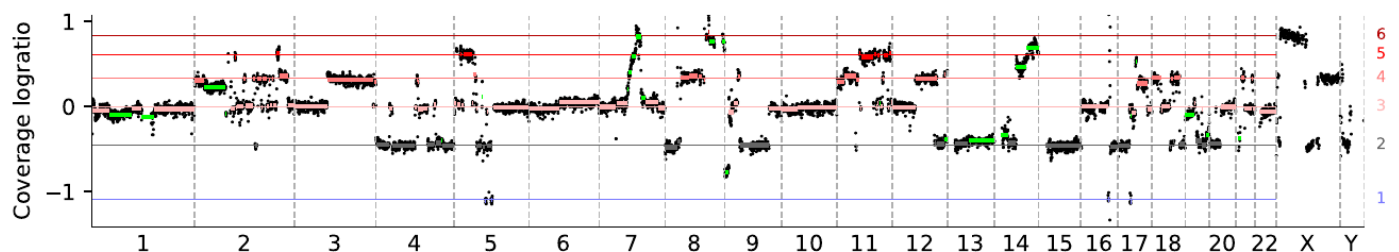
AE-035 clonal expansion timeline



Patient AE-046

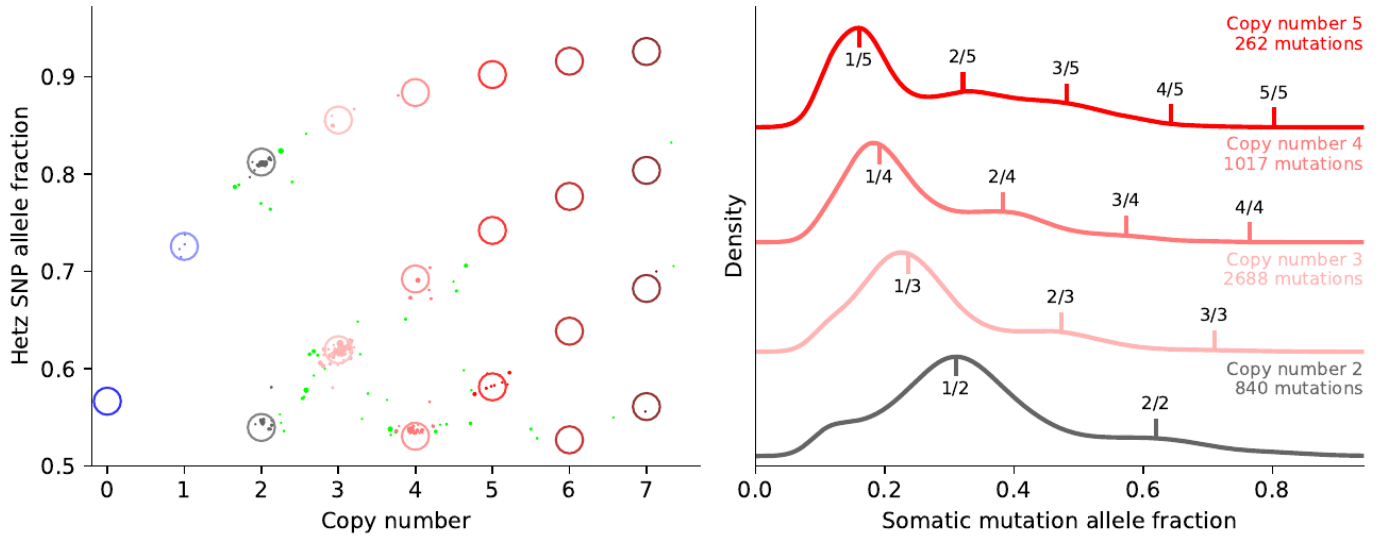
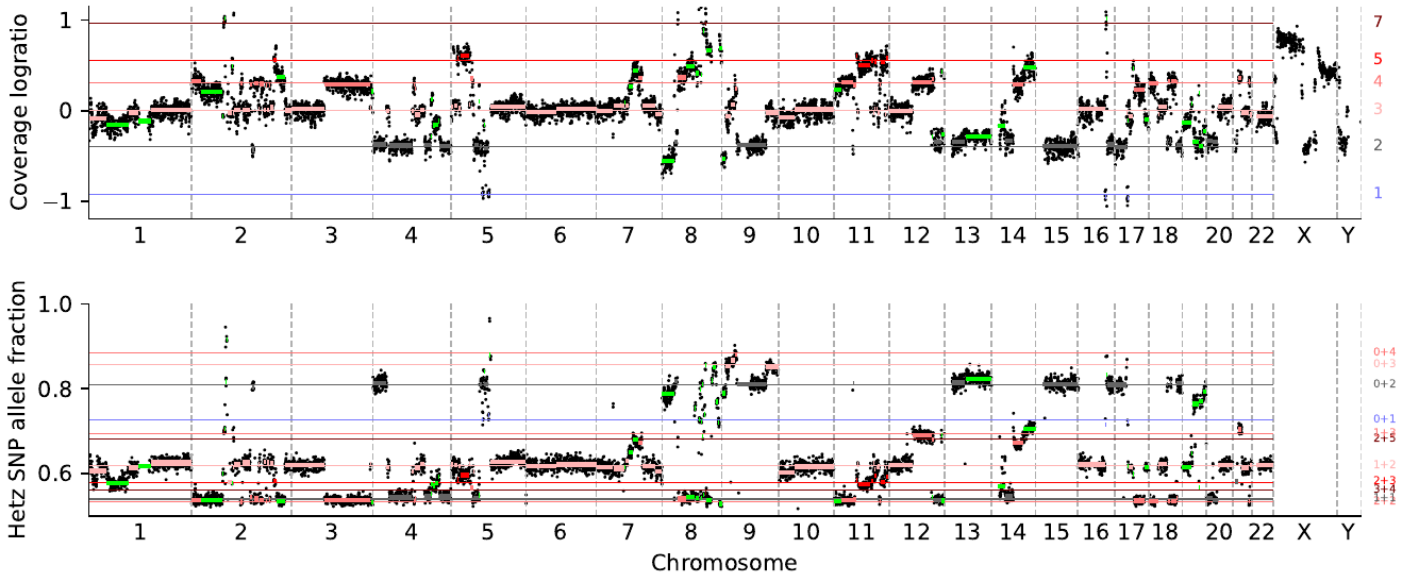
AE-046-Baseline-cfDNA

Cancer fraction: 72%, Diploid level: -0.45



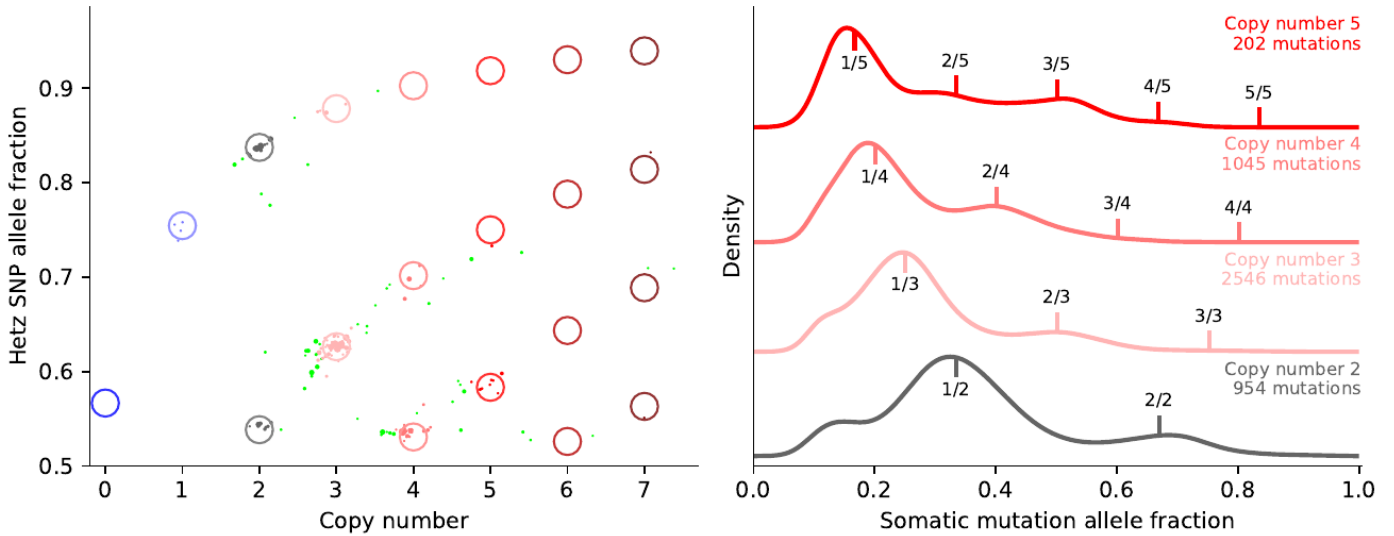
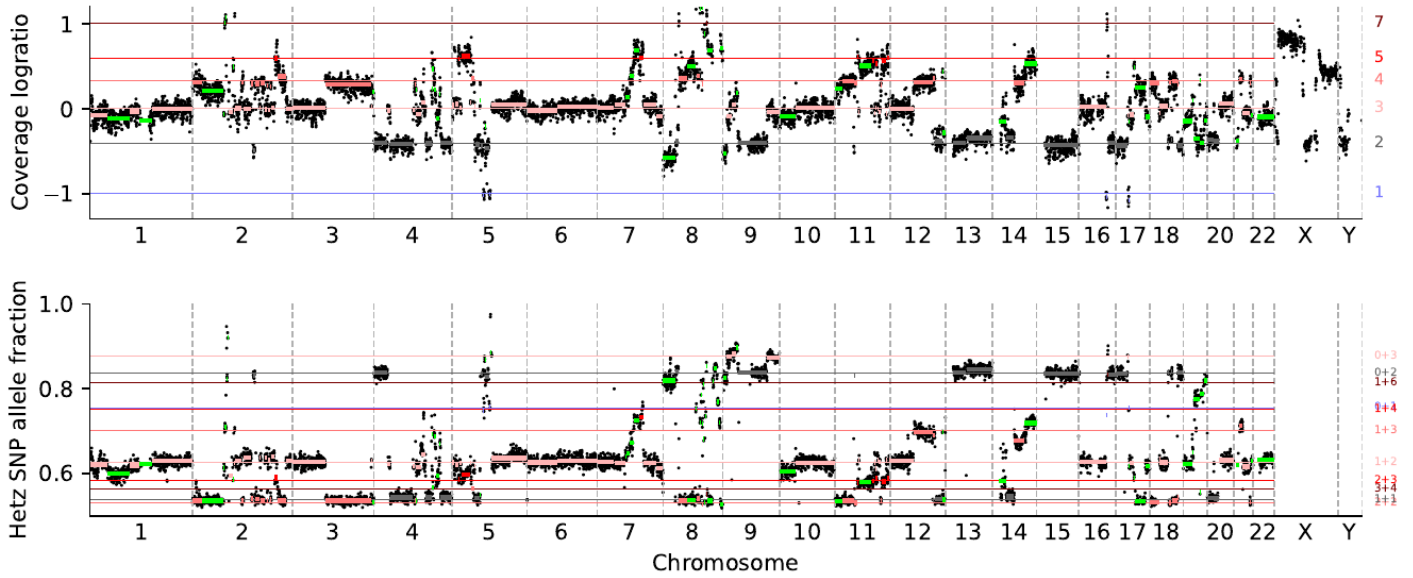
AE-046-Progression-cfDNA

Cancer fraction: 62%, Diploid level: -0.39

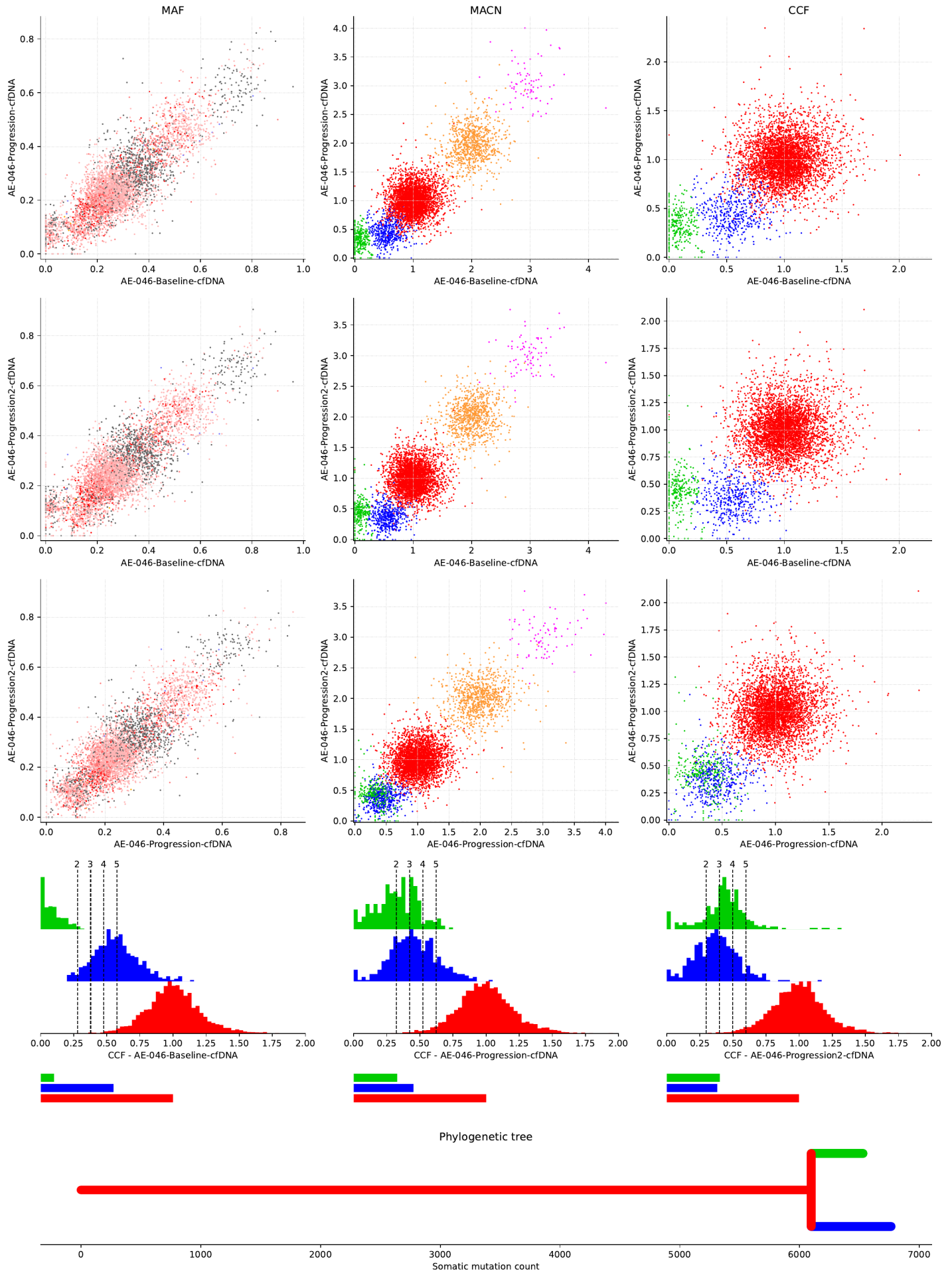


AE-046-Progression2-cfDNA

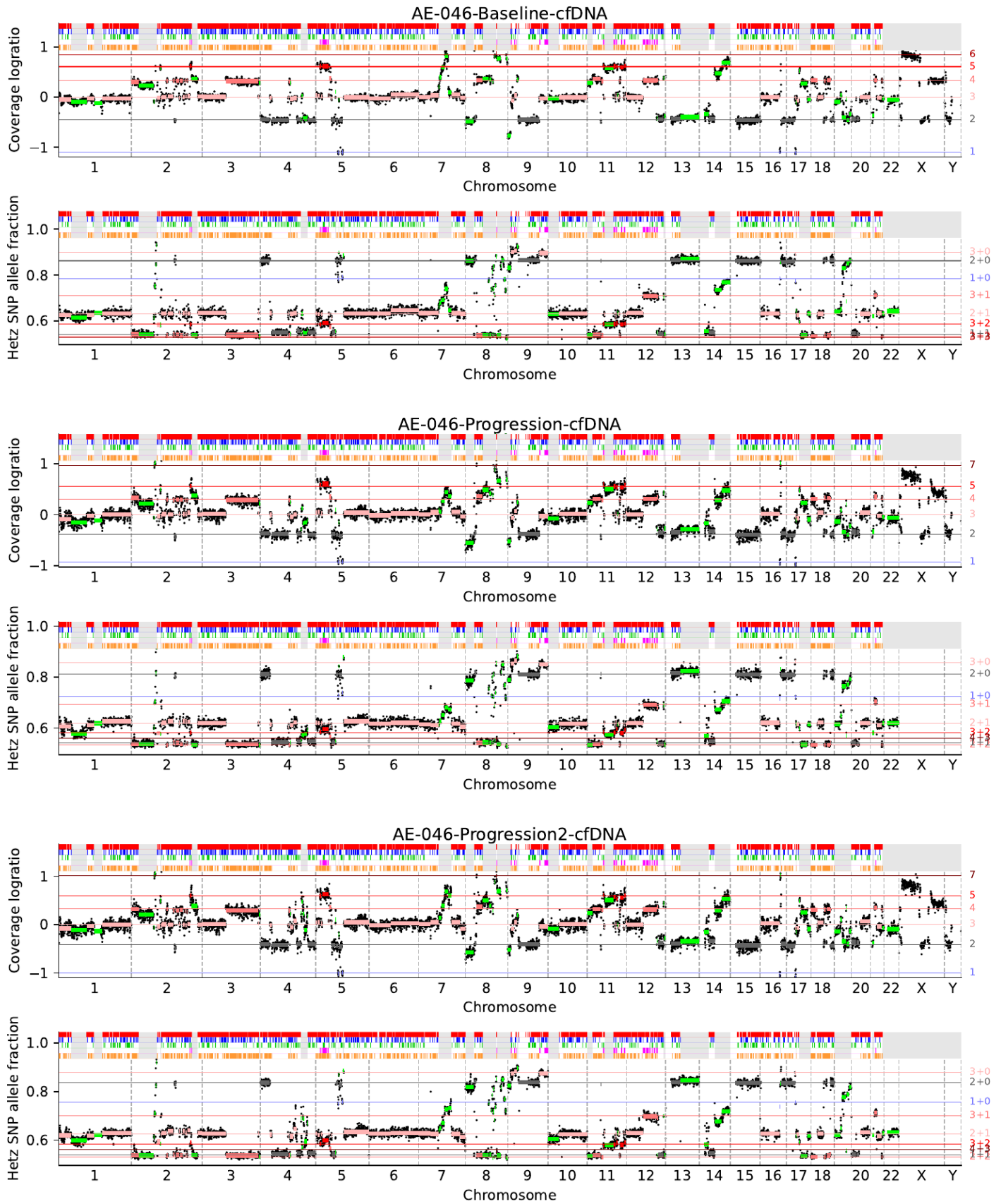
Cancer fraction: 67%, Diploid level: -0.41



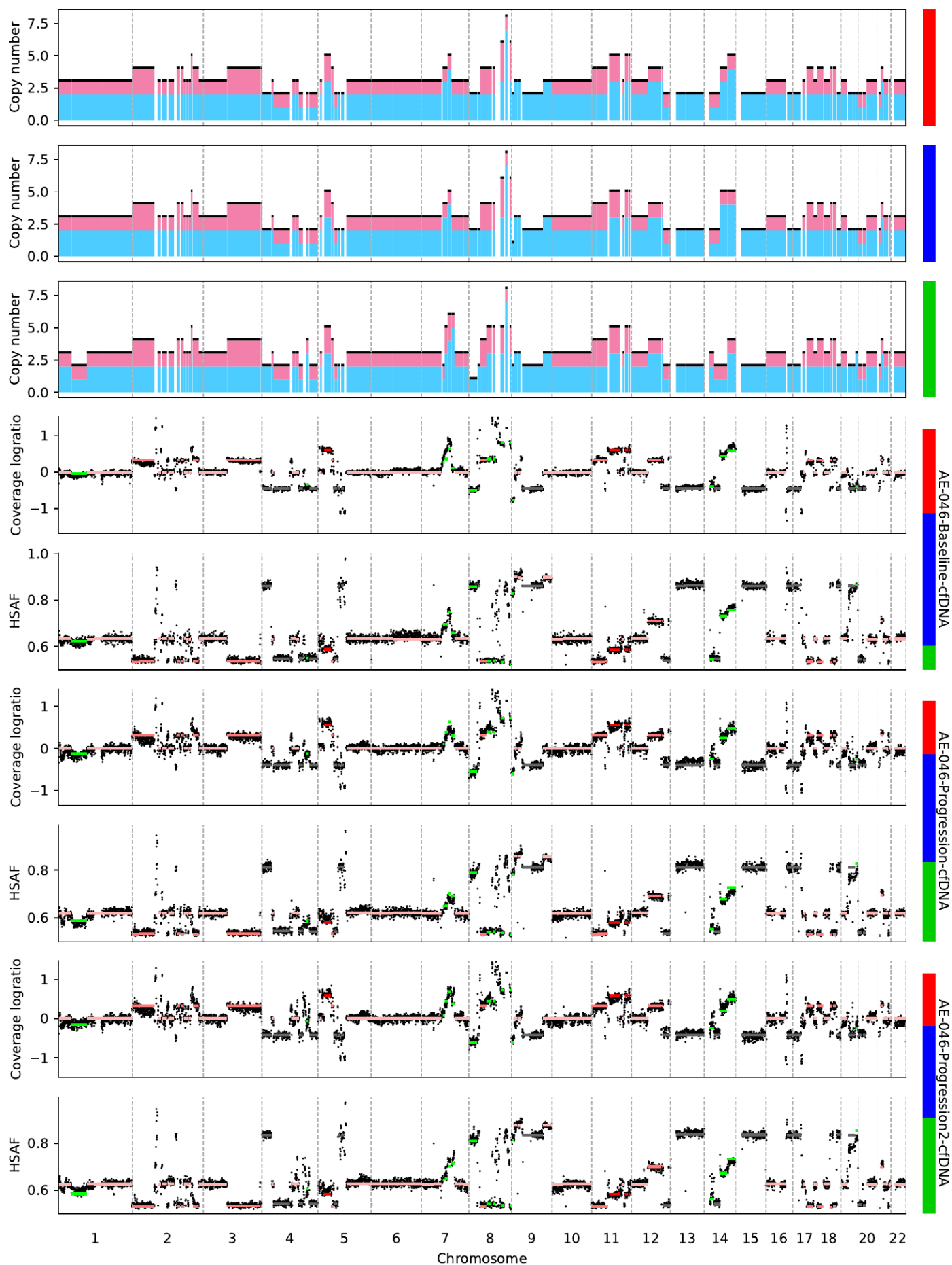
AE-046 mutation clustering



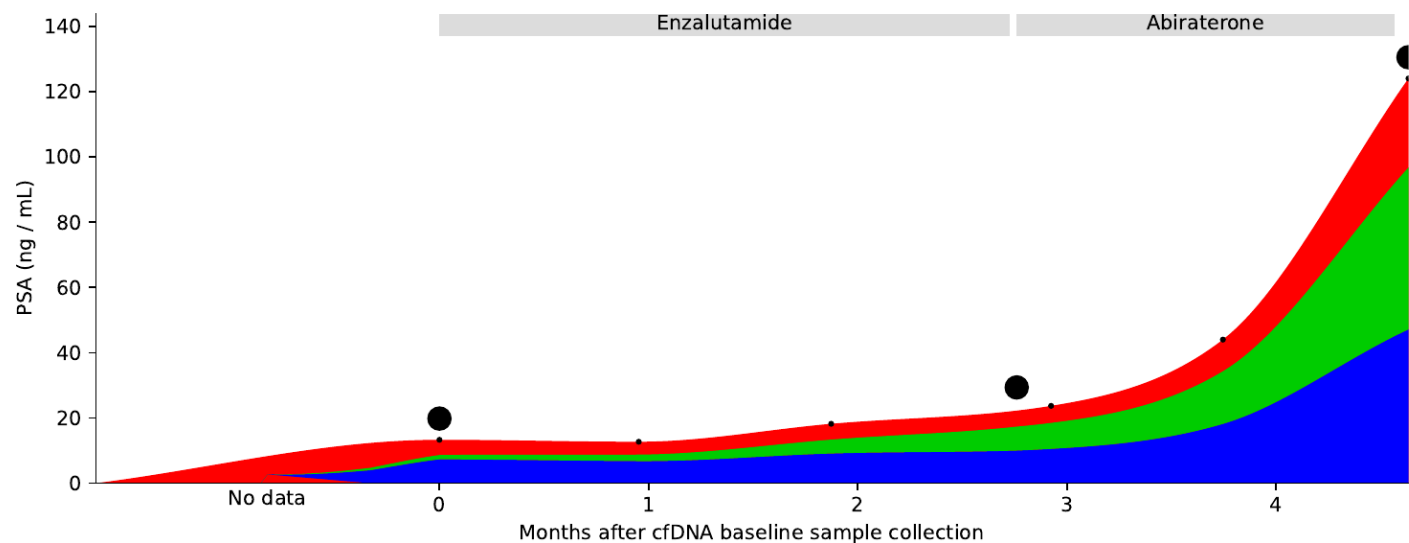
AE-046 mutation cluster positions



AE-046 copy number deconvolution



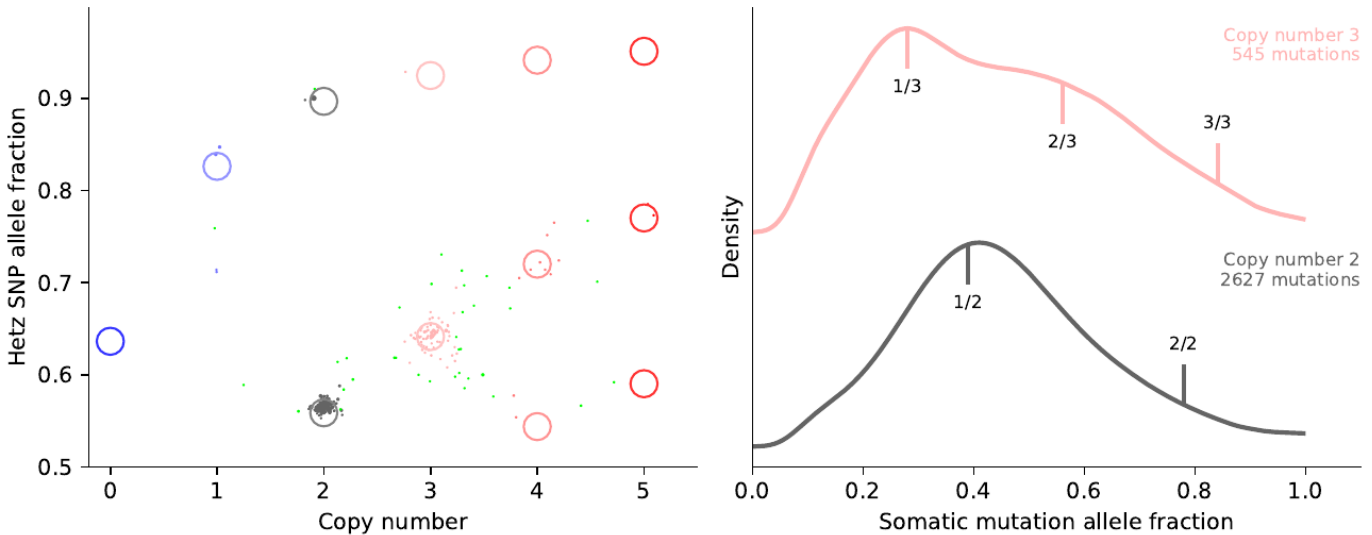
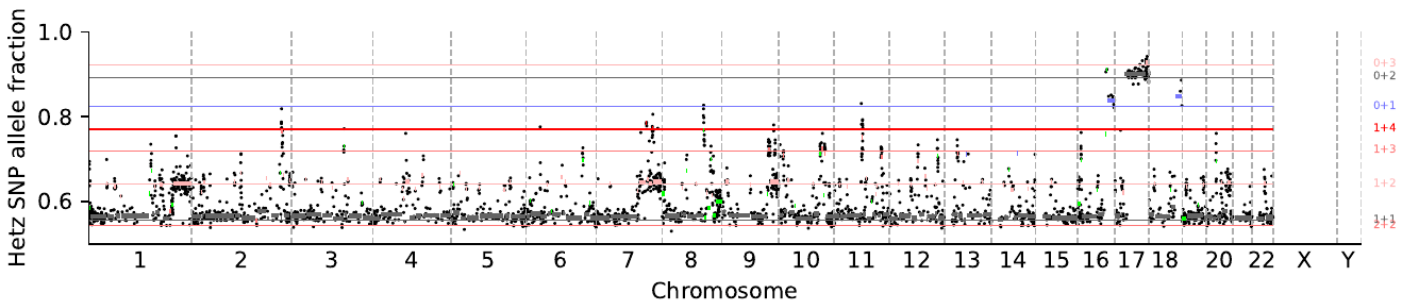
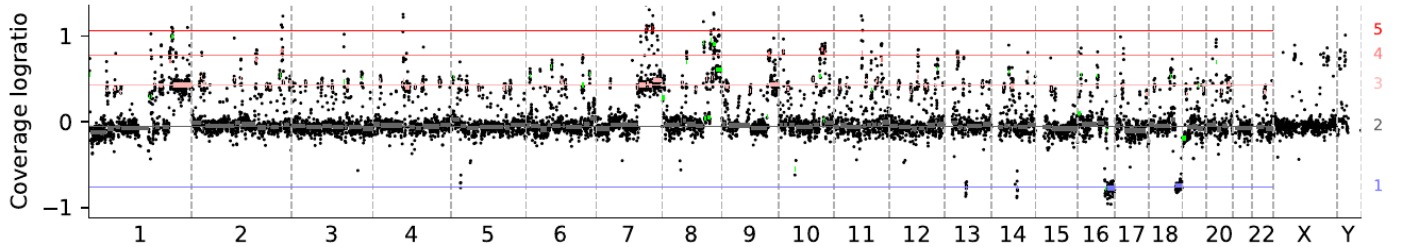
AE-046 clonal expansion timeline



Patient AE-092

AE-092-Baseline-cfDNA

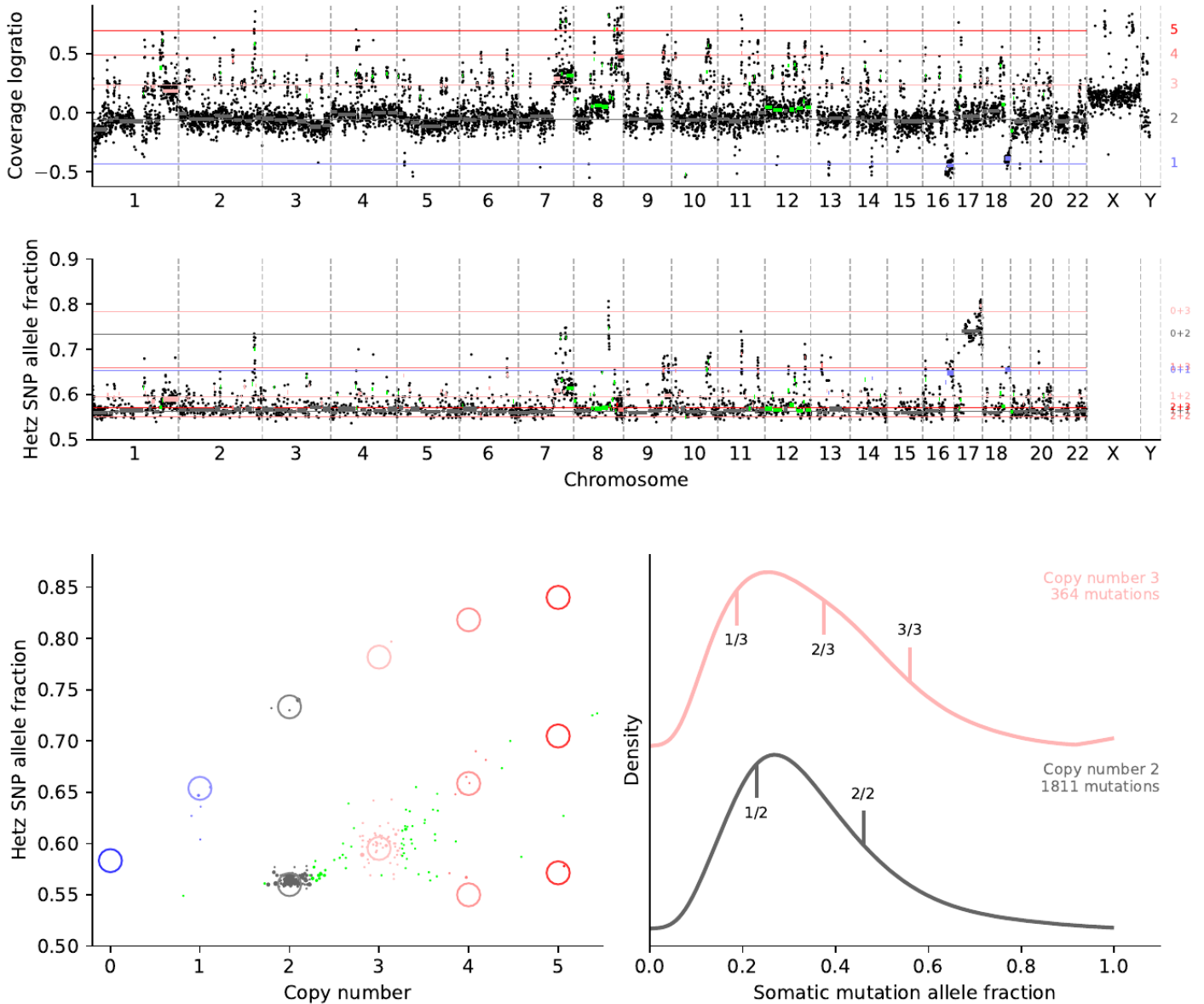
Cancer fraction: 78%, Diploid level: -0.05



PLEASE NOTE: Low median sequencing depth (24x) of this sample has resulted in higher-than-usual noise levels.

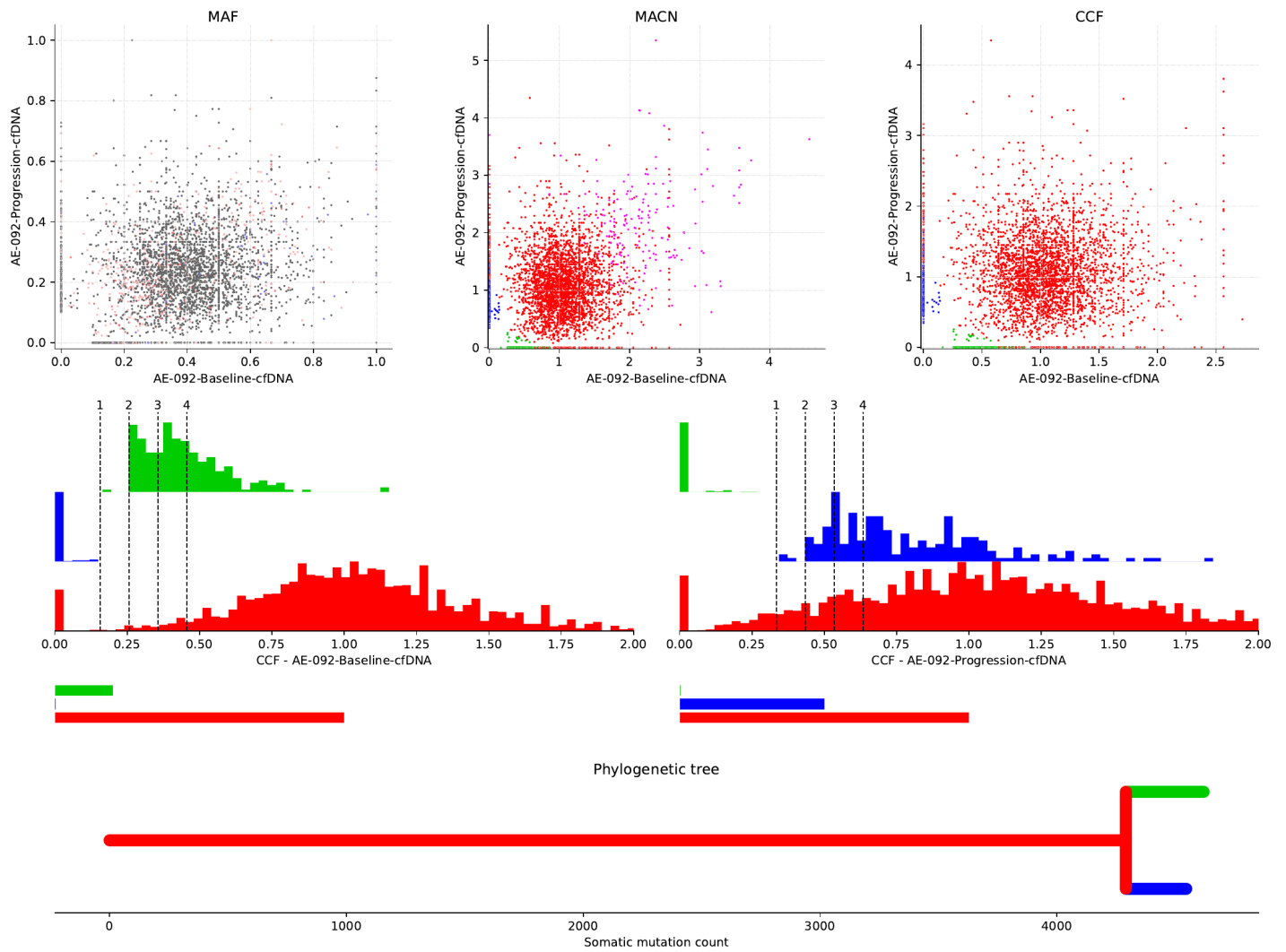
AE-092-Progression-cfDNA

Cancer fraction: 46%, Diploid level: -0.06

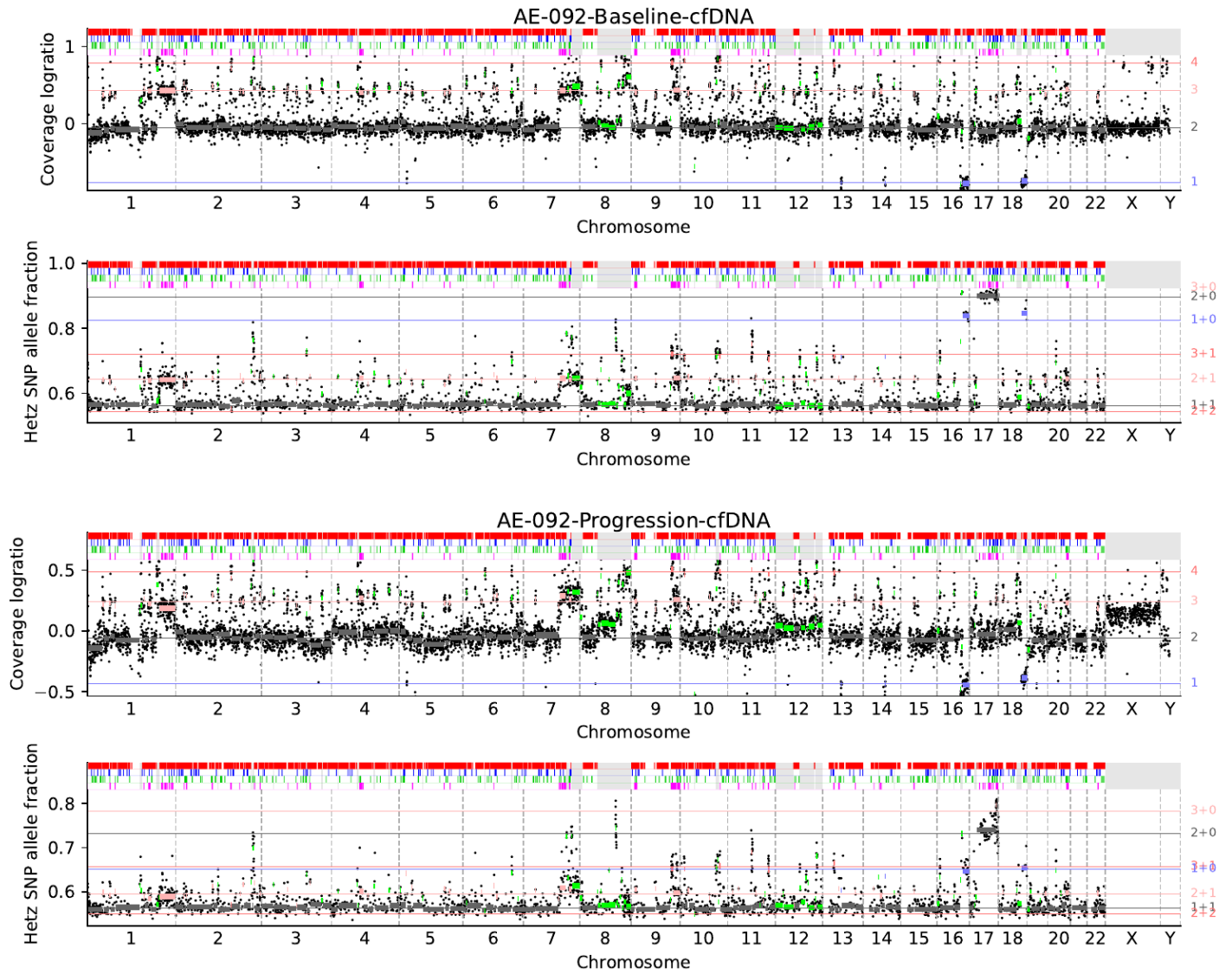


PLEASE NOTE: Low median sequencing depth (28x) of this sample has resulted in higher-than-usual noise levels.

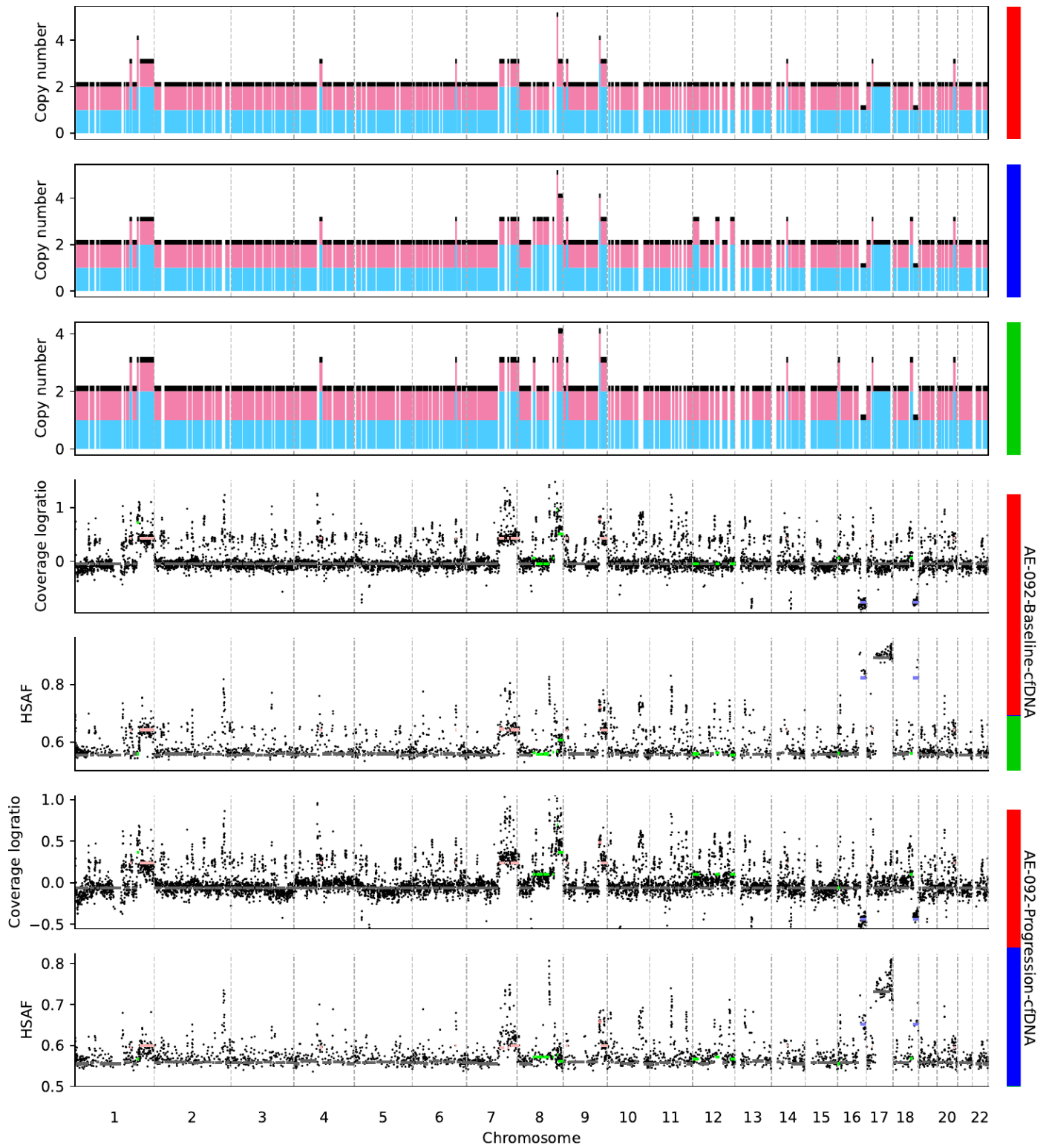
AE-092 mutation clustering



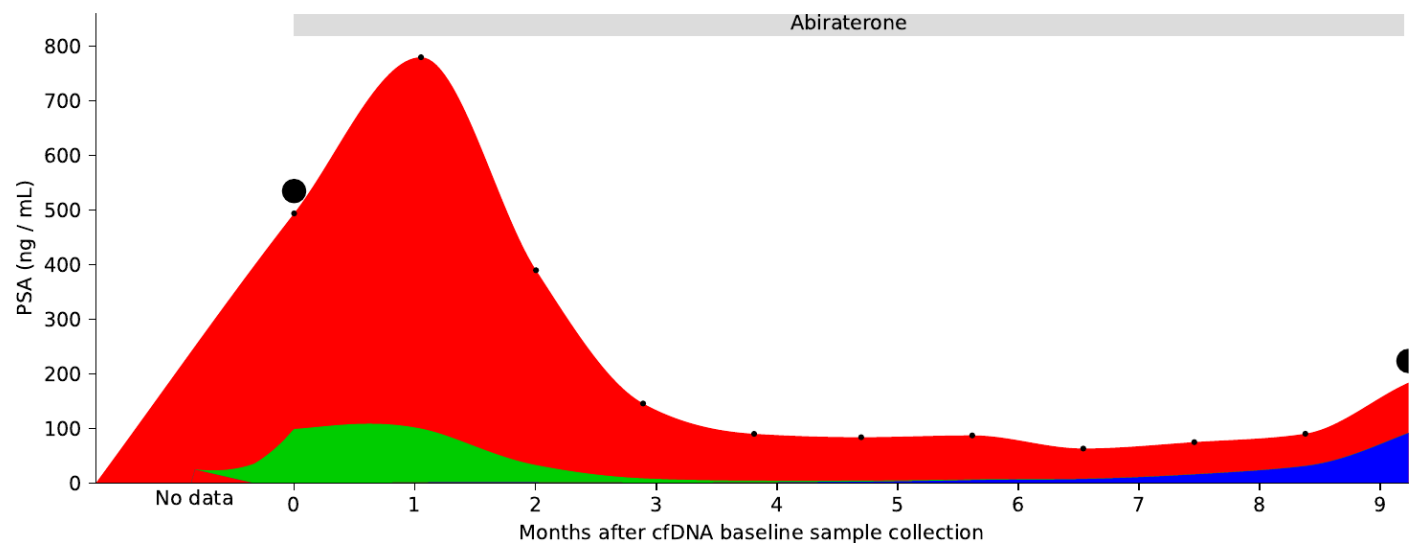
AE-092 mutation cluster positions



AE-092 copy number deconvolution



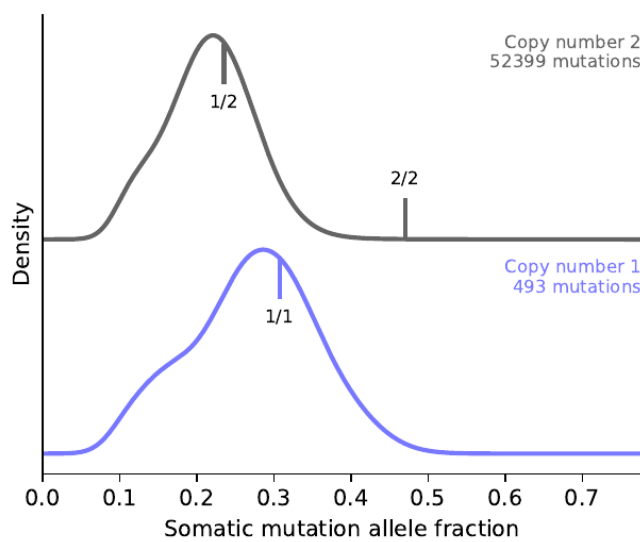
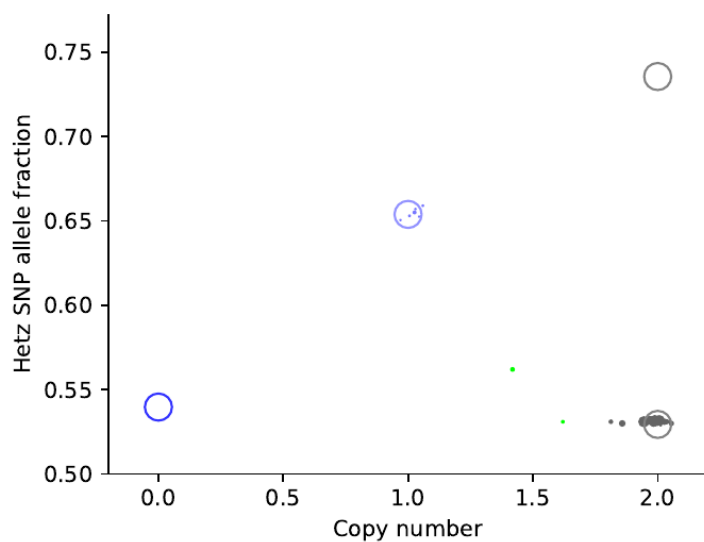
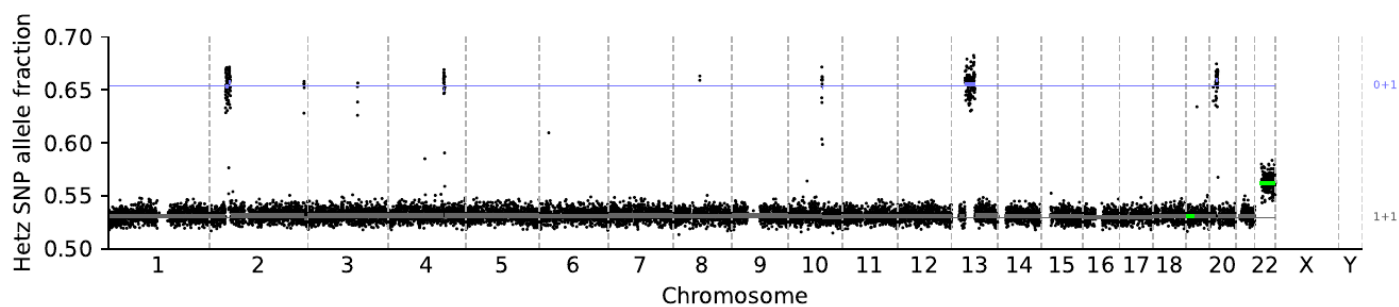
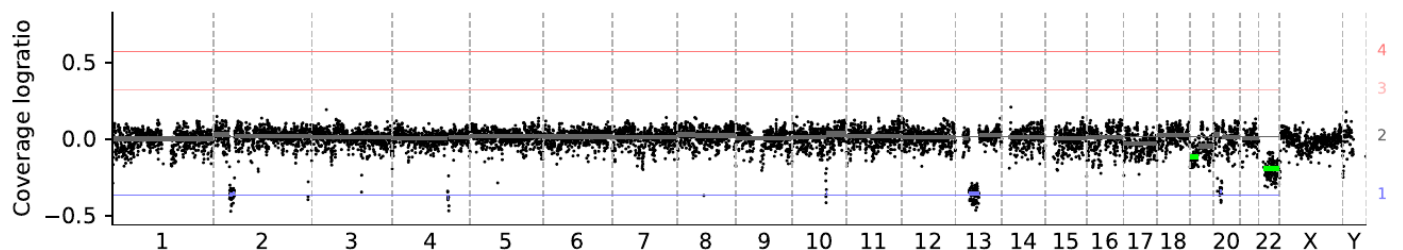
AE-092 clonal expansion timeline



Patient AE-098

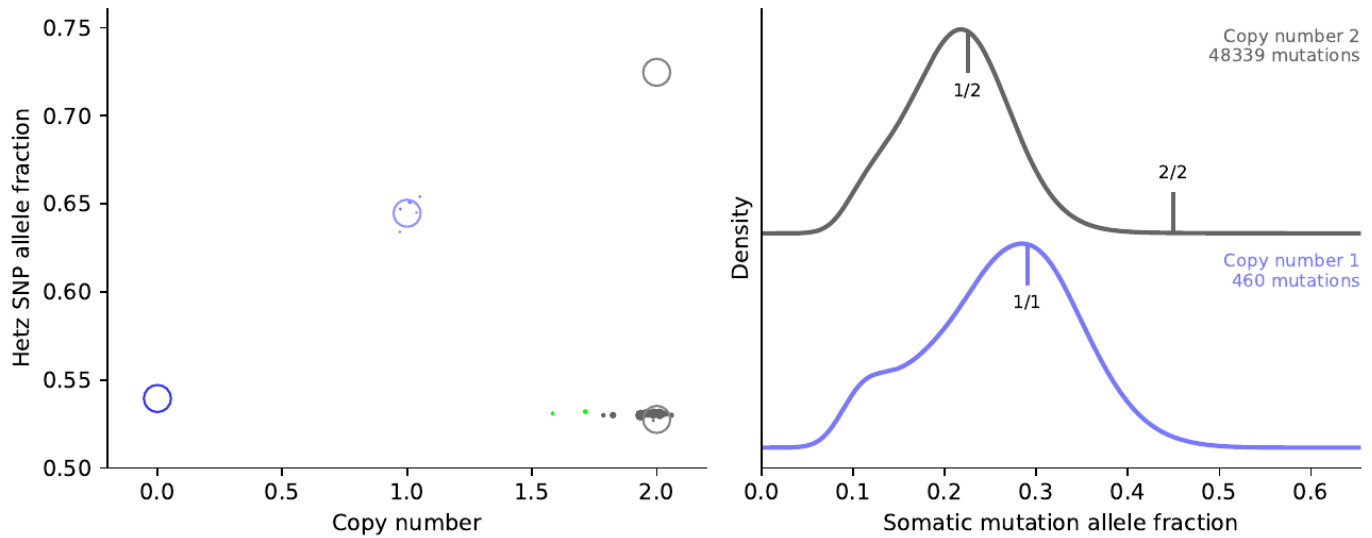
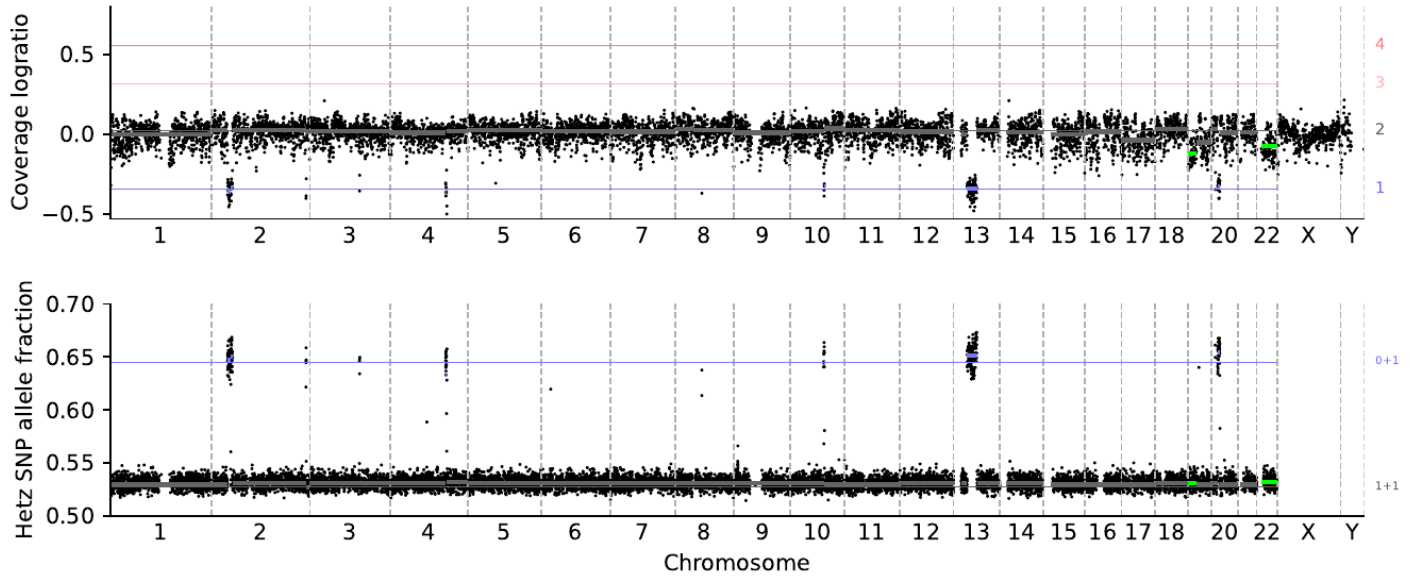
AE-098-Baseline-cfDNA

Cancer fraction: 47%, Diploid level: 0.02

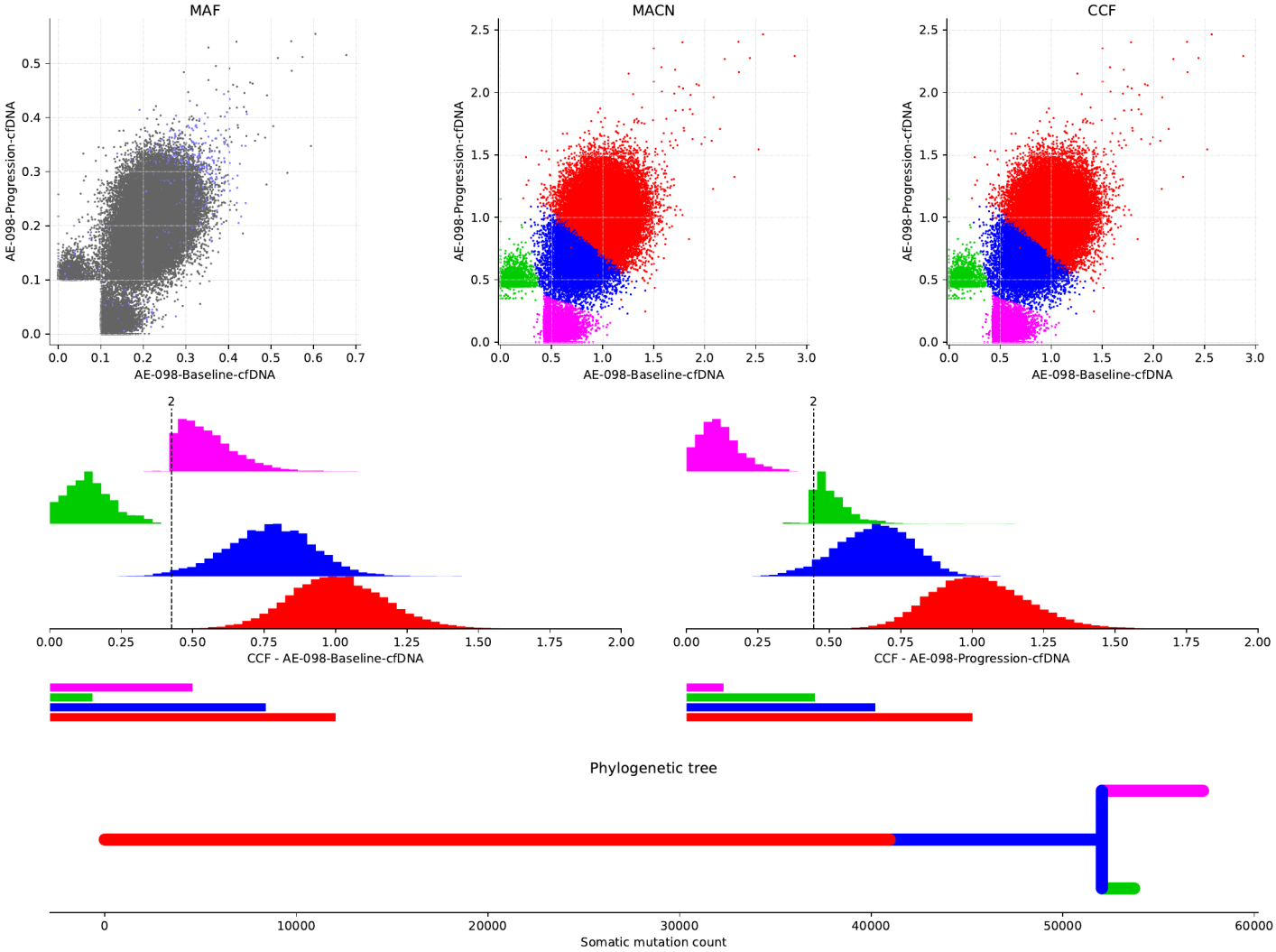


AE-098-Progression-cfDNA

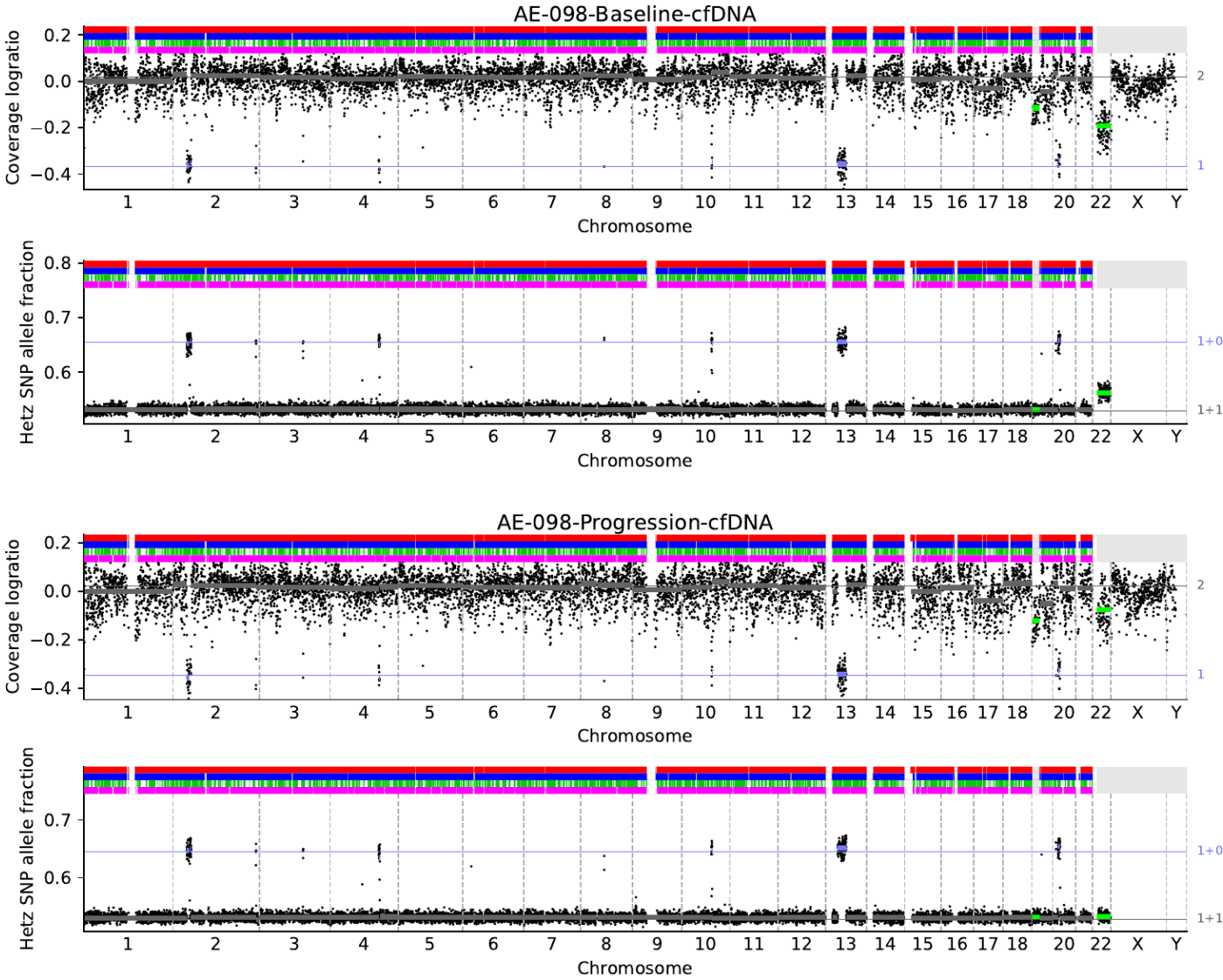
Cancer fraction: 45%, Diploid level: 0.02



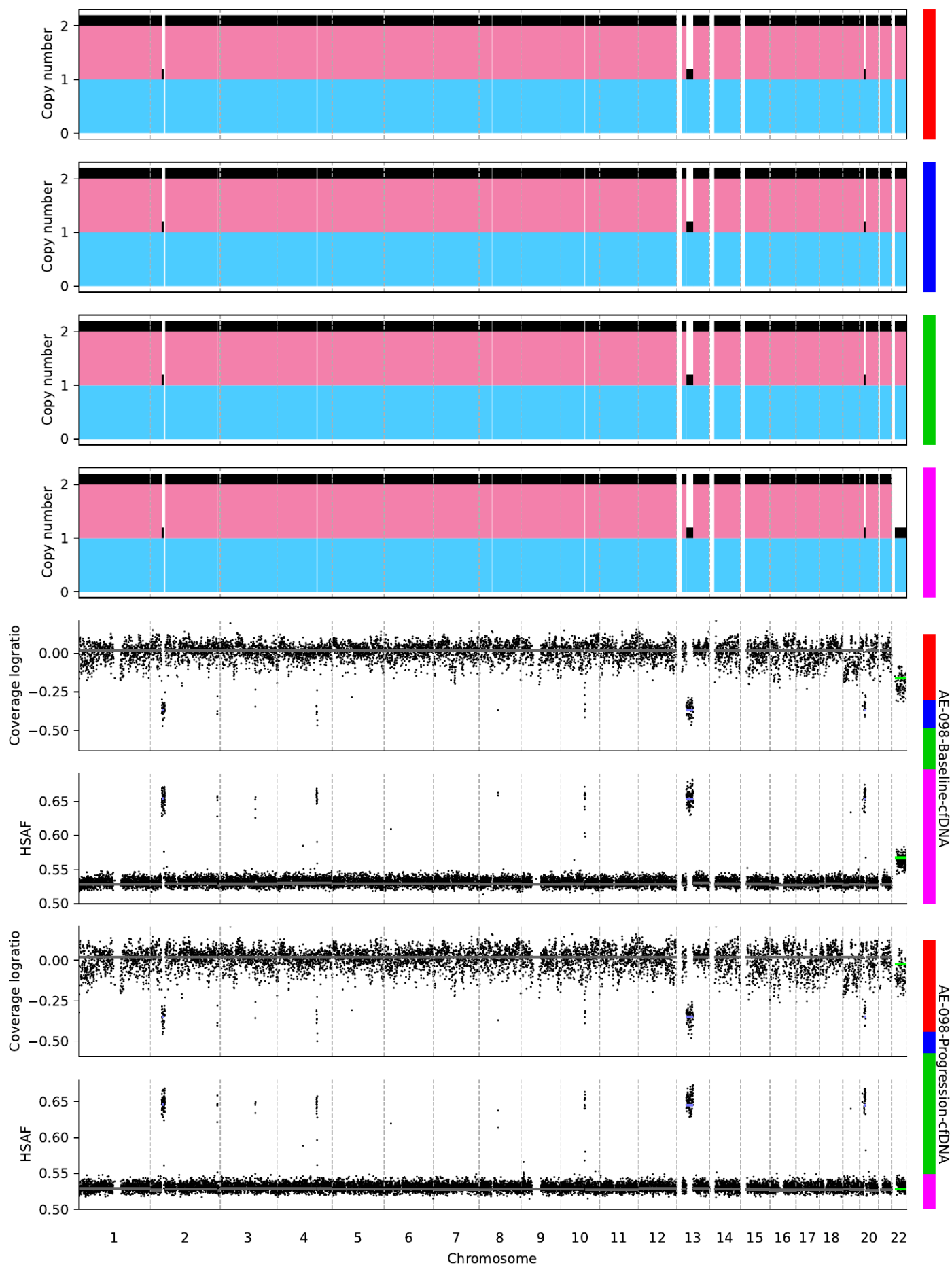
AE-098 mutation clustering



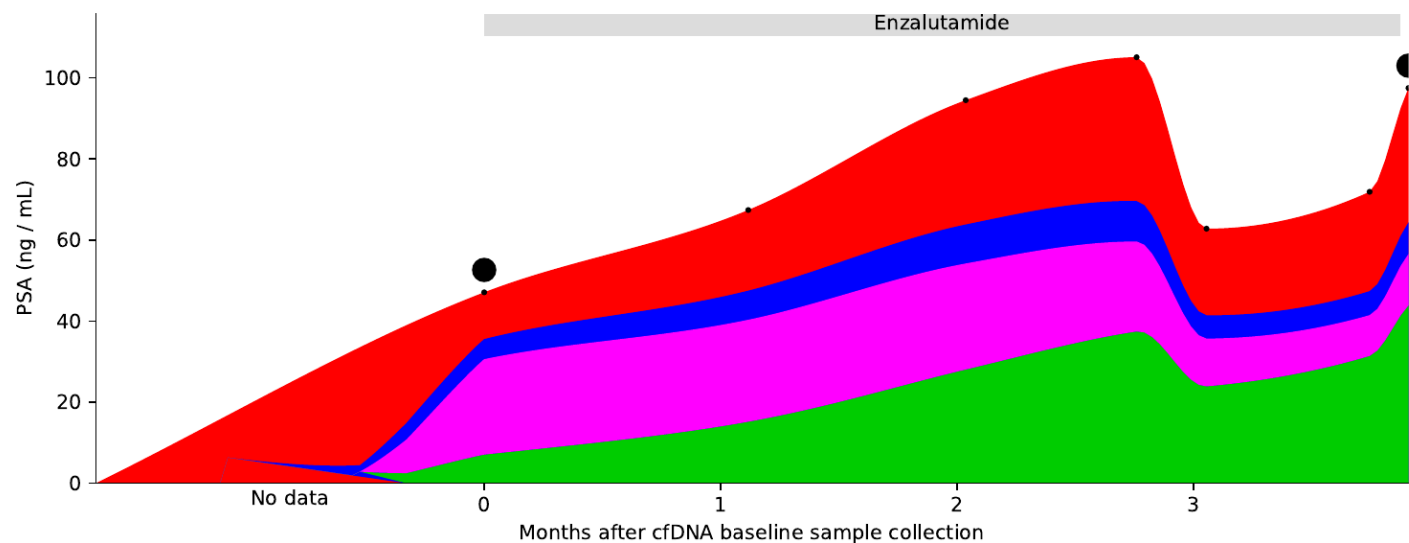
AE-098 mutation cluster positions



AE-098 copy number deconvolution



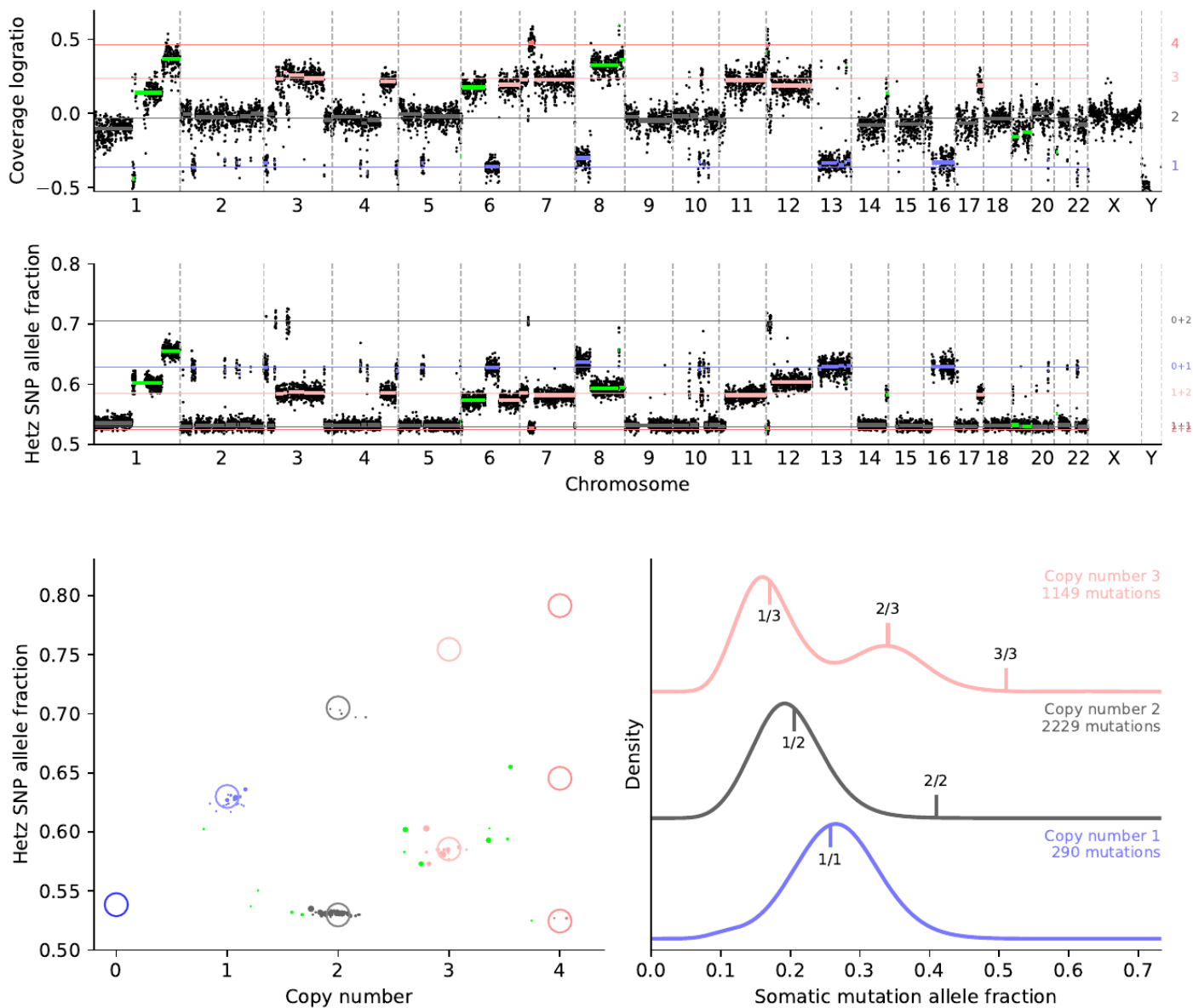
AE-098 clonal expansion timeline



Patient AE-099

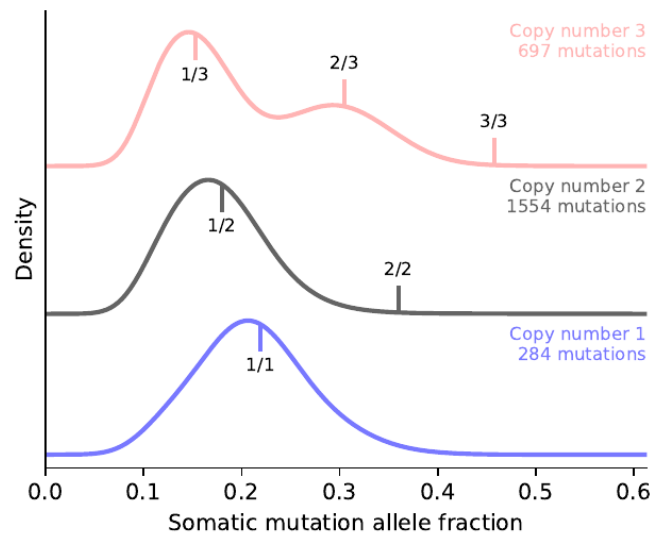
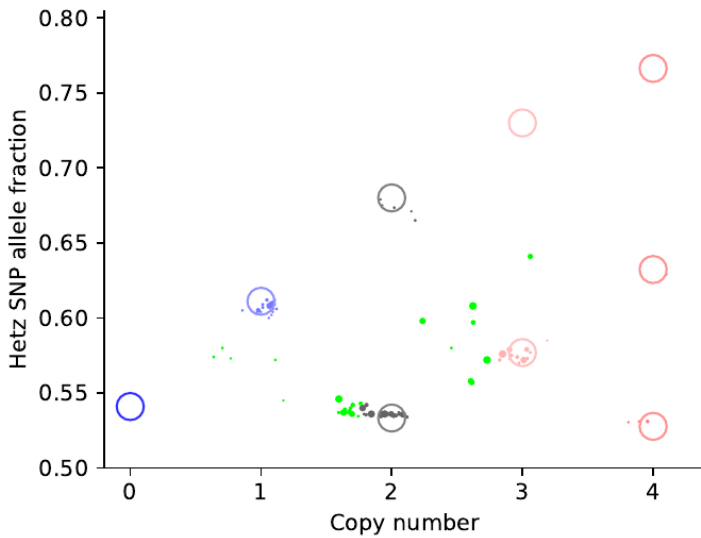
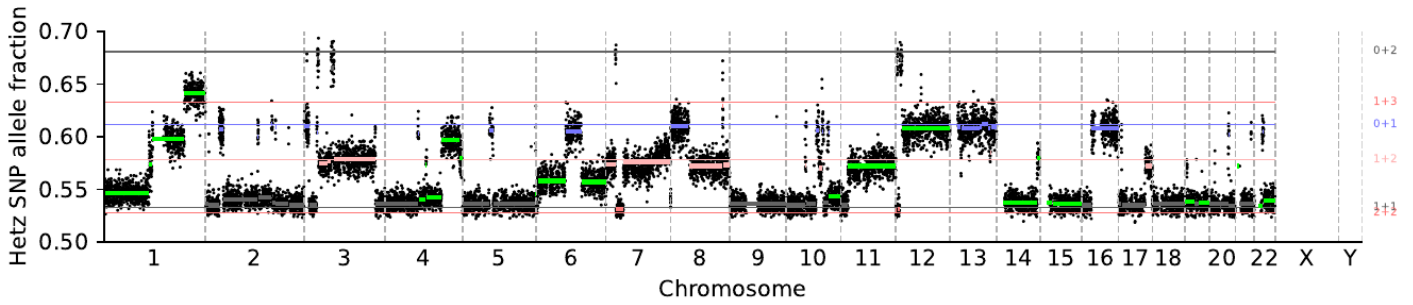
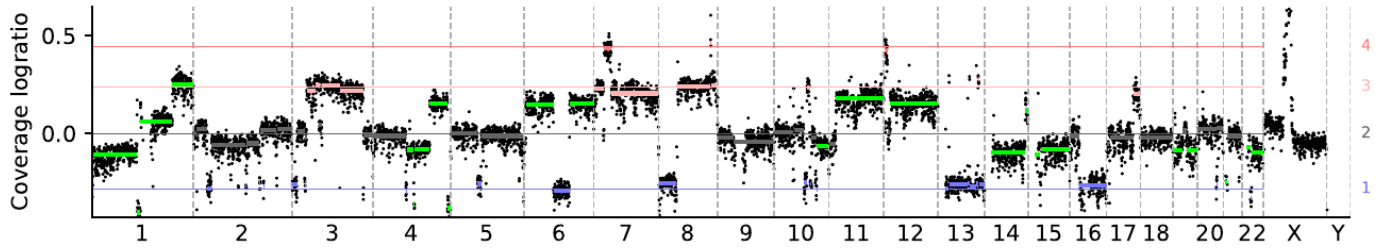
AE-099-Baseline-cfDNA

Cancer fraction: 41%, Diploid level: -0.03



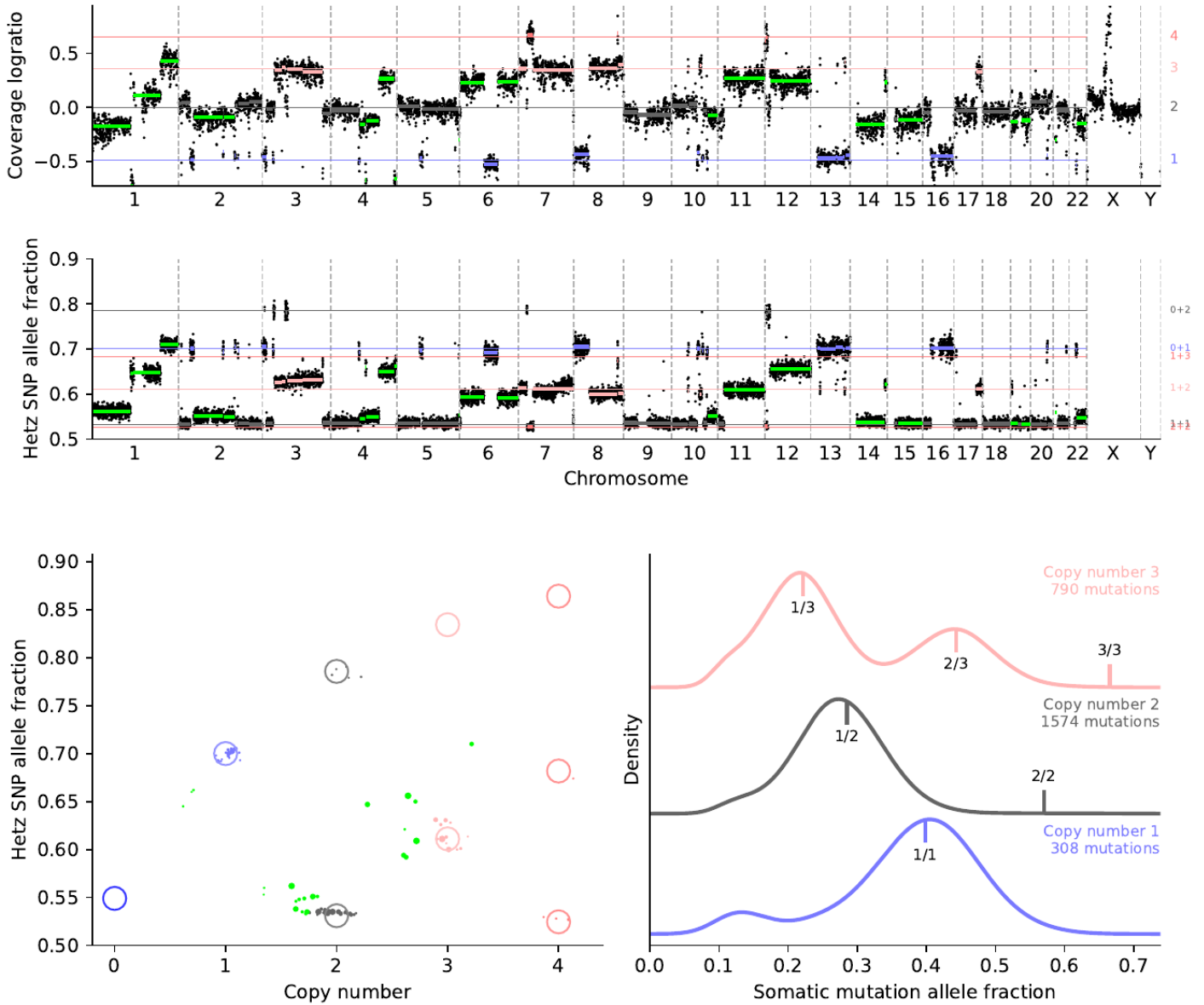
AE-099-Progression-cfDNA

Cancer fraction: 36%, Diploid level: 0.00

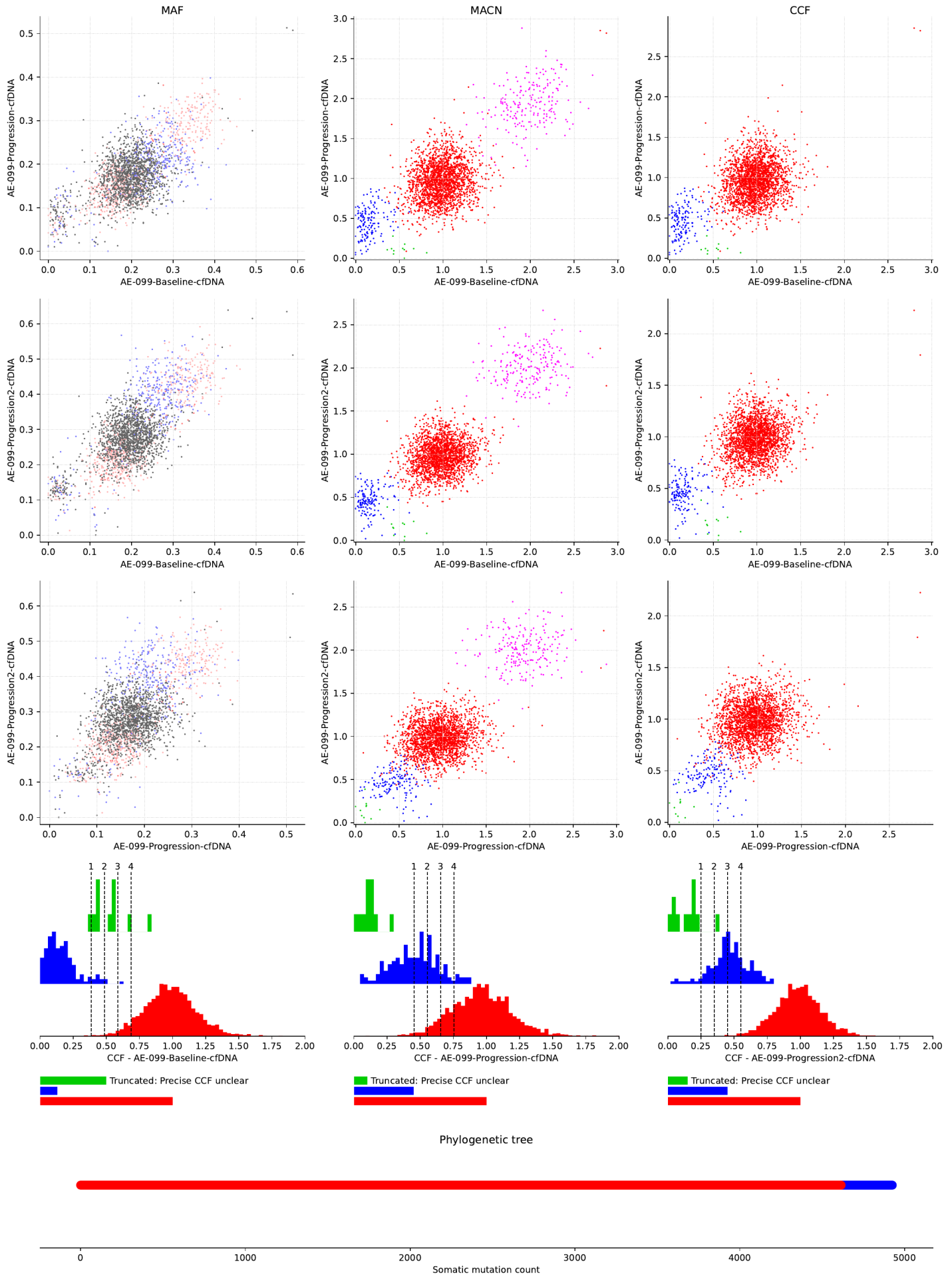


AE-099-Progression2-cfDNA

Cancer fraction: 57%, Diploid level: 0.00

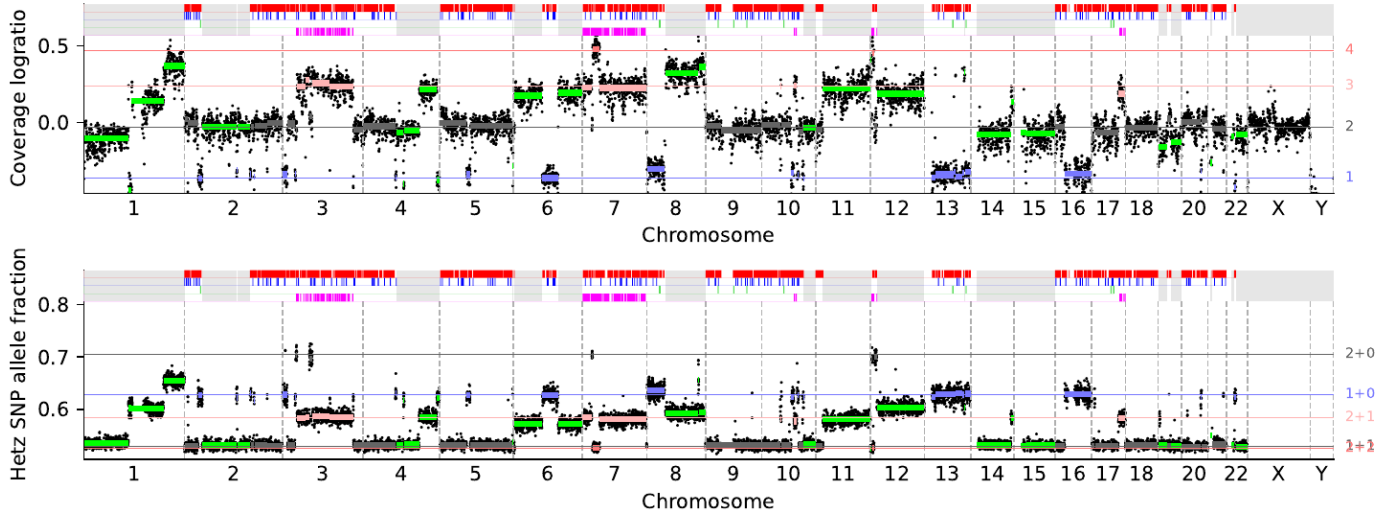


AE-099 mutation clustering

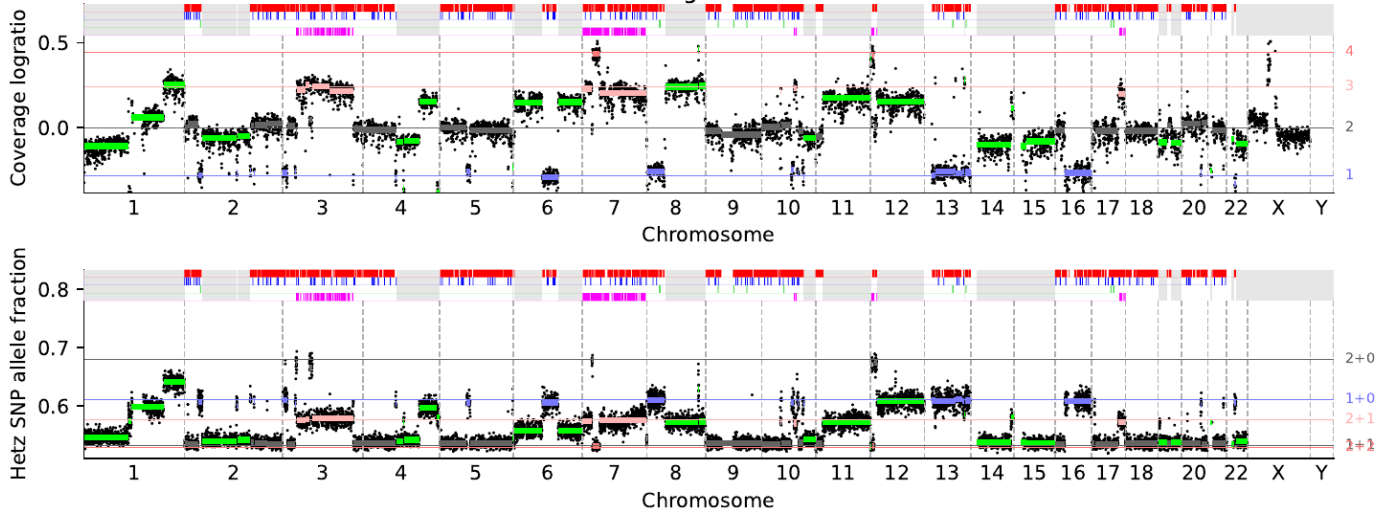


AE-099 mutation cluster positions

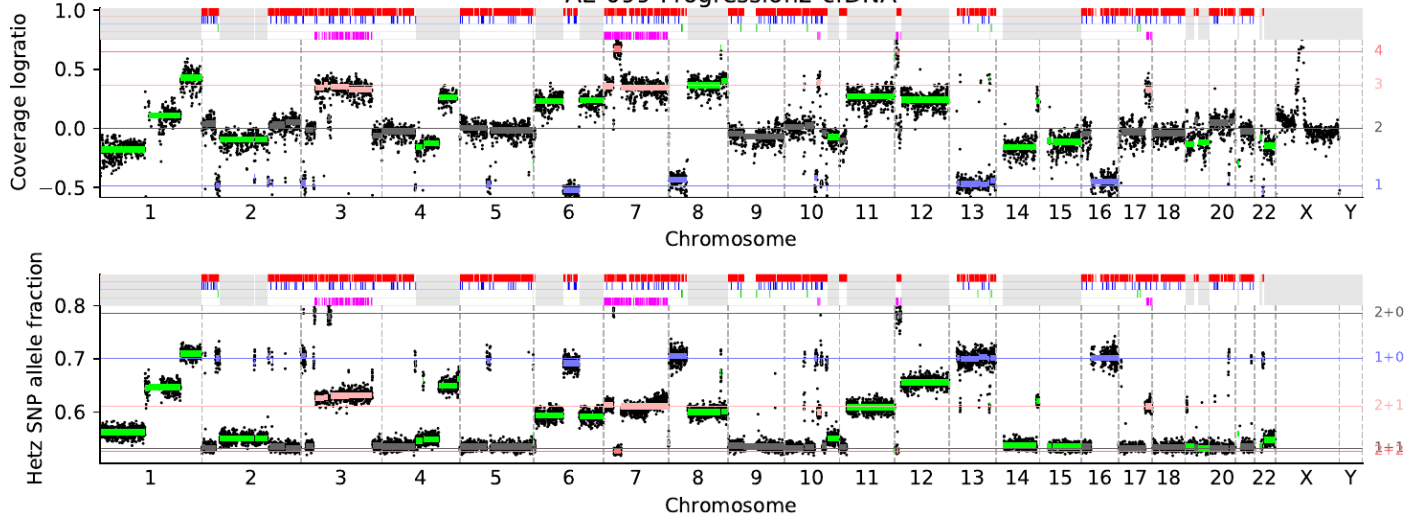
AE-099-Baseline-cfDNA



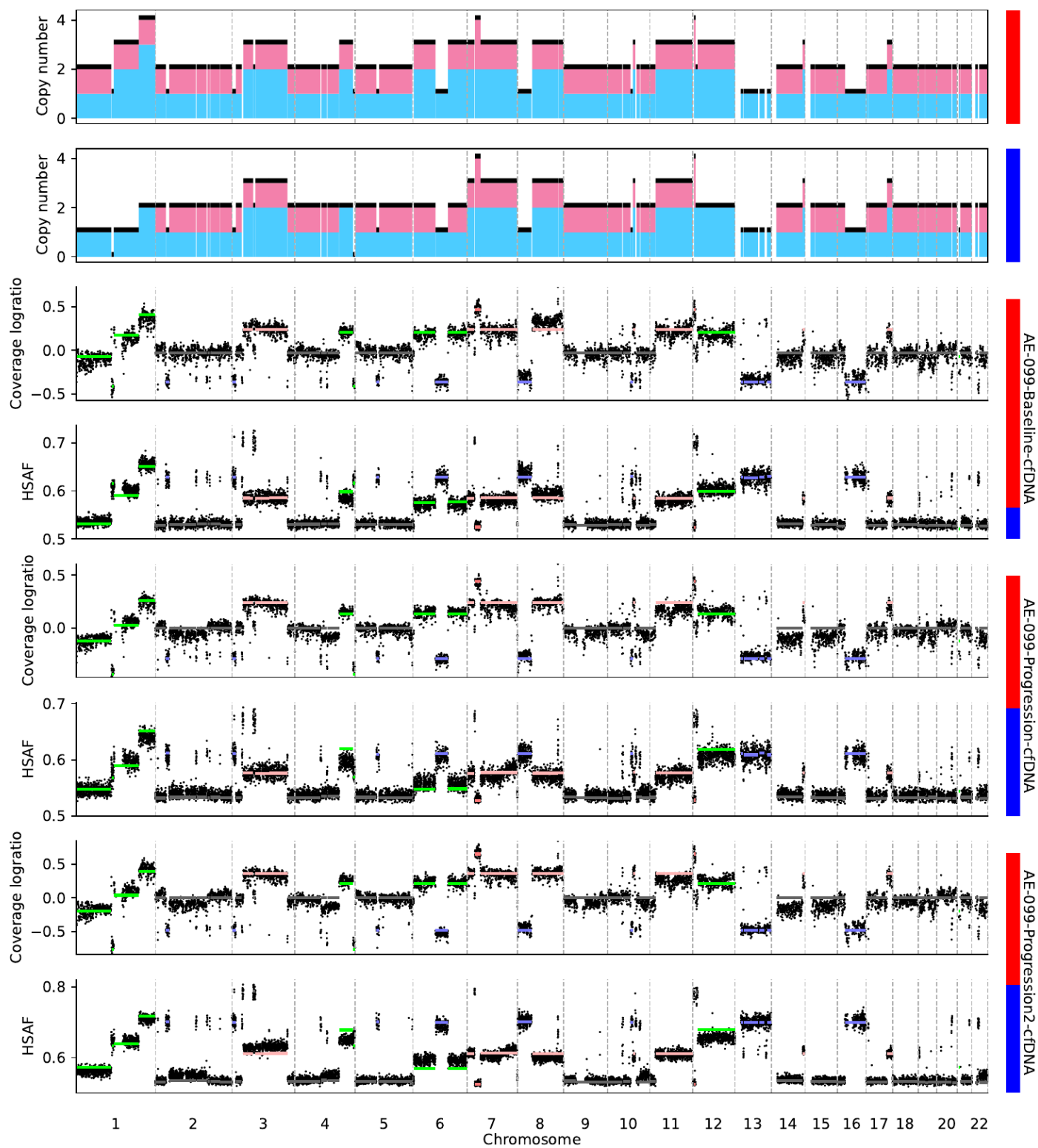
AE-099-Progression-cfDNA



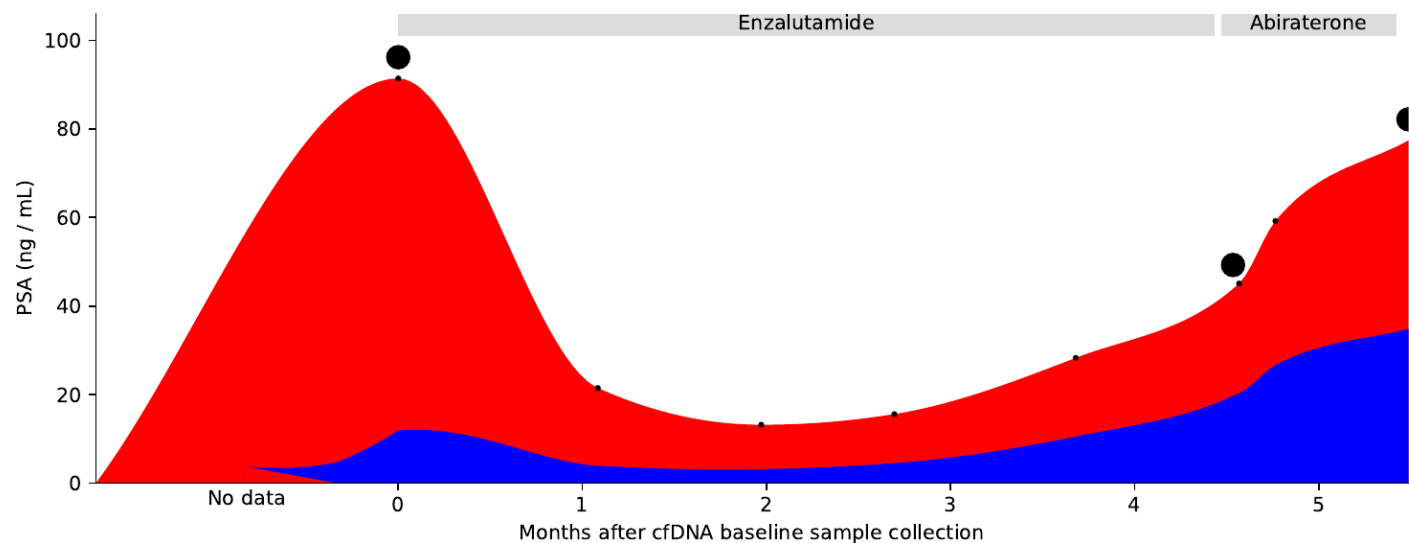
AE-099-Progression2-cfDNA



AE-099 copy number deconvolution



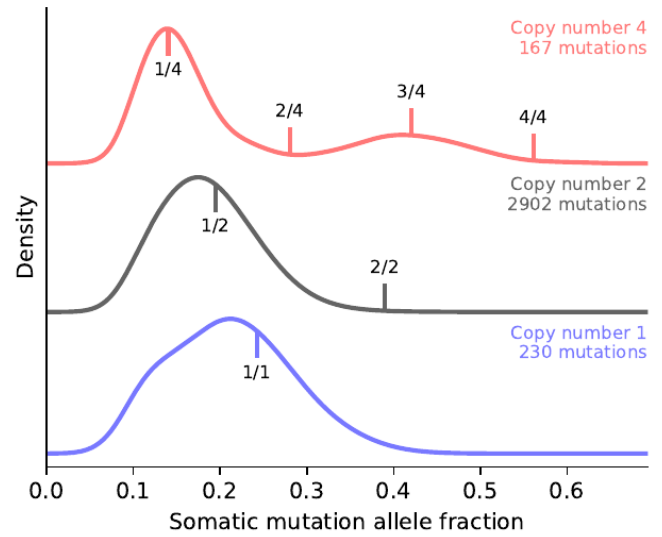
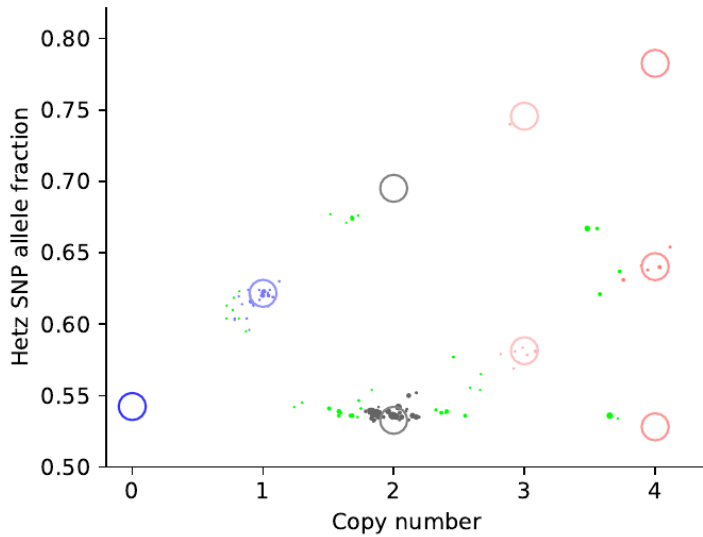
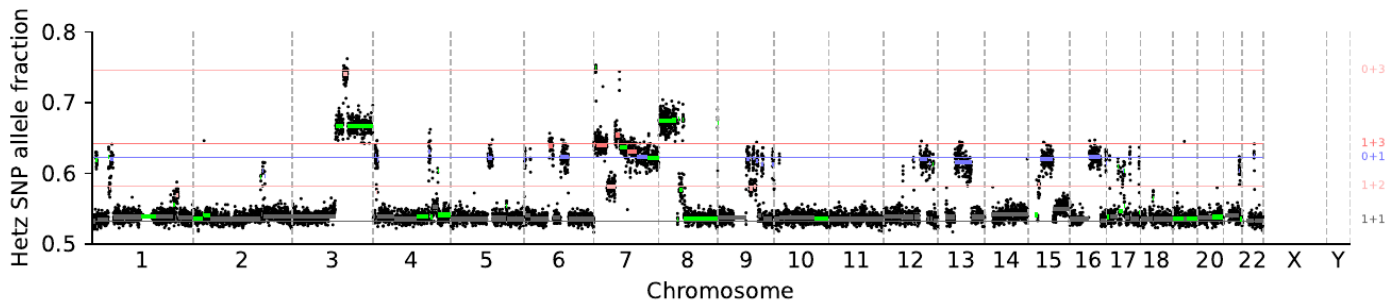
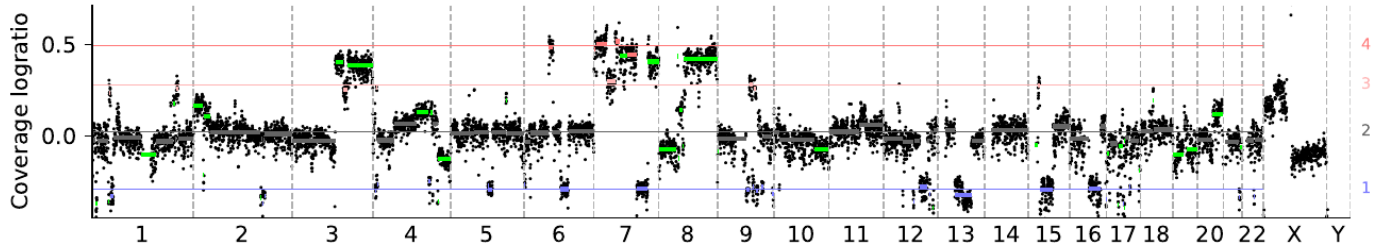
AE-099 clonal expansion timeline



Patient AE-110

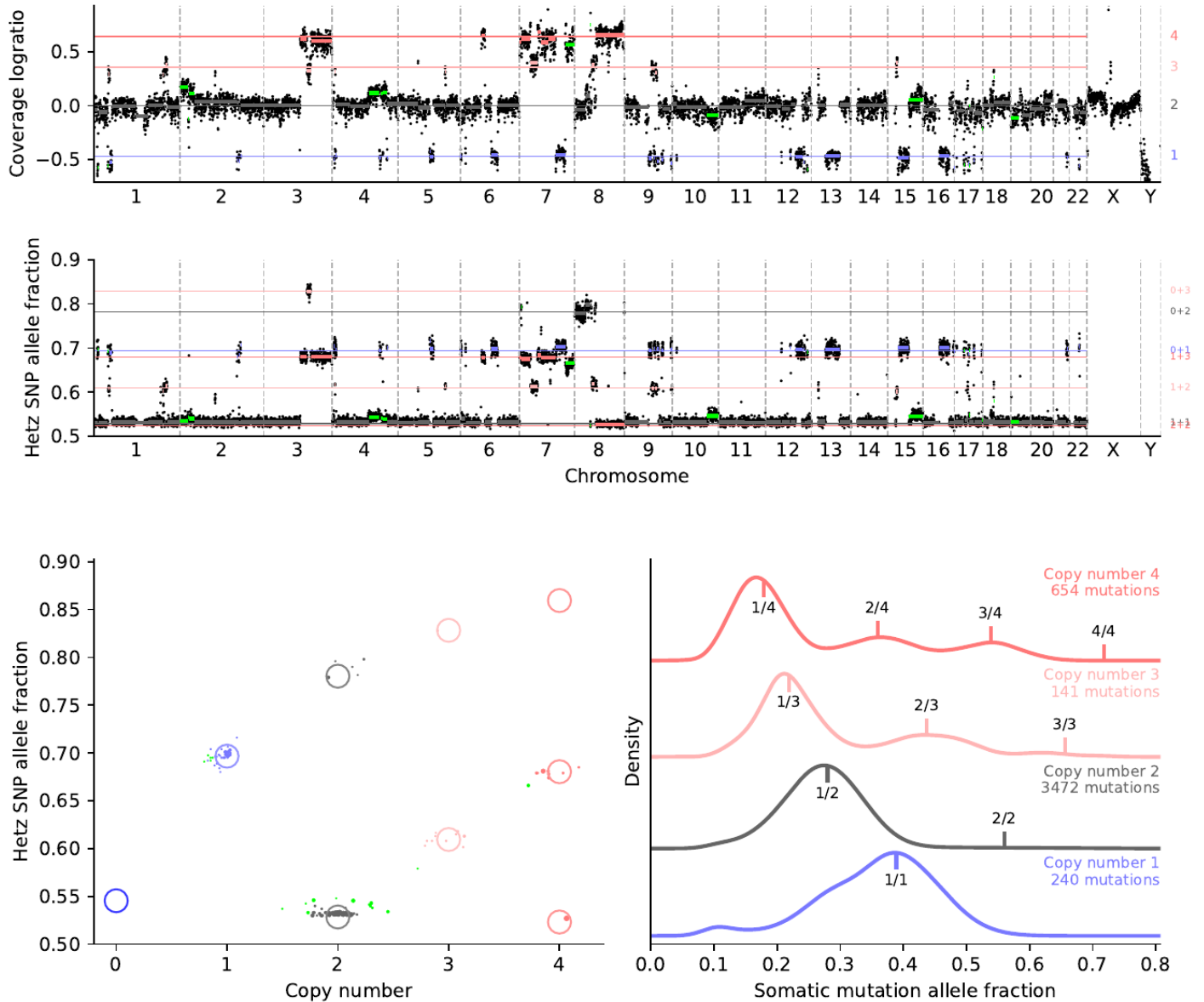
AE-110-Baseline-cfDNA

Cancer fraction: 39%, Diploid level: 0.02

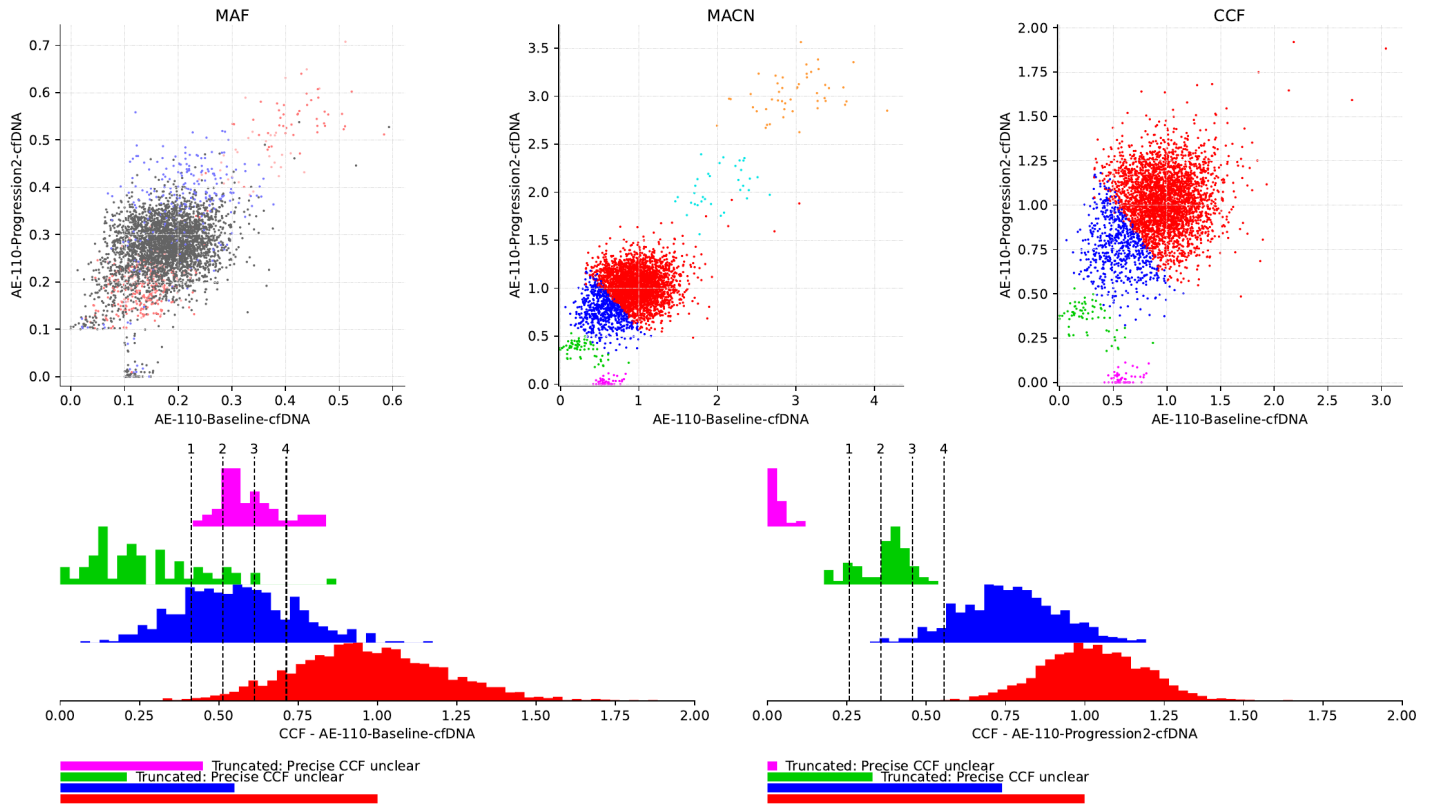


AE-110-Progression2-cfDNA

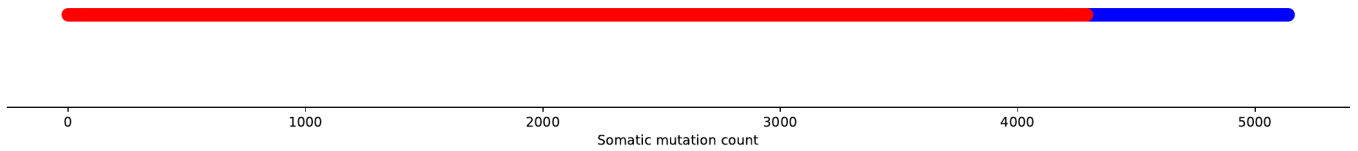
Cancer fraction: 56%, Diploid level: 0.00



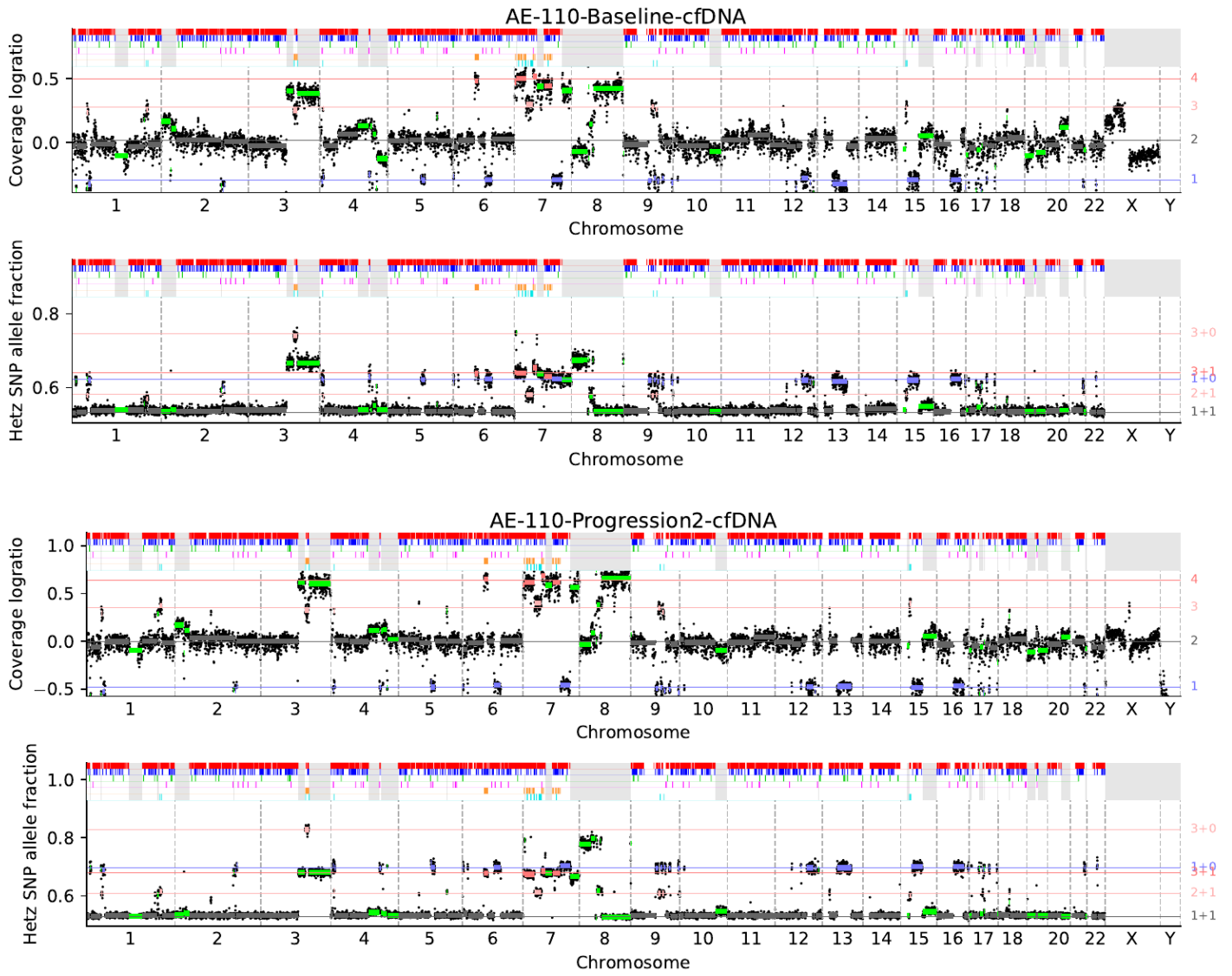
AE-110 mutation clustering



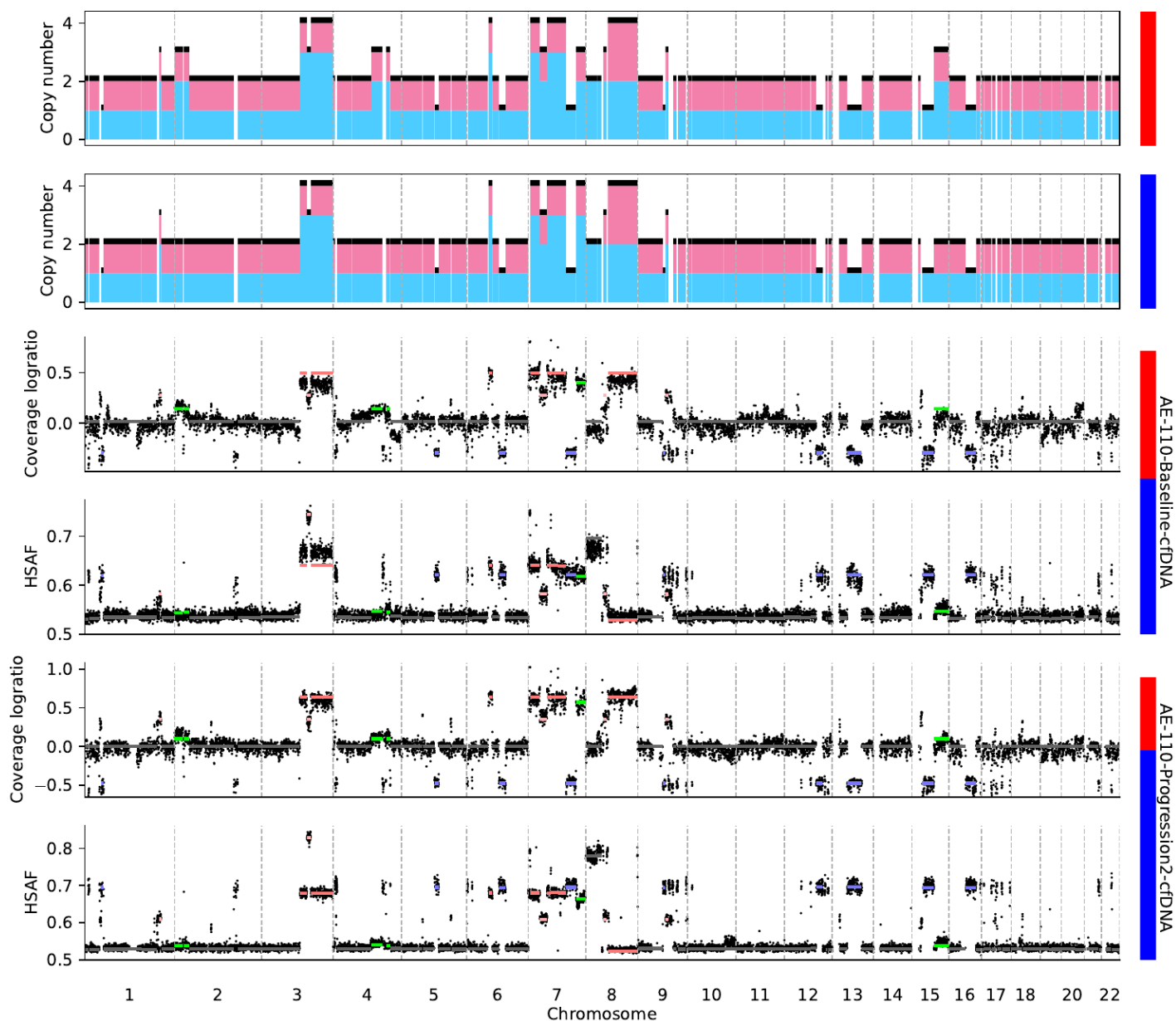
Phylogenetic tree



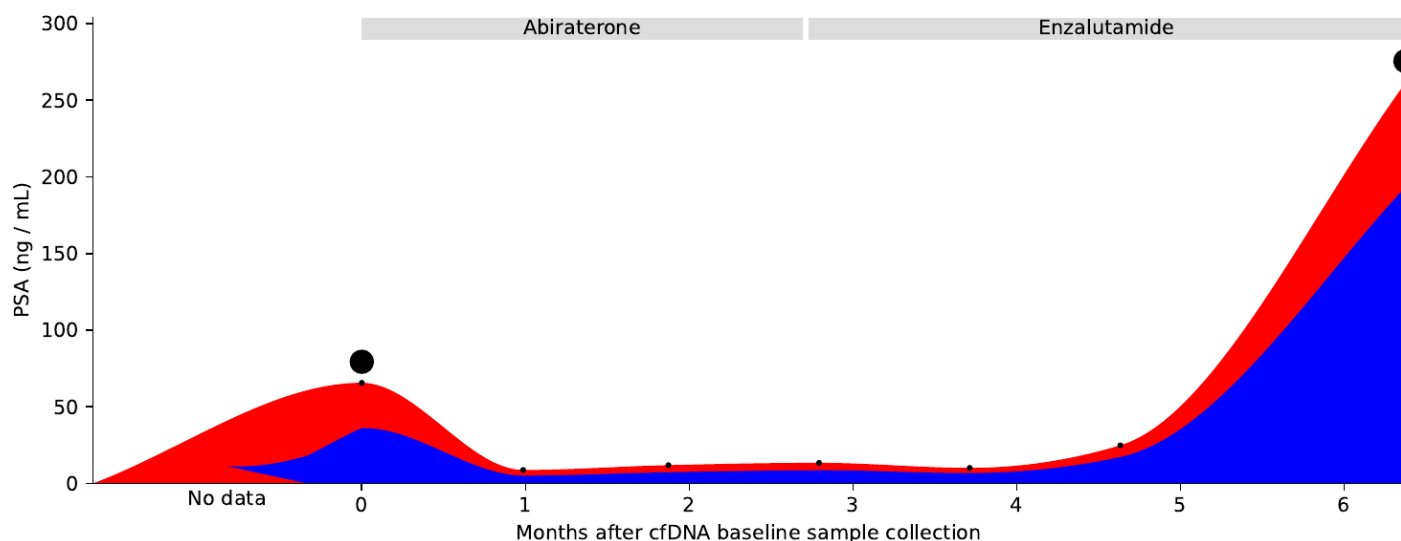
AE-110 mutation cluster positions



AE-110 copy number deconvolution



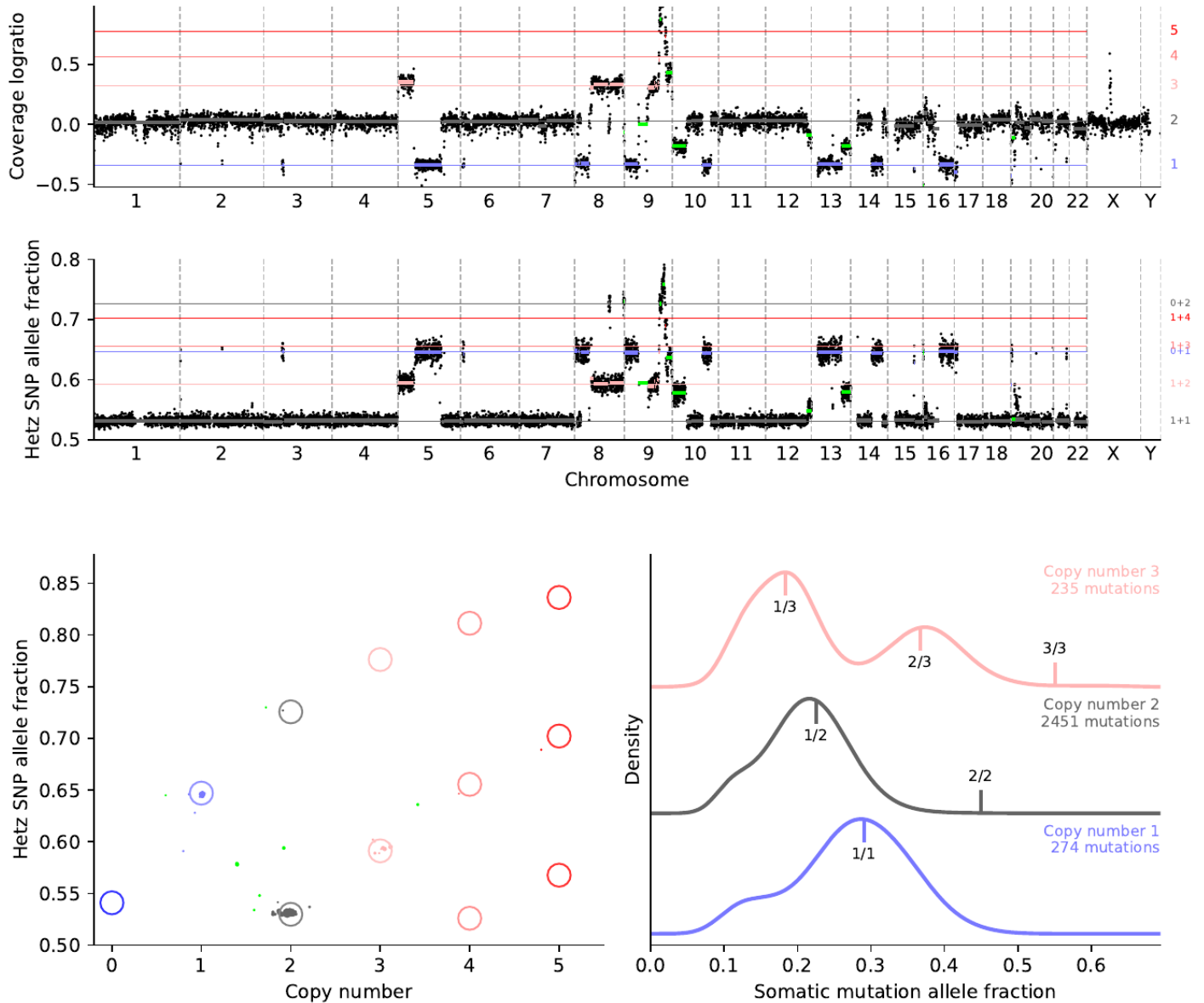
AE-110 clonal expansion timeline



Patient AE-124

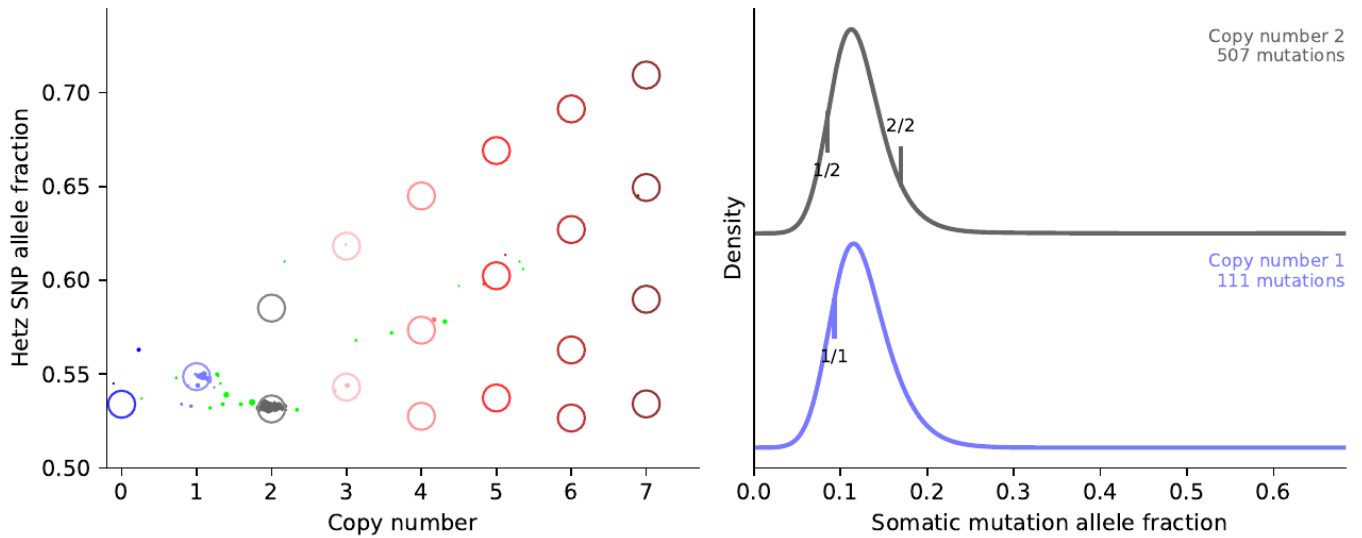
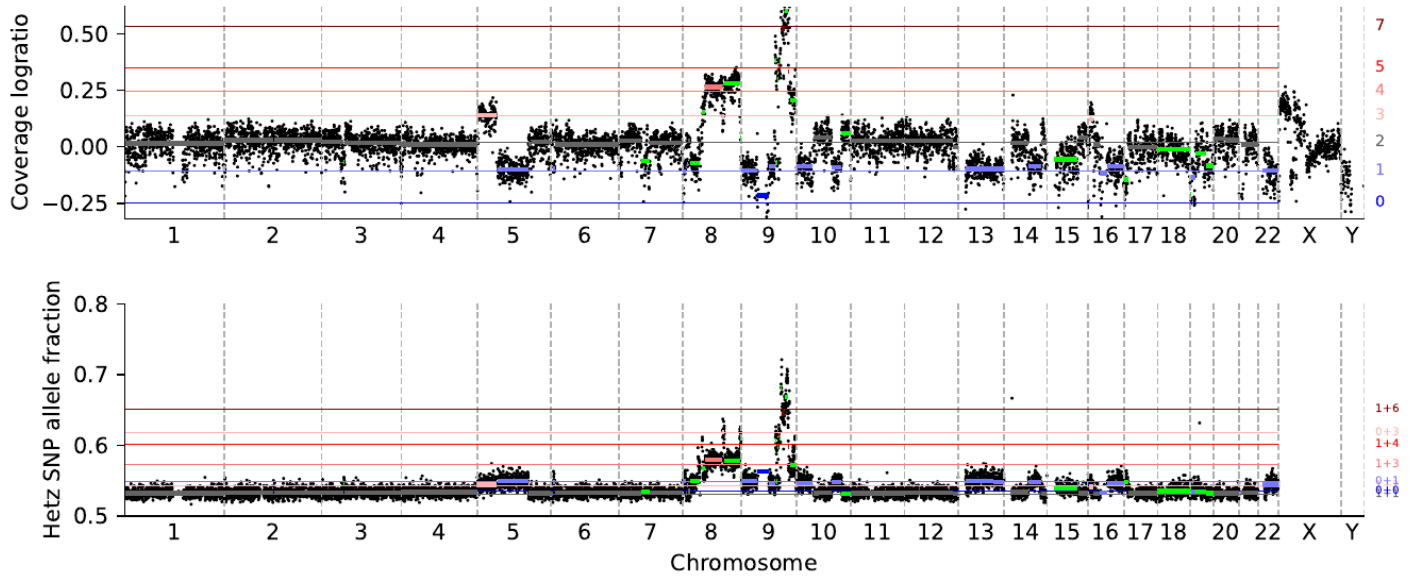
AE-124-Baseline-cfDNA

Cancer fraction: 45%, Diploid level: 0.03

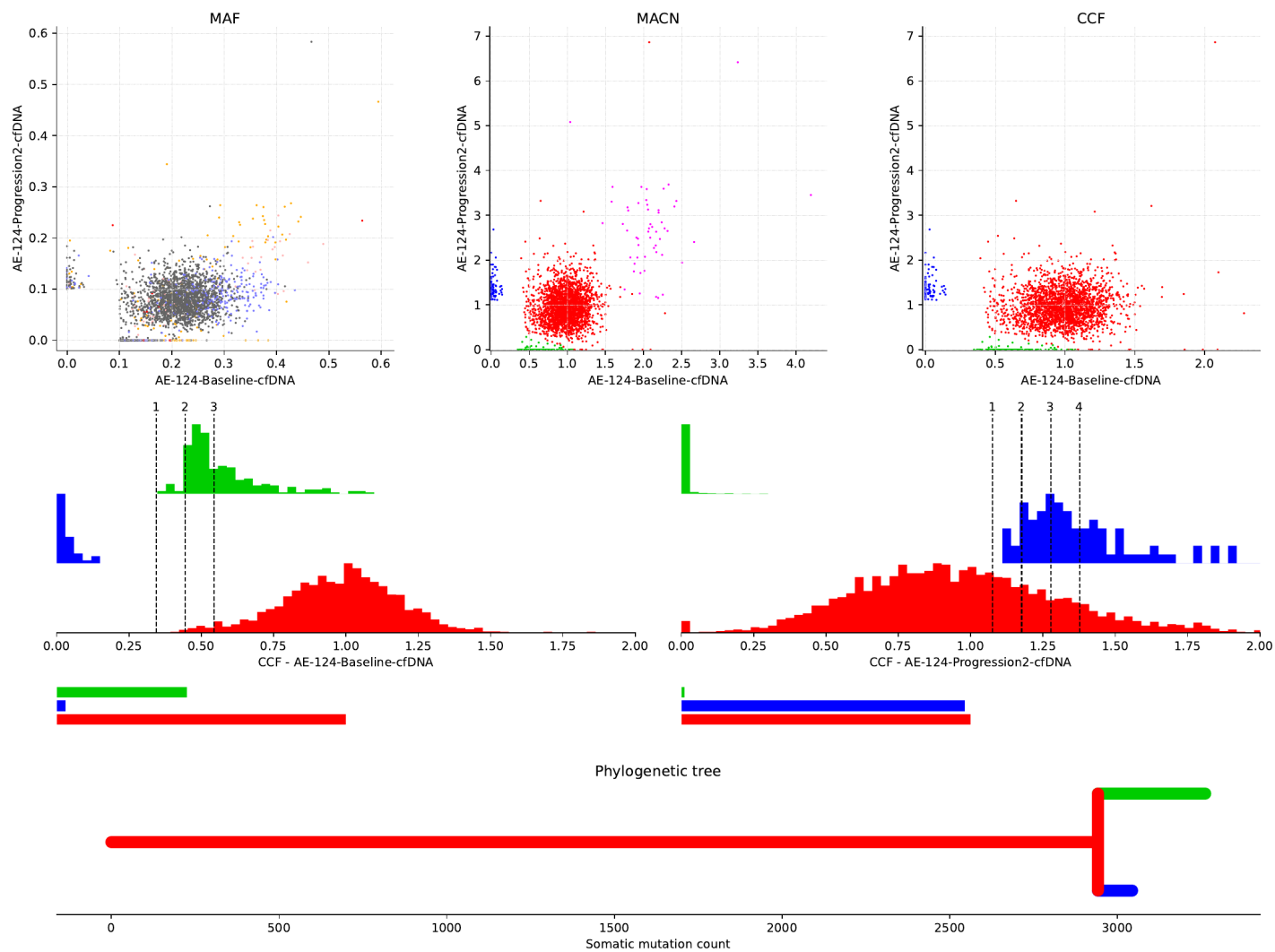


AE-124-Progression2-cfDNA

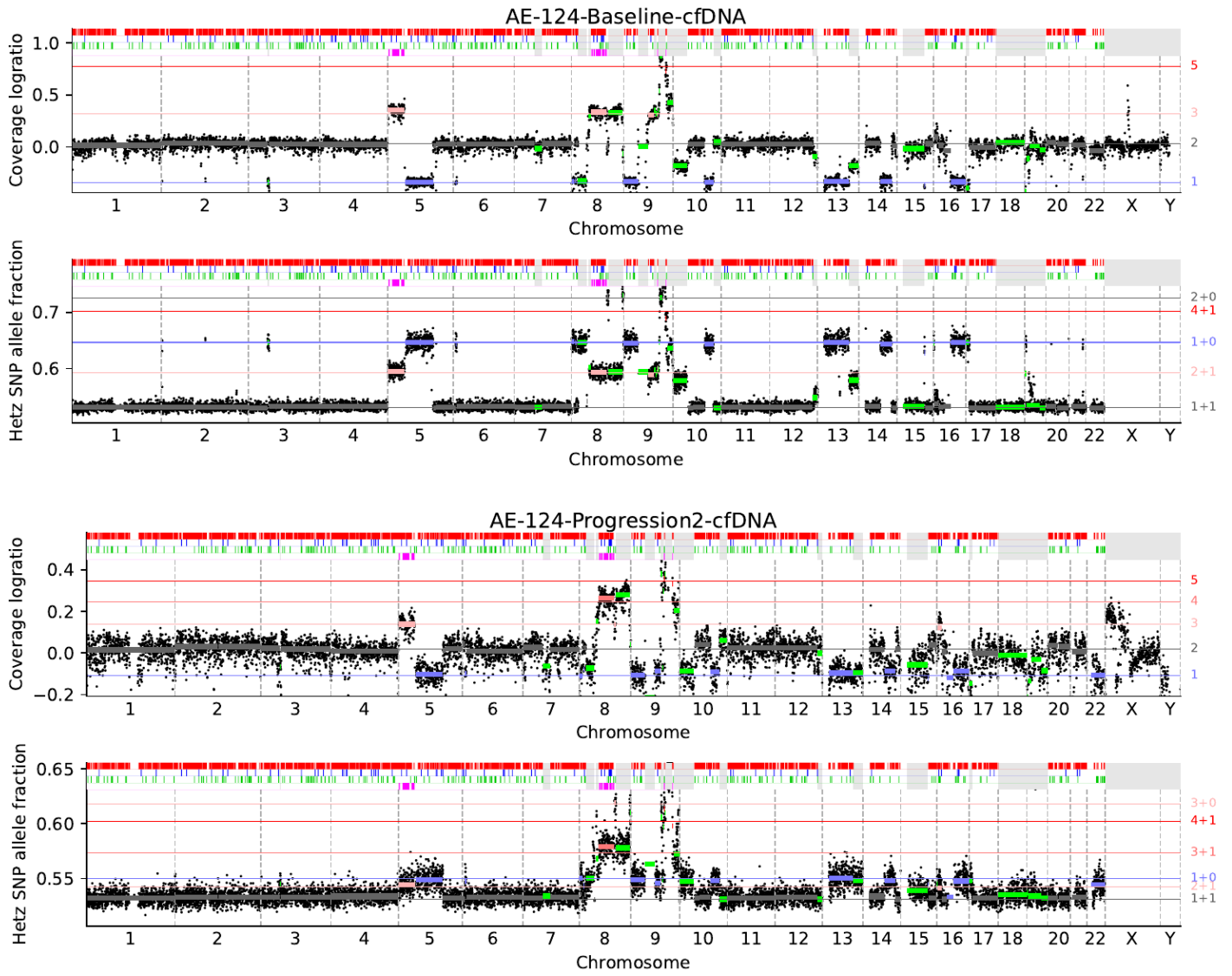
Cancer fraction: 17%, Diploid level: 0.02



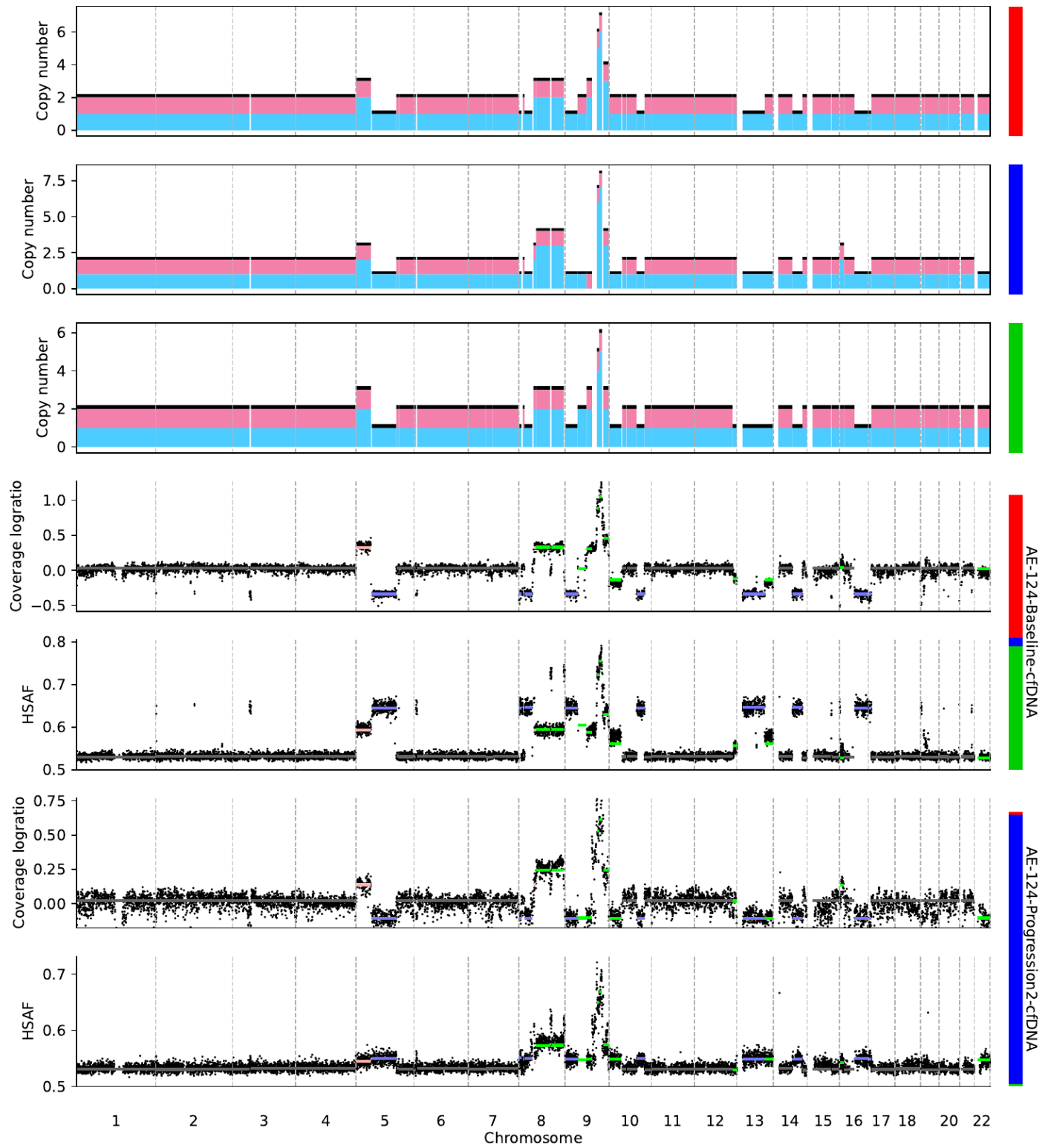
AE-124 mutation clustering



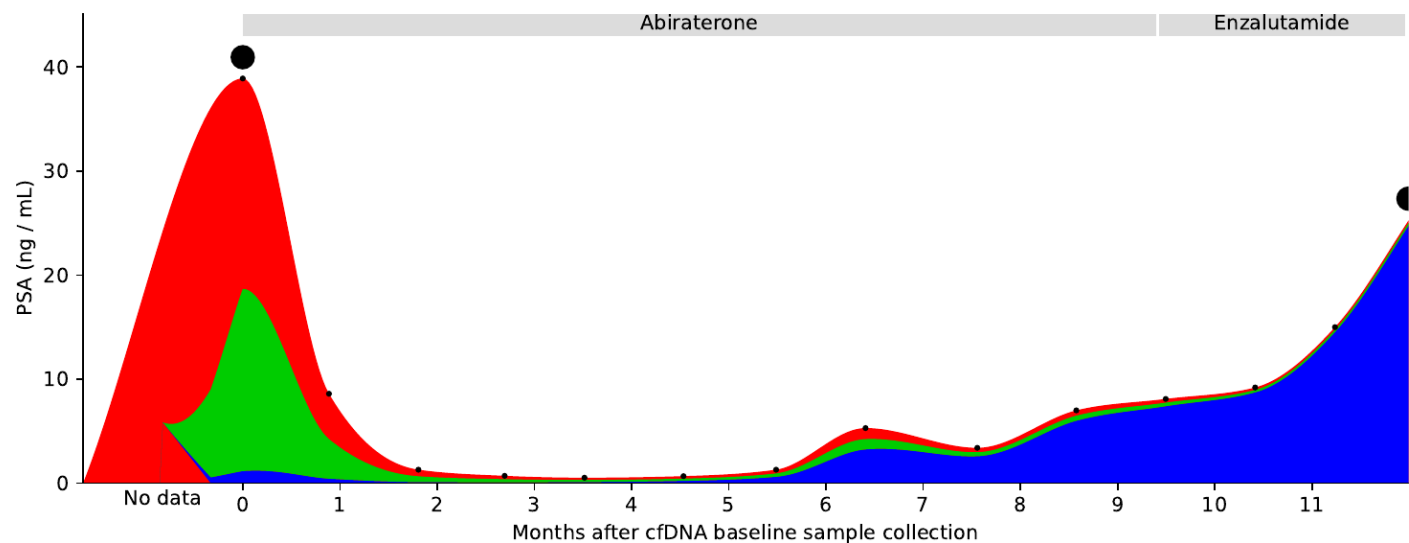
AE-124 mutation cluster positions



AE-124 copy number deconvolution



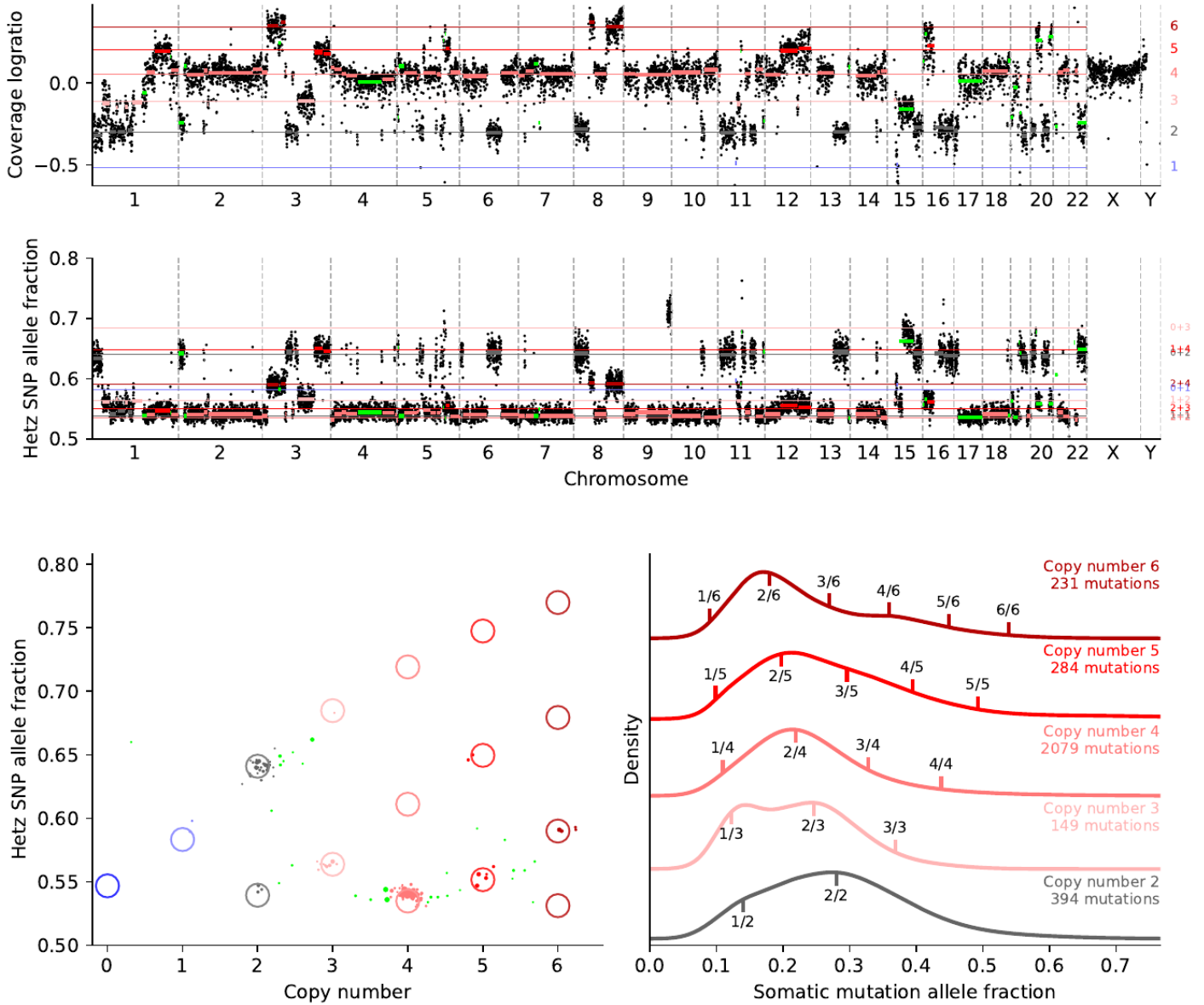
AE-124 clonal expansion timeline



Patient AE-129

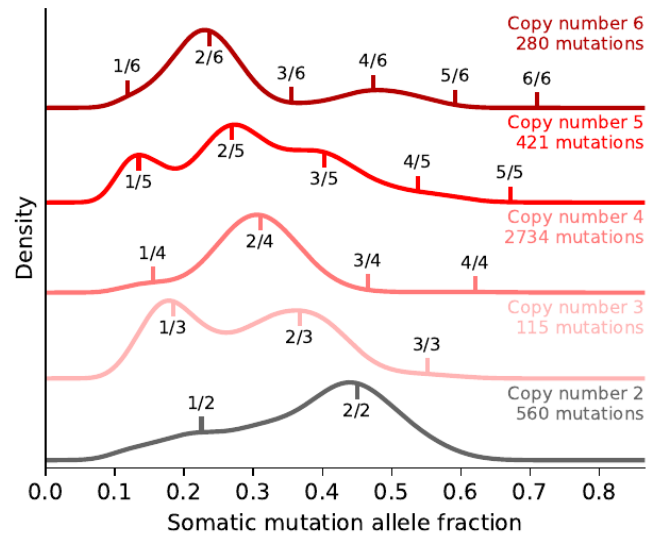
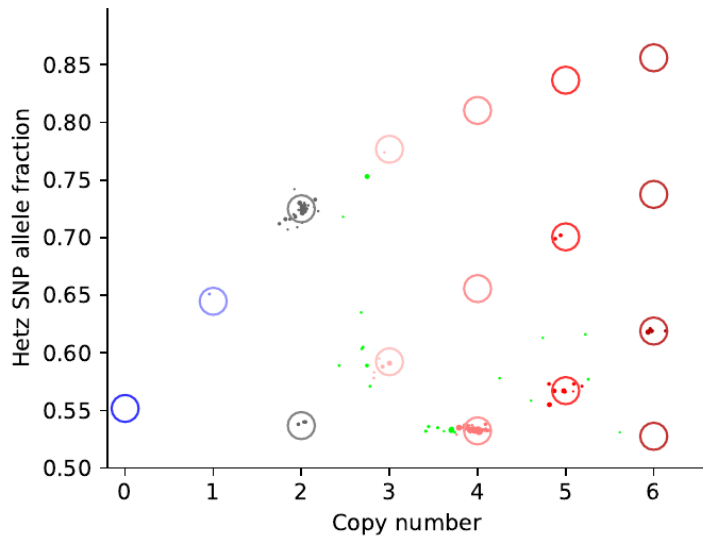
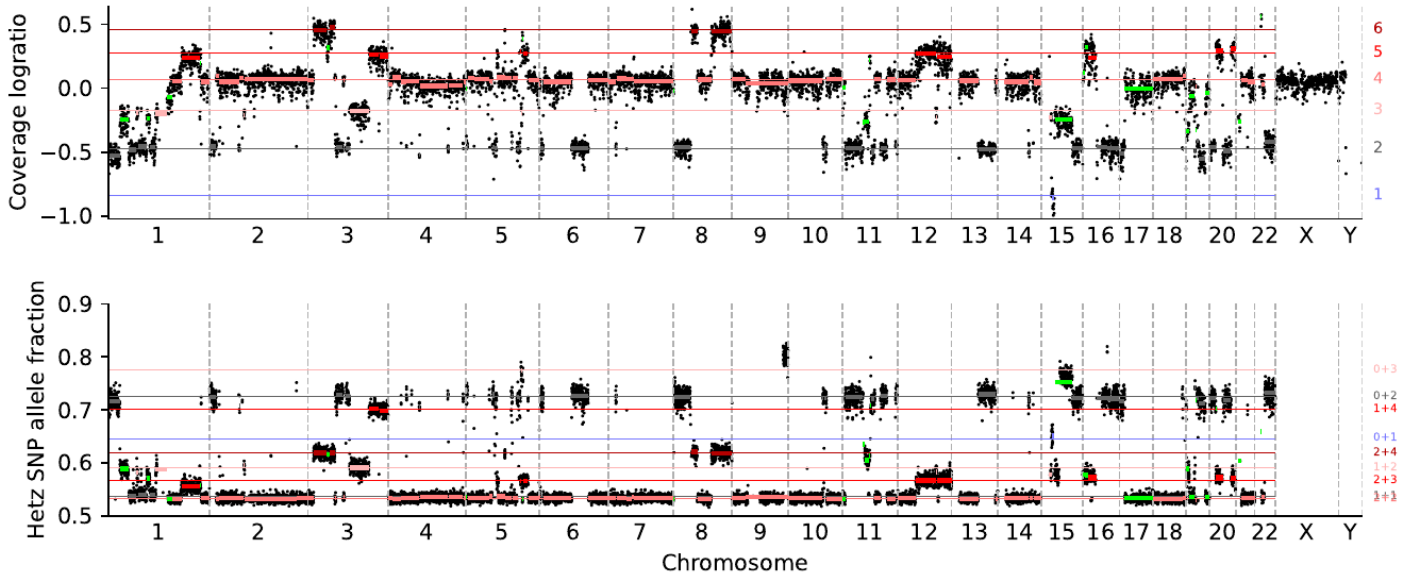
AE-129-Baseline-cfDNA

Cancer fraction: 28%, Diploid level: -0.30

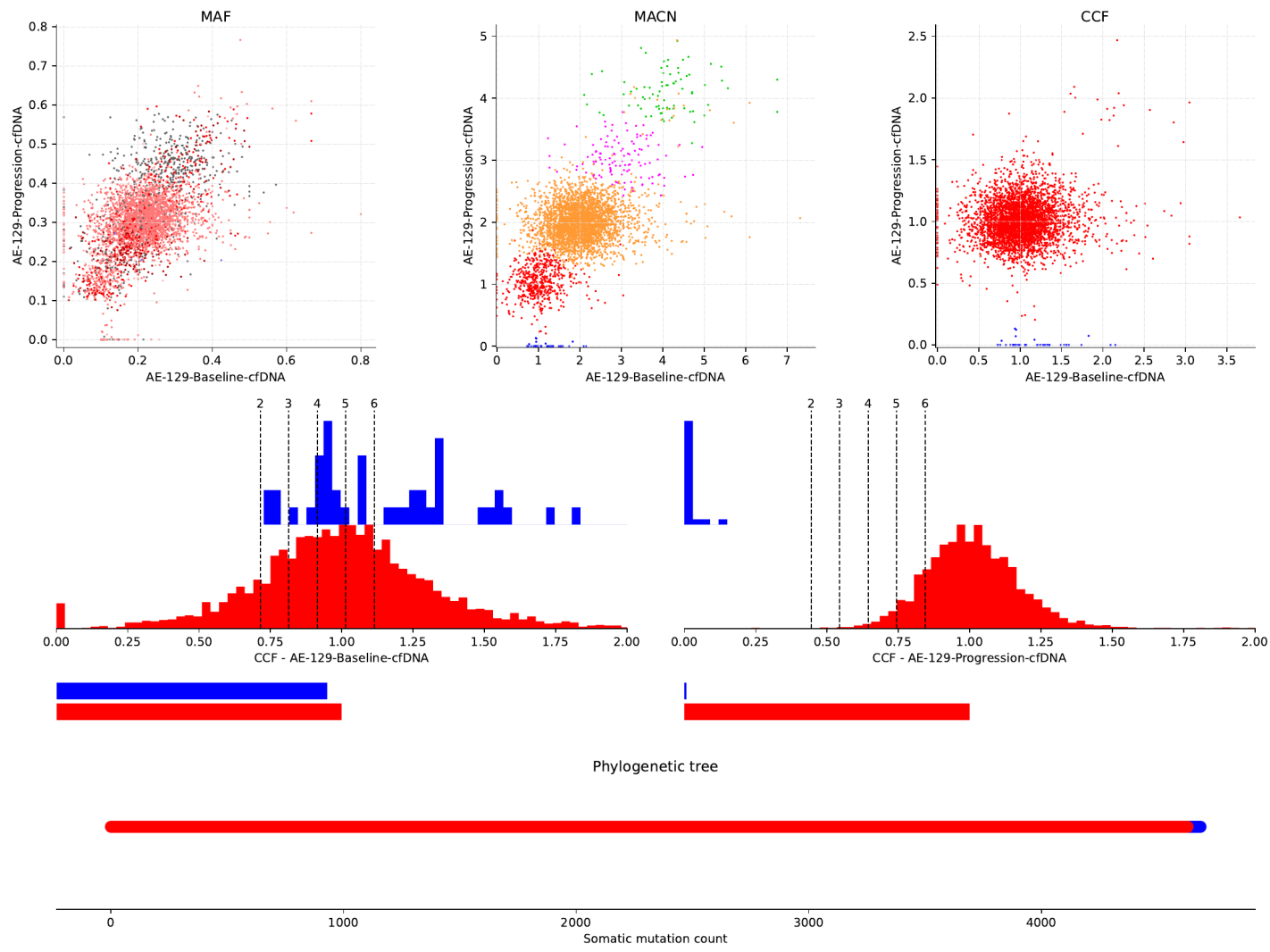


AE-129-Progression-cfDNA

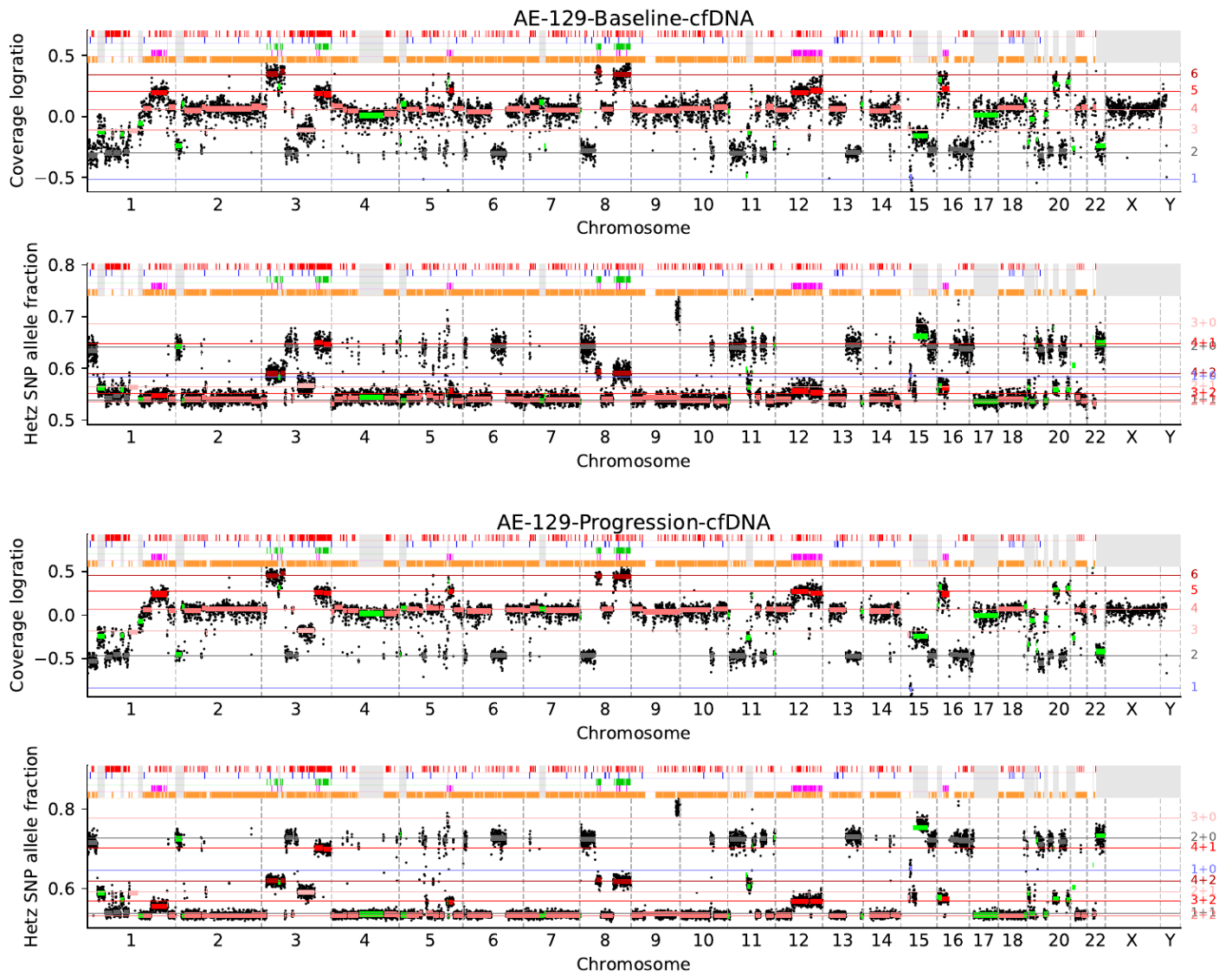
Cancer fraction: 45%, Diploid level: -0.47



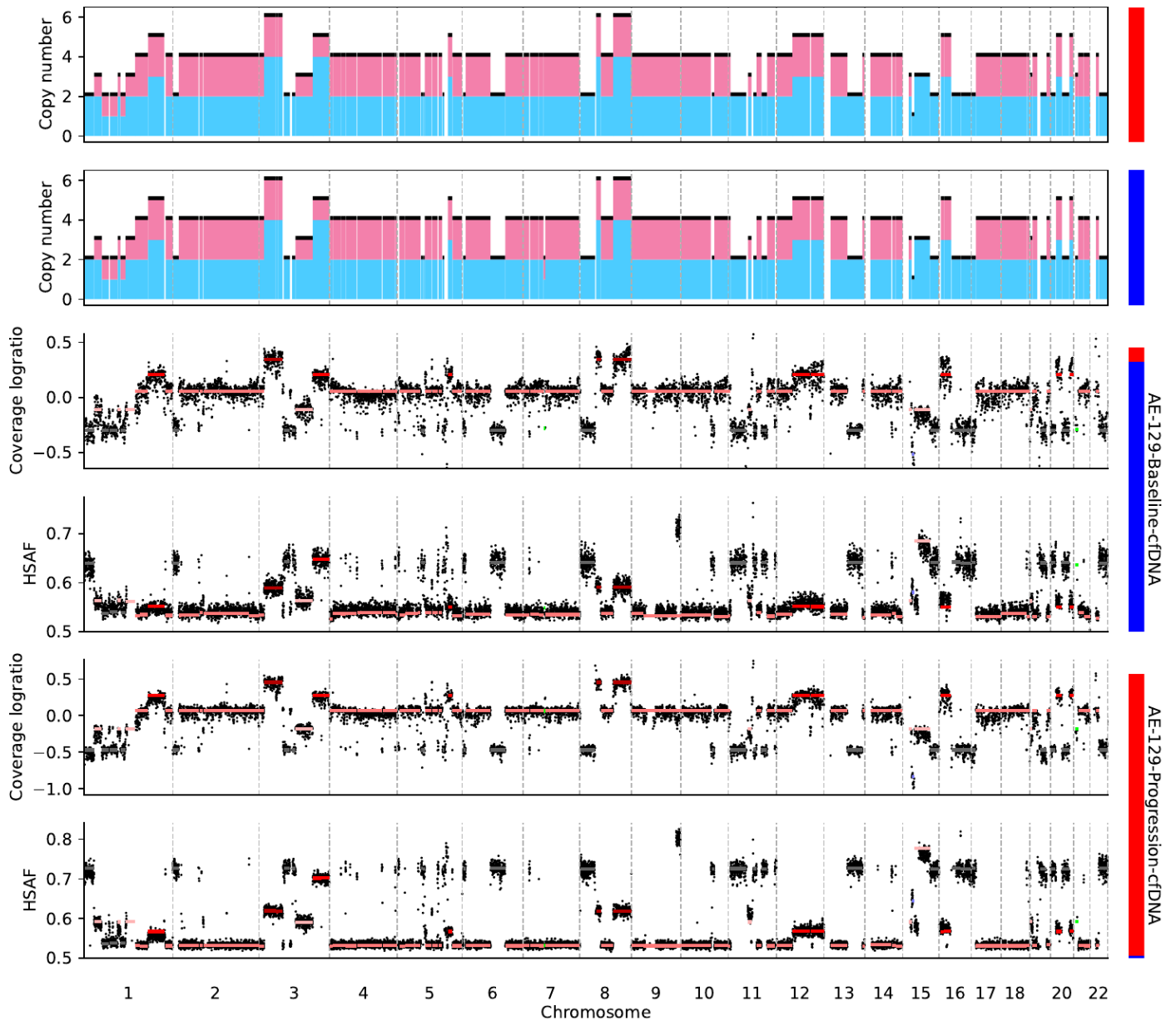
AE-129 mutation clustering



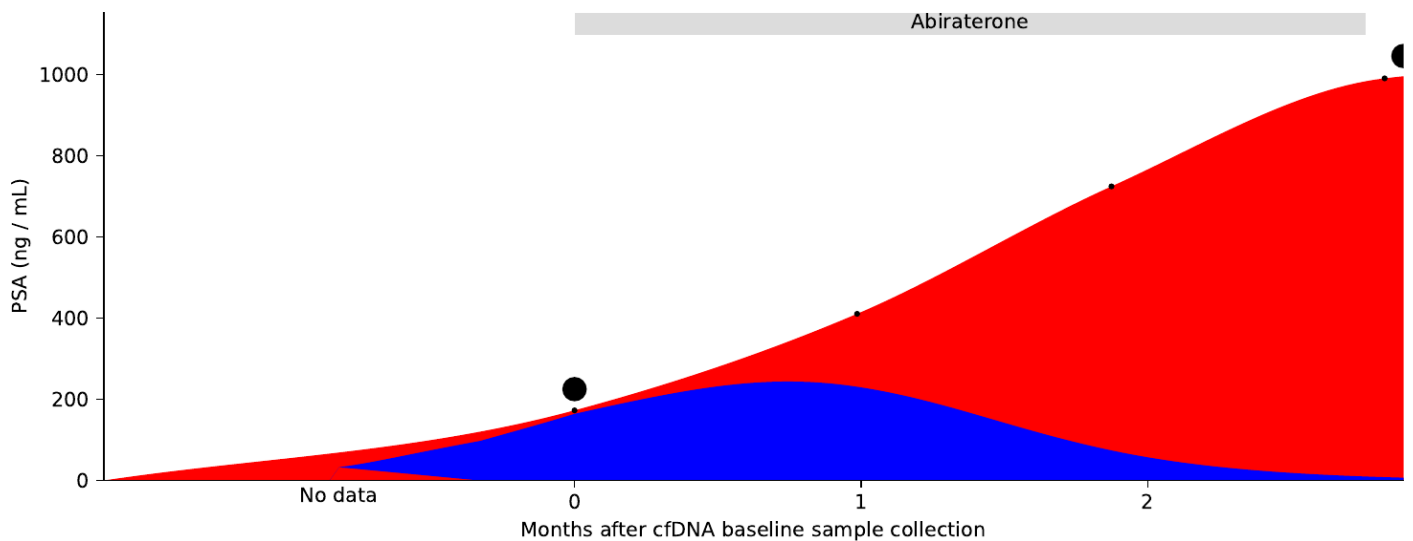
AE-129 mutation cluster positions



AE-129 copy number deconvolution



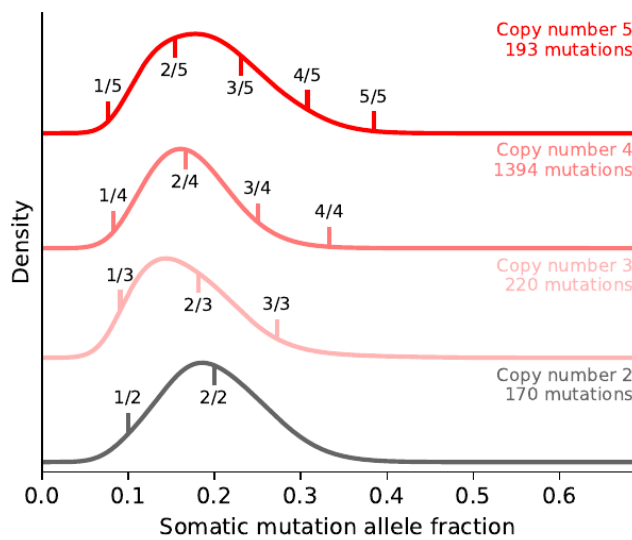
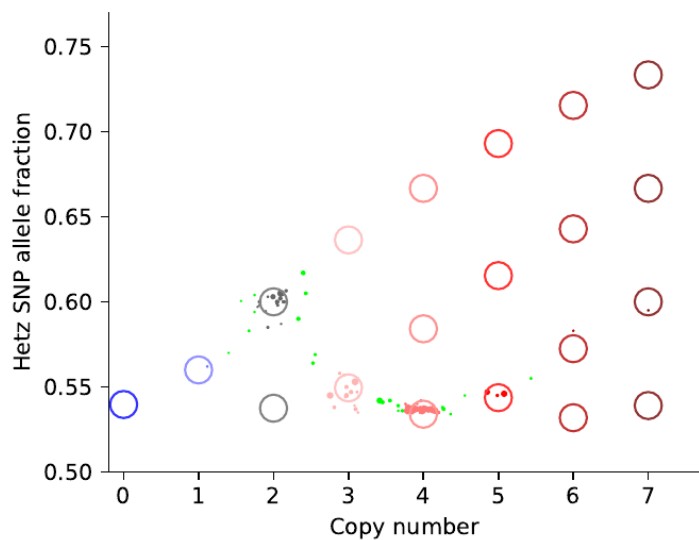
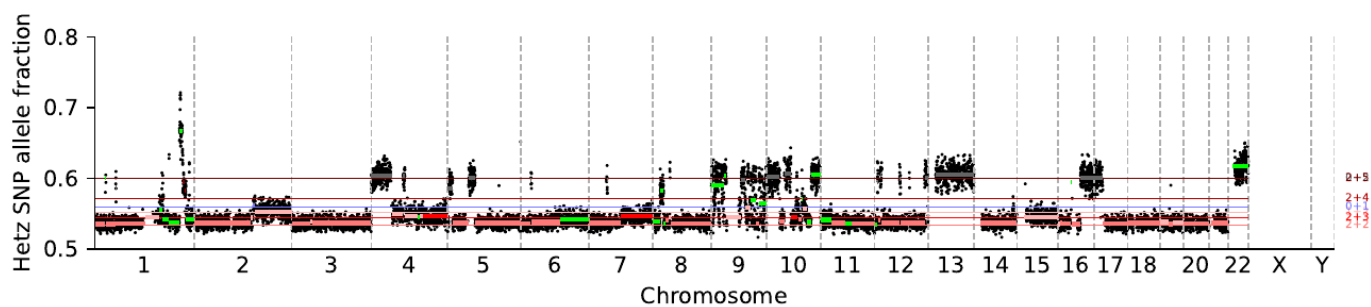
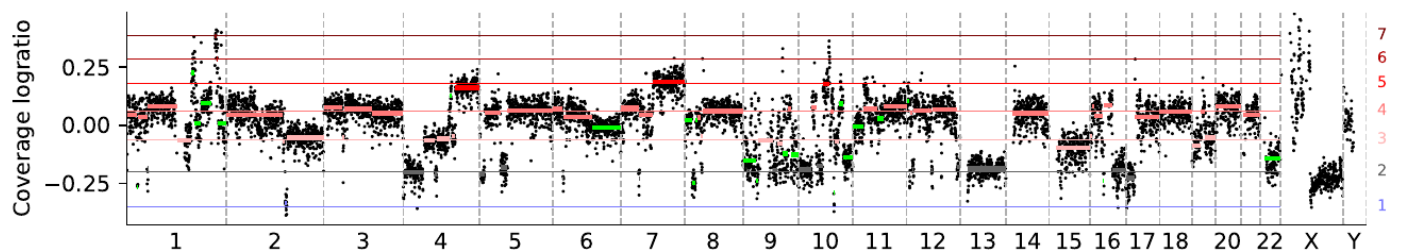
AE-129 clonal expansion timeline



Patient AE-132

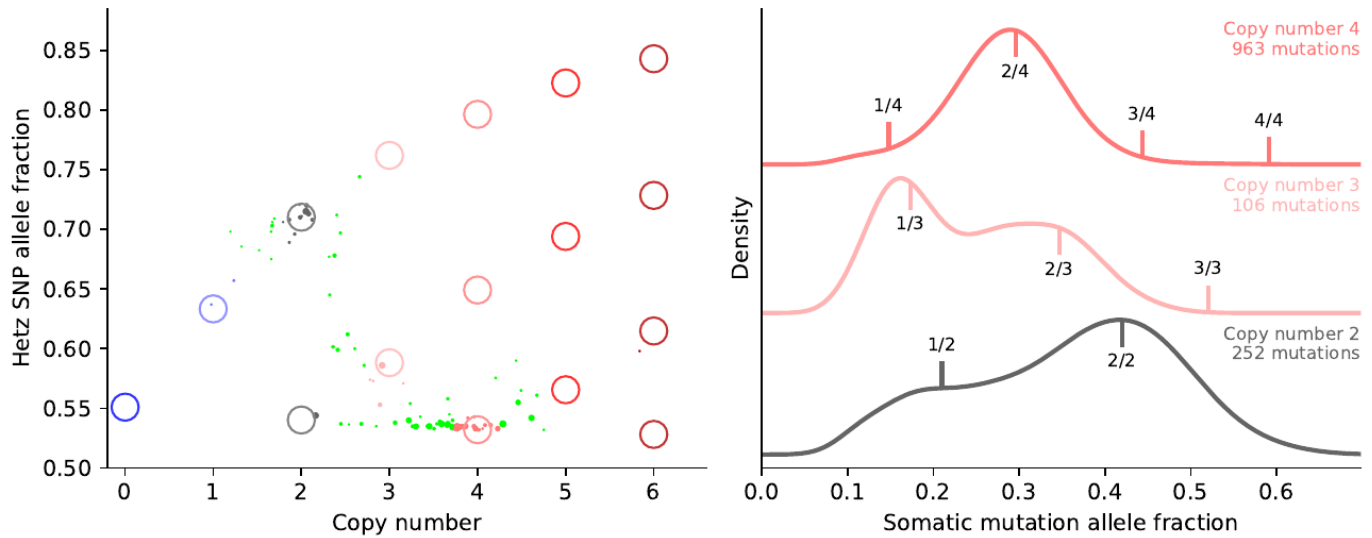
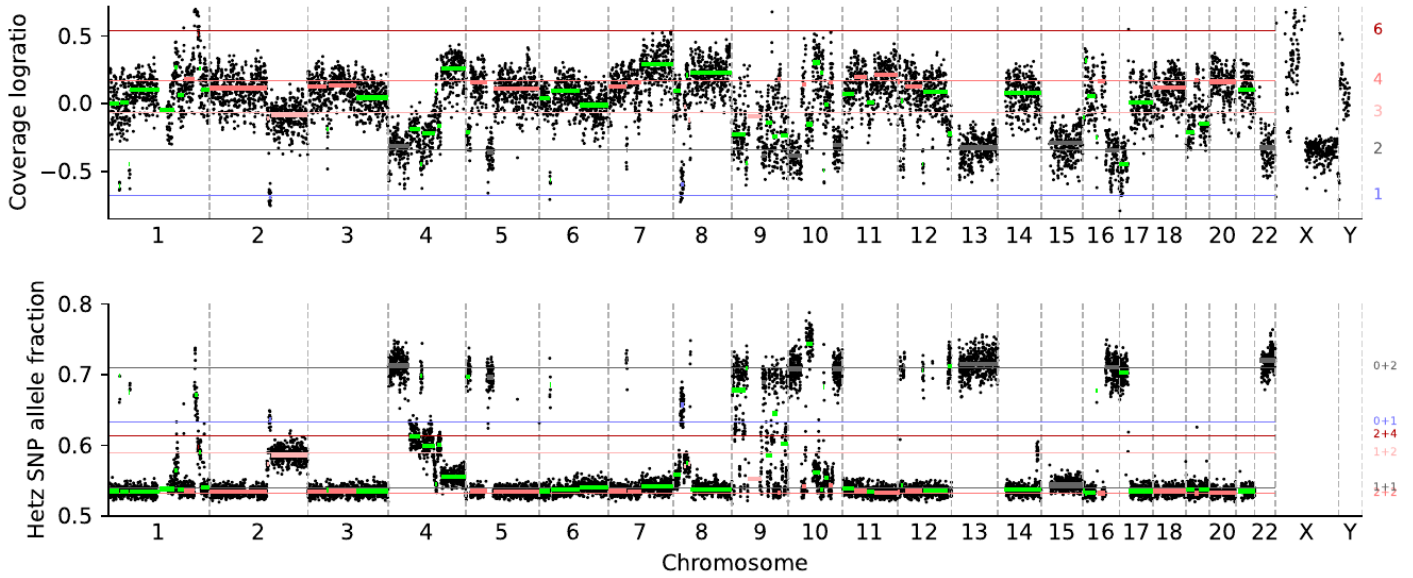
AE-132-Baseline-cfDNA

Cancer fraction: 20%, Diploid level: -0.20

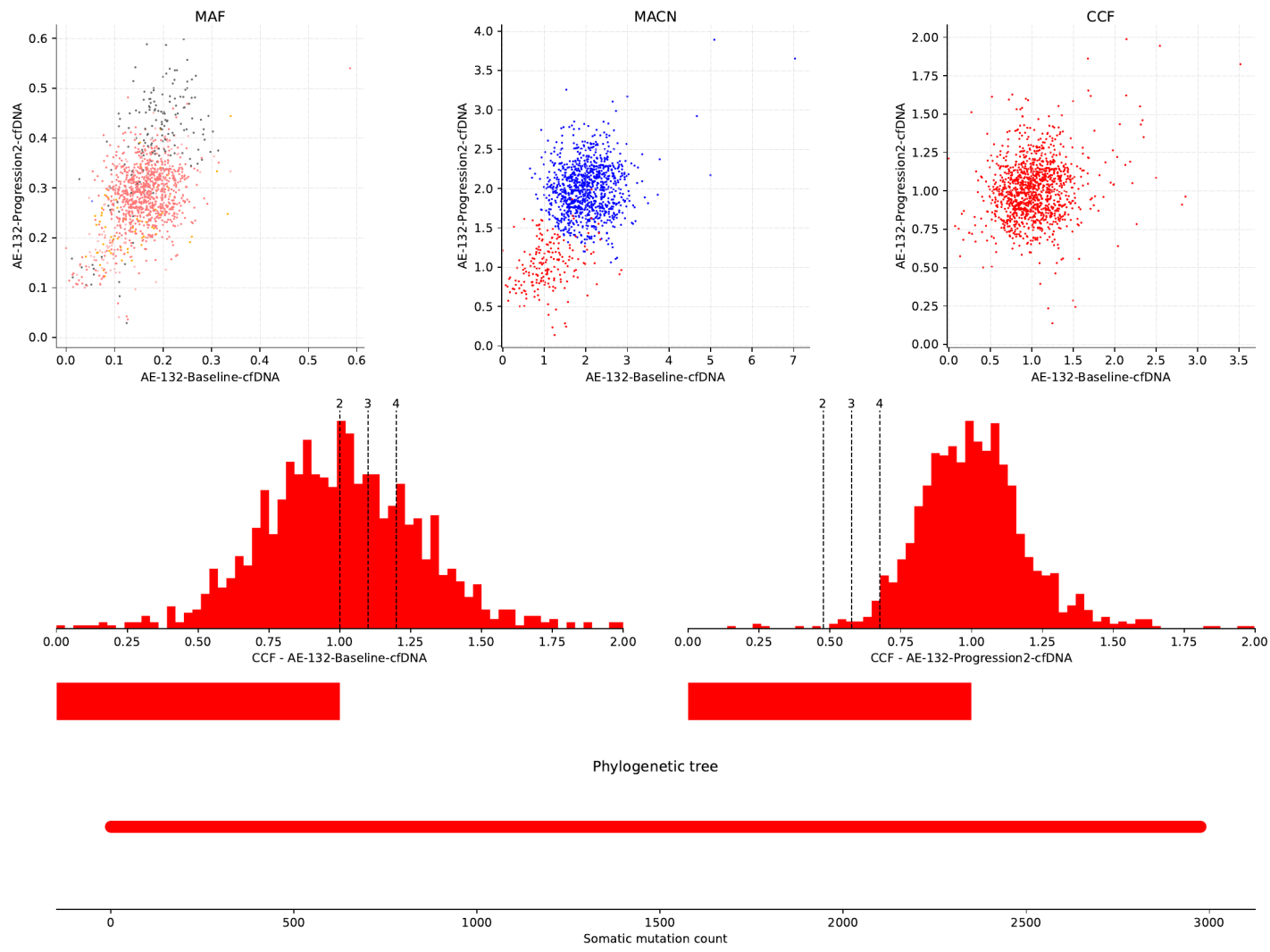


AE-132-Progression2-cfDNA

Cancer fraction: 42%, Diploid level: -0.34

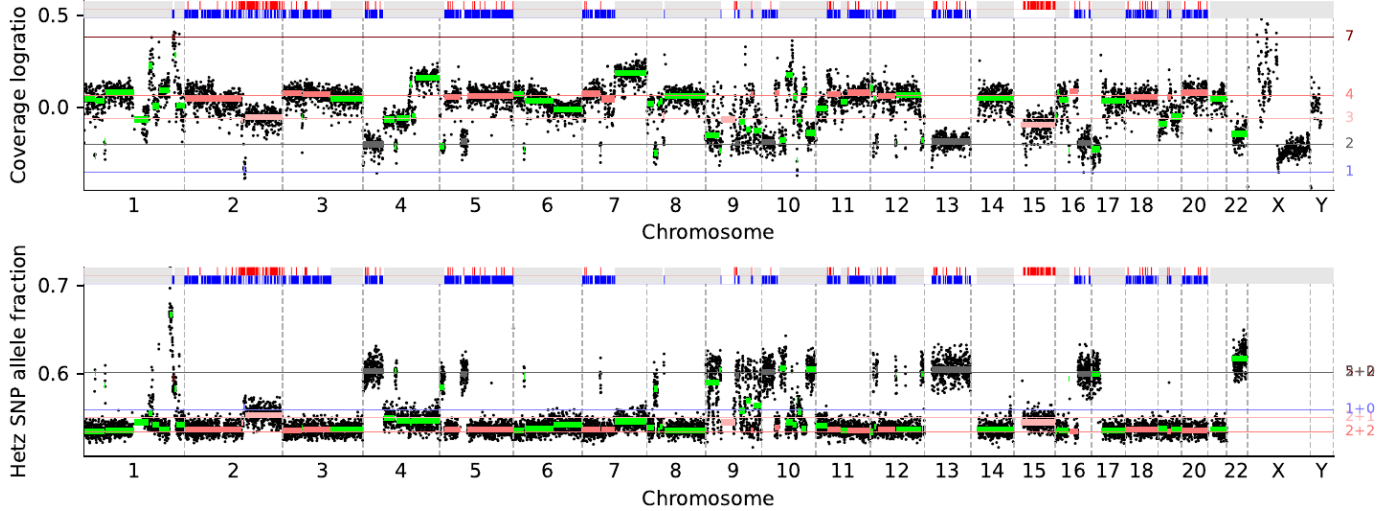


AE-132 mutation clustering

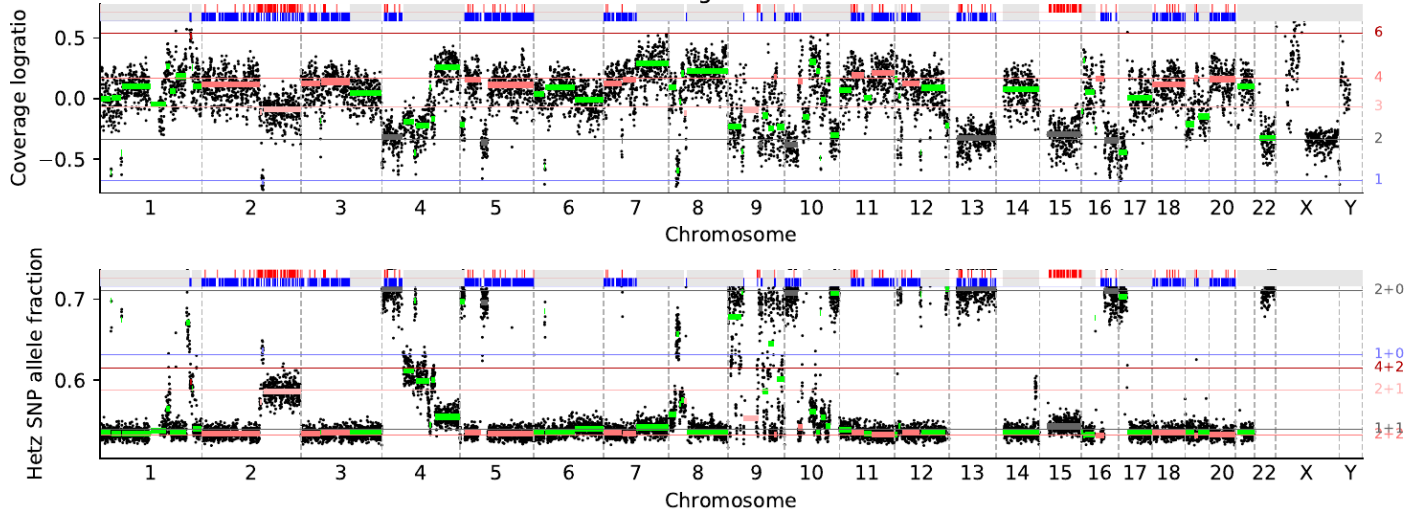


AE-132 mutation cluster positions

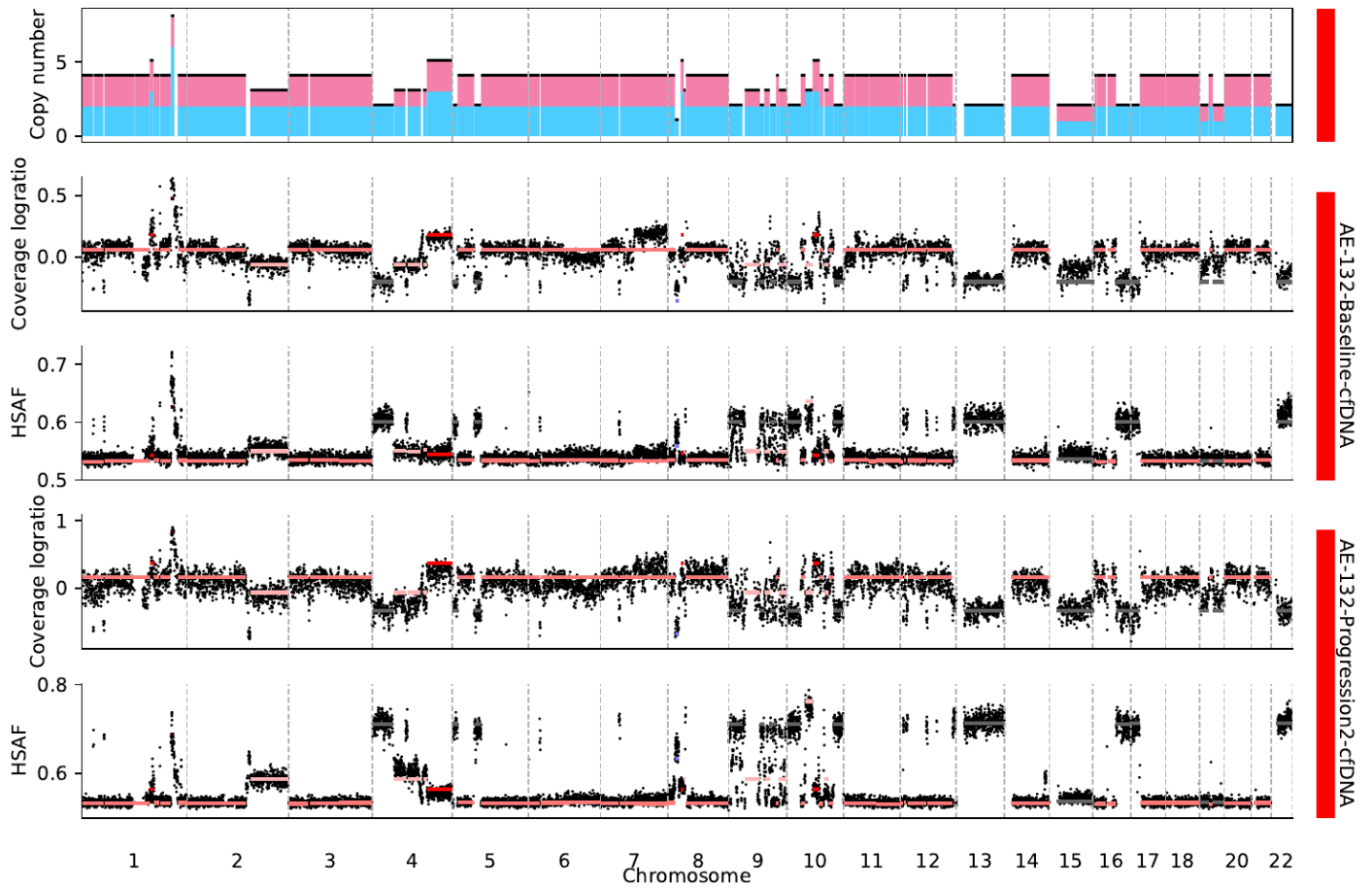
AE-132-Baseline-cfDNA



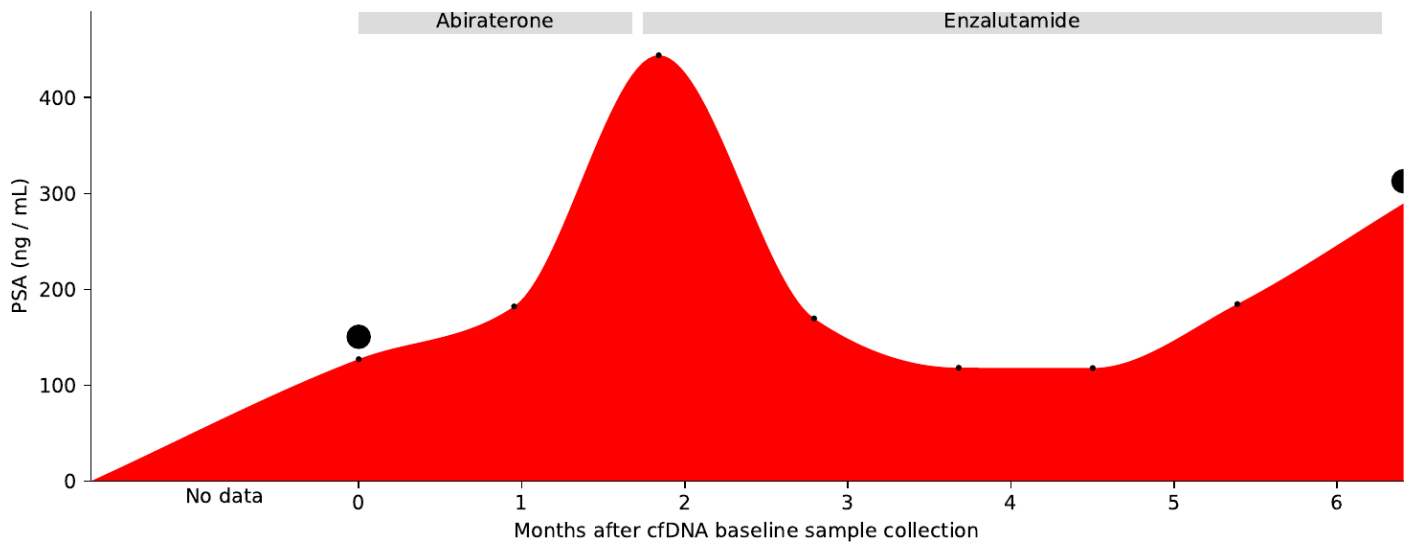
AE-132-Progression2-cfDNA



AE-132 copy number deconvolution



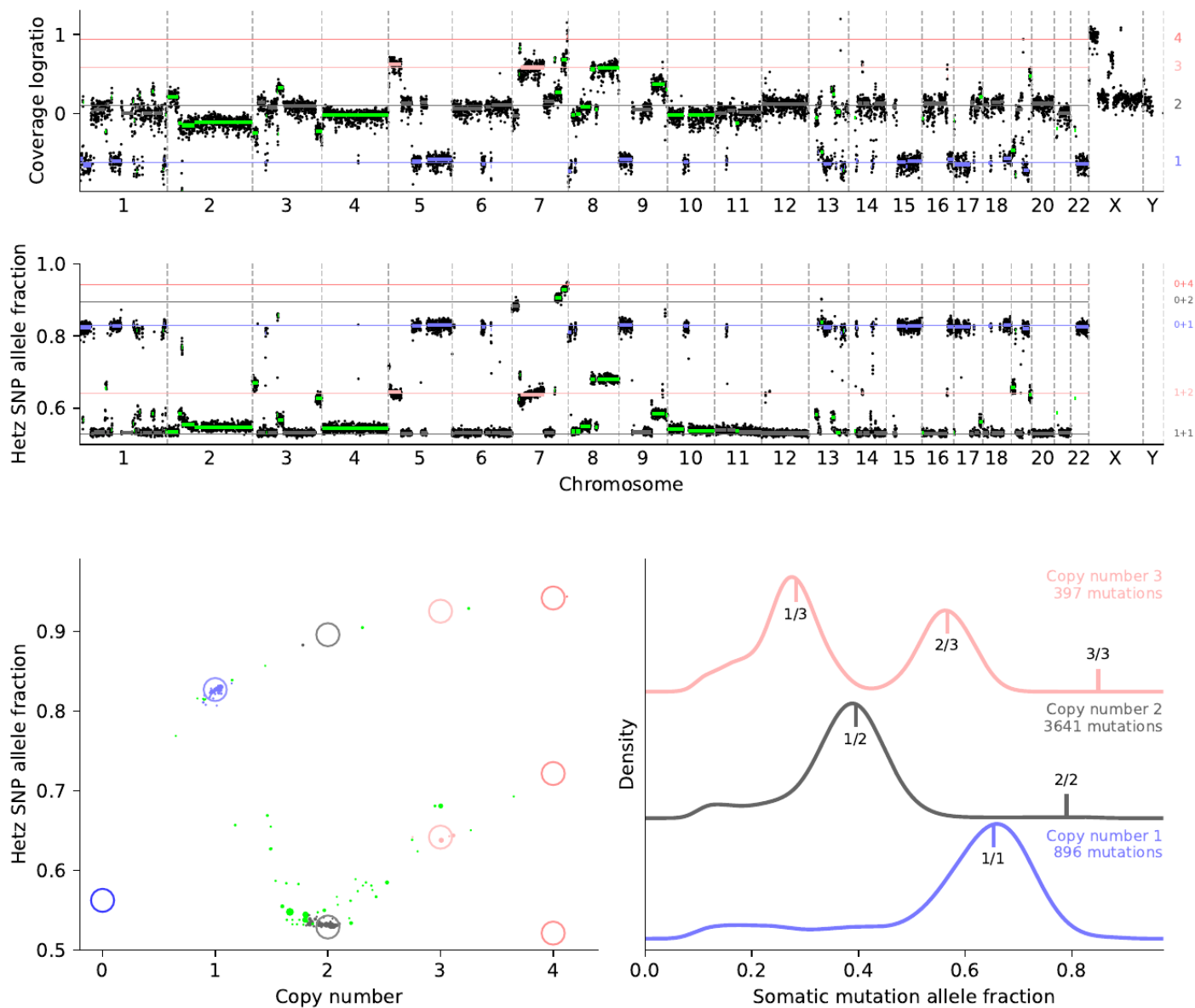
AE-132 clonal expansion timeline



Patient AE-136

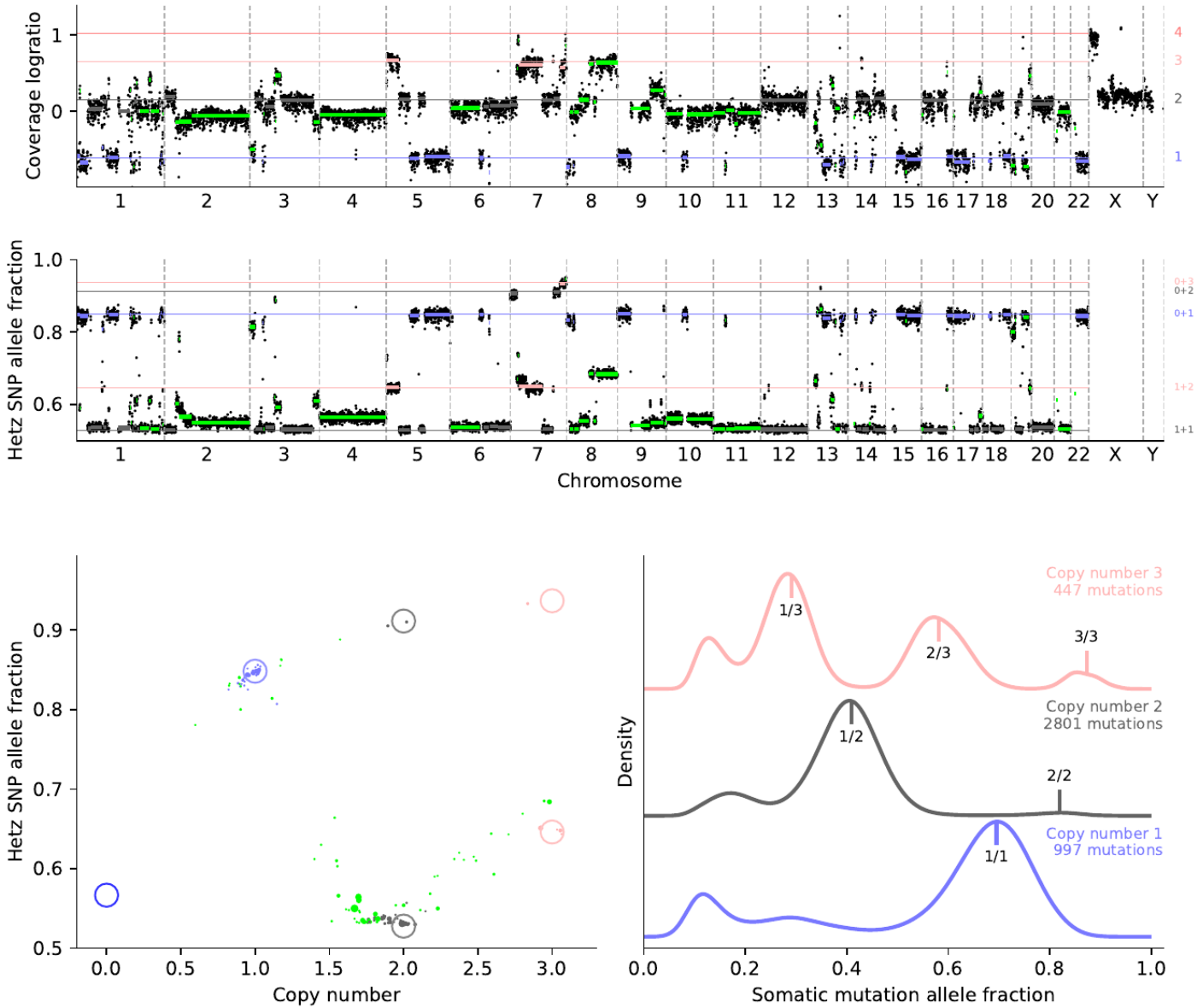
AE-136-Baseline-cfDNA

Cancer fraction: 79%, Diploid level: 0.10

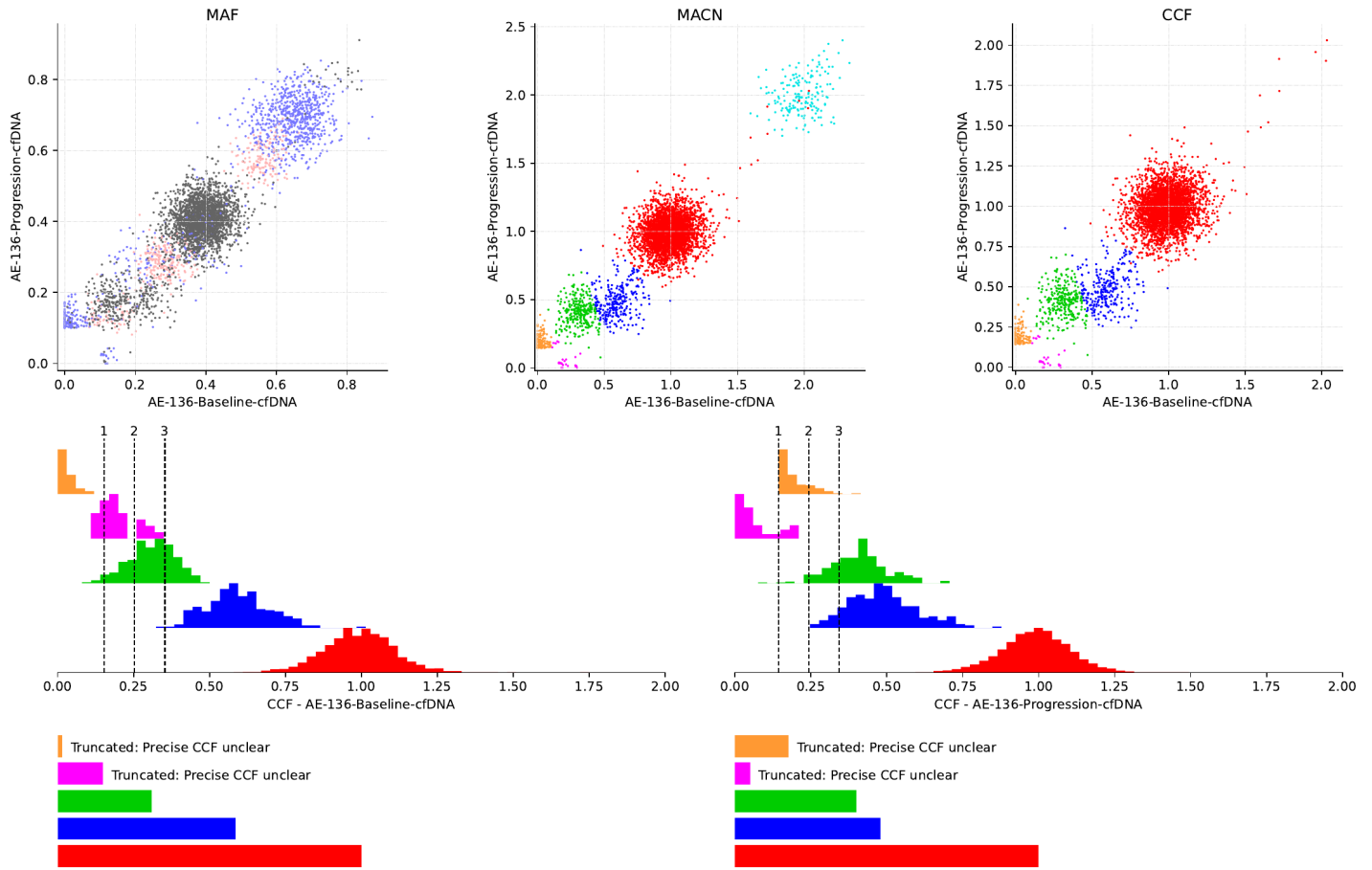


AE-136-Progression-cfDNA

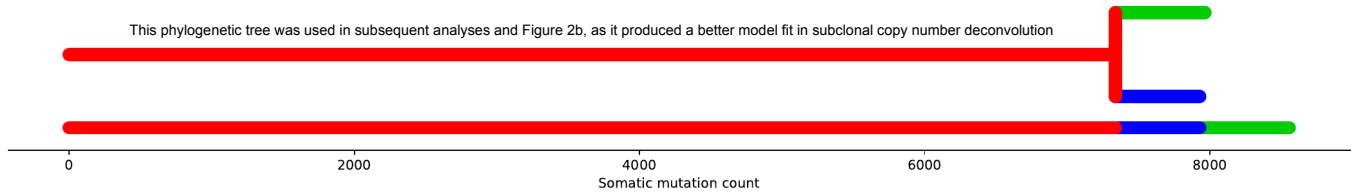
Cancer fraction: 82%, Diploid level: 0.15



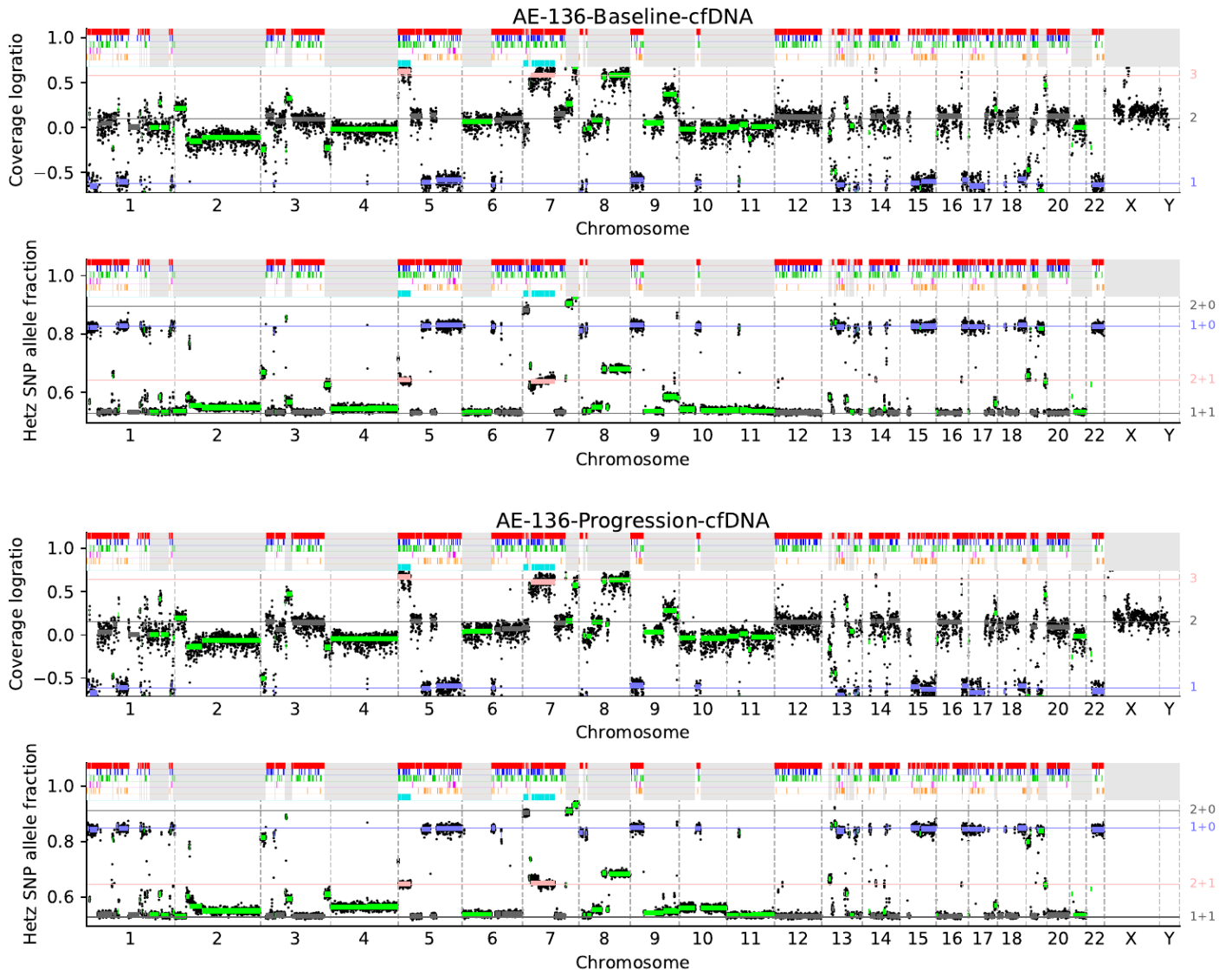
AE-136 mutation clustering



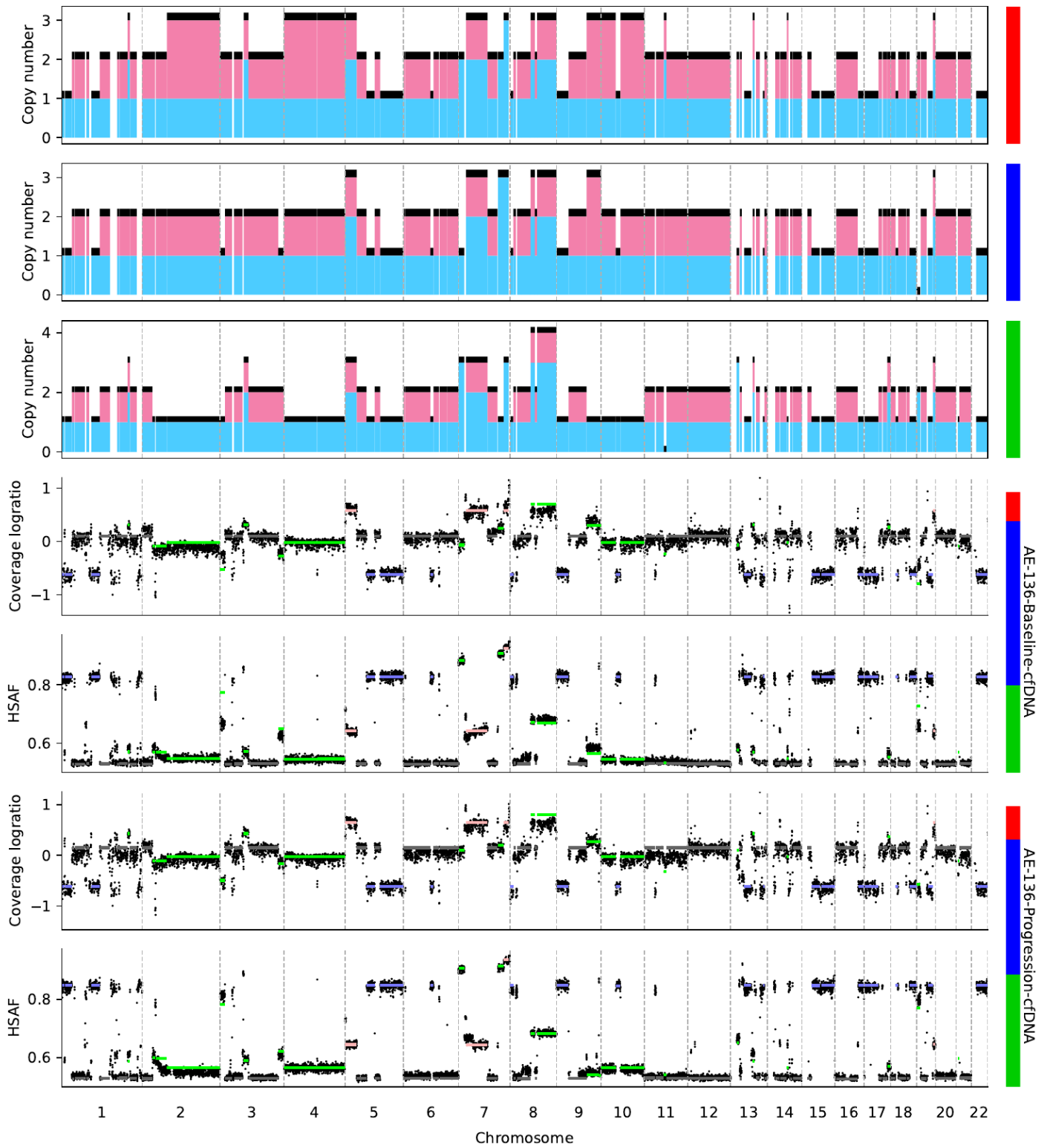
Possible phylogenetic trees



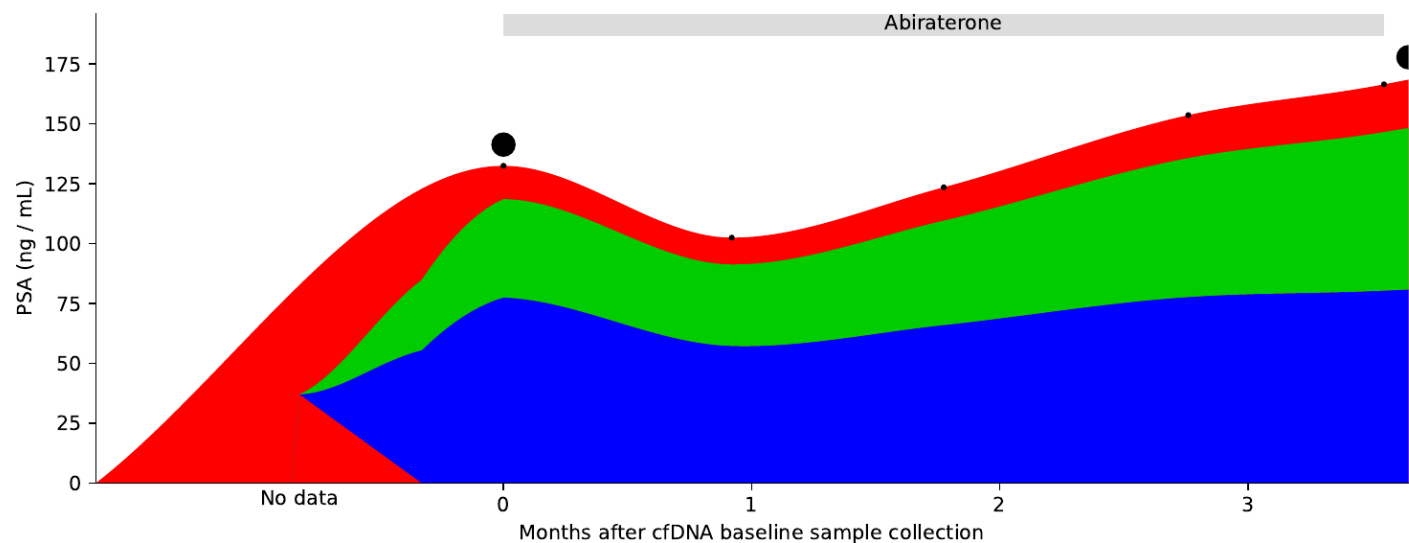
AE-136 mutation cluster positions



AE-136 copy number deconvolution



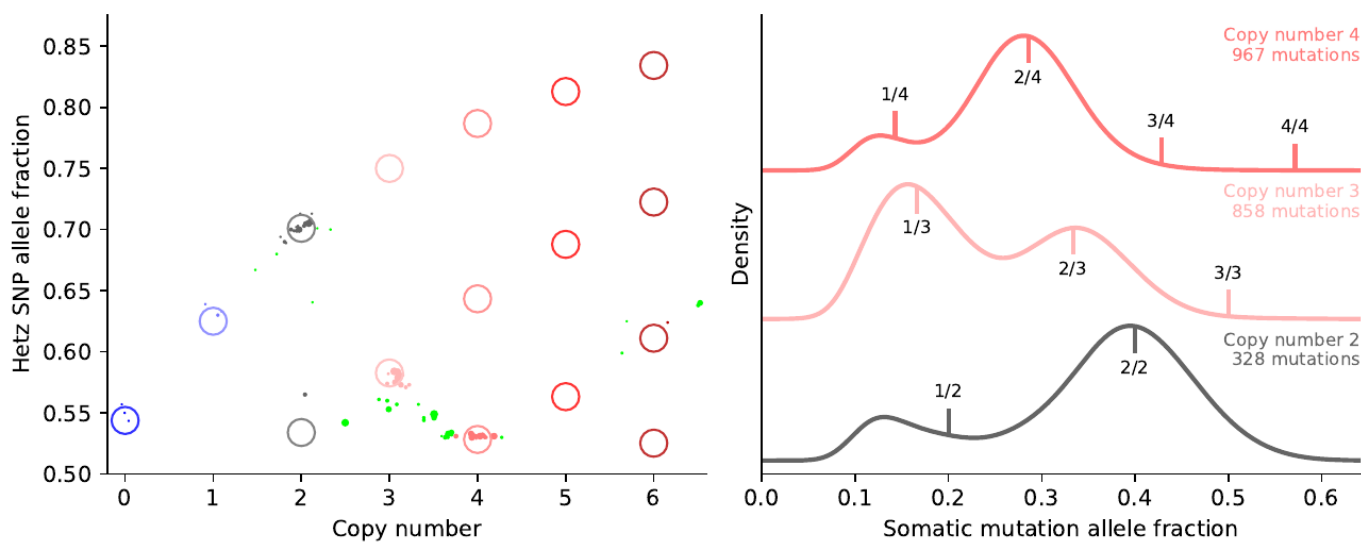
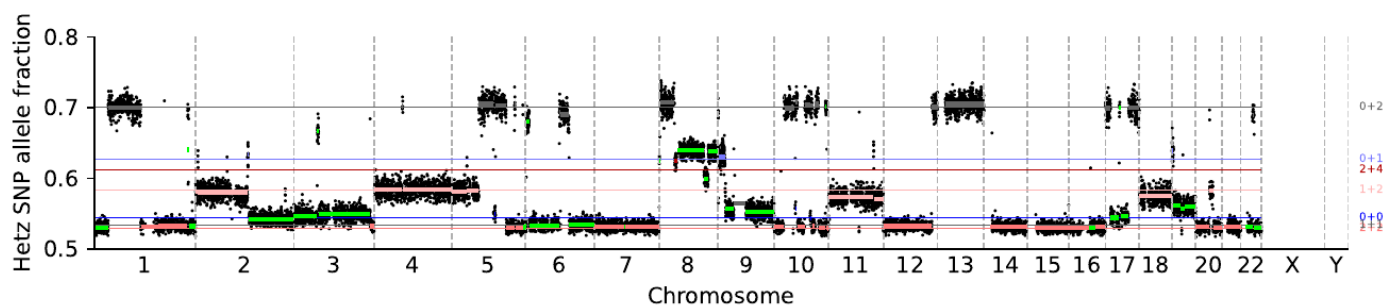
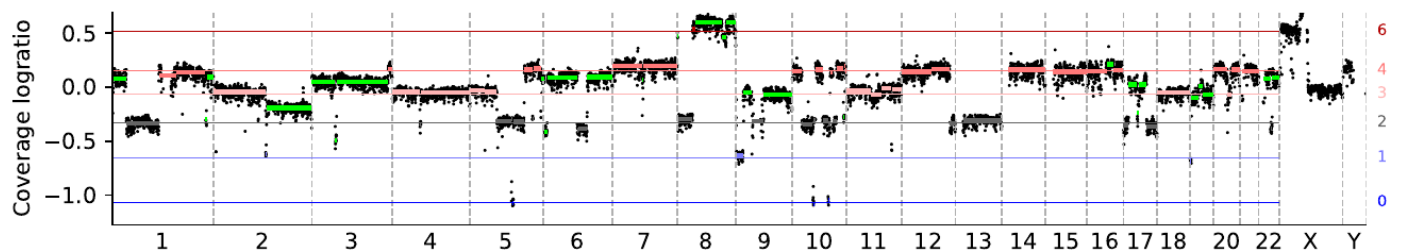
AE-136 clonal expansion timeline



Patient AE-151

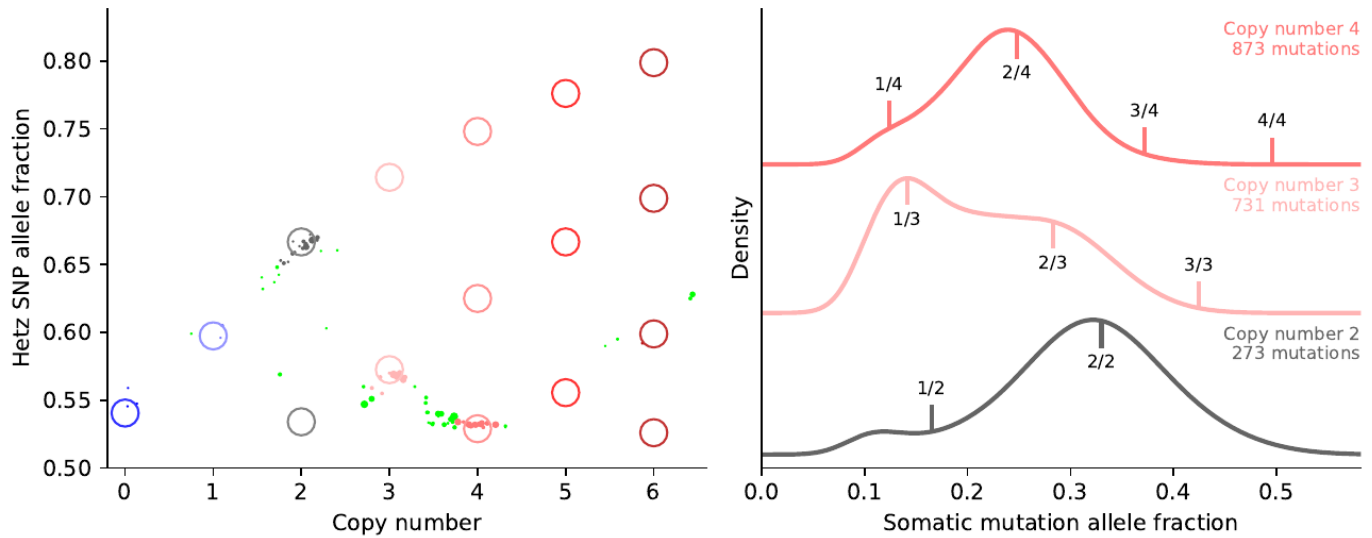
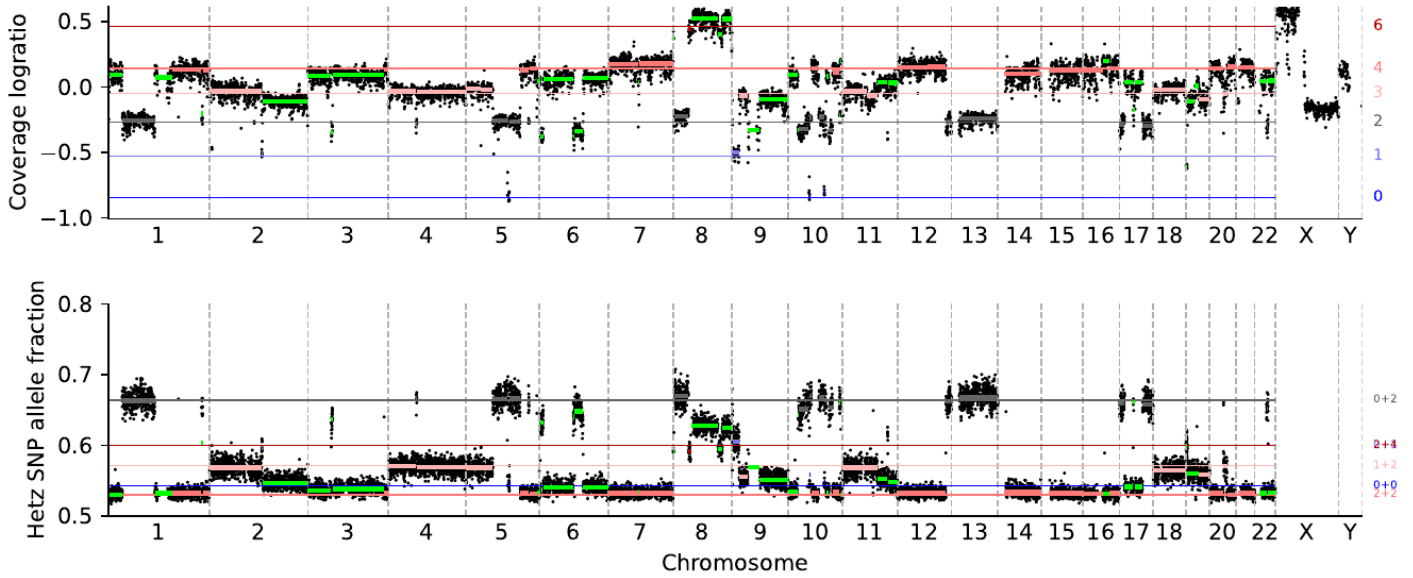
AE-151-Baseline-cfDNA

Cancer fraction: 40%, Diploid level: -0.33



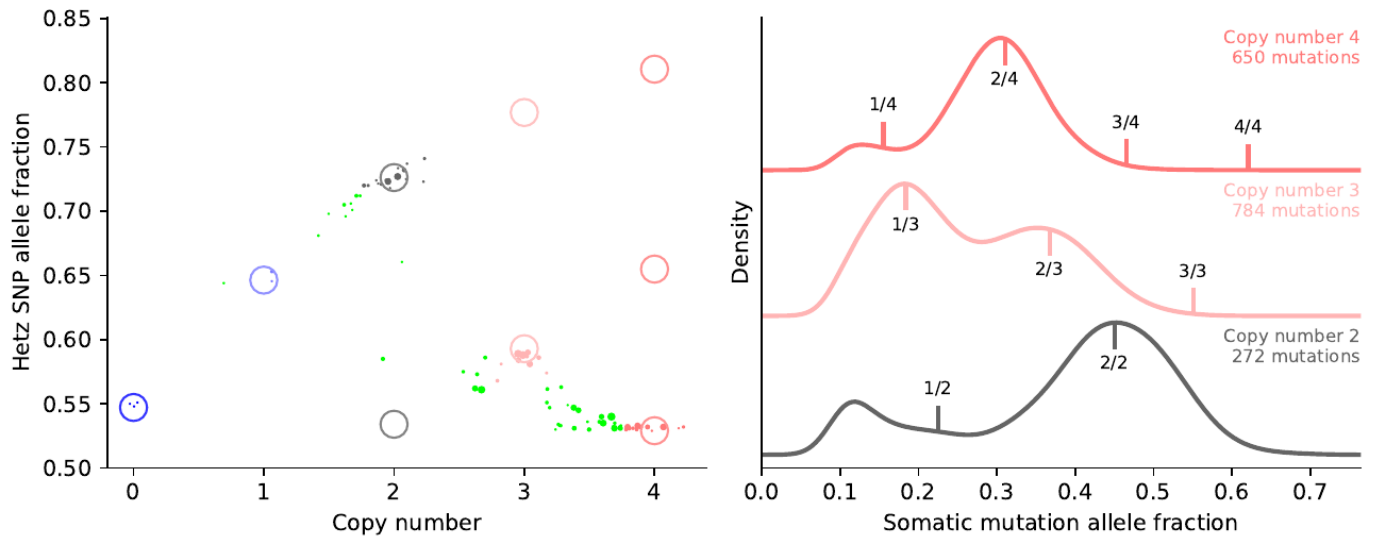
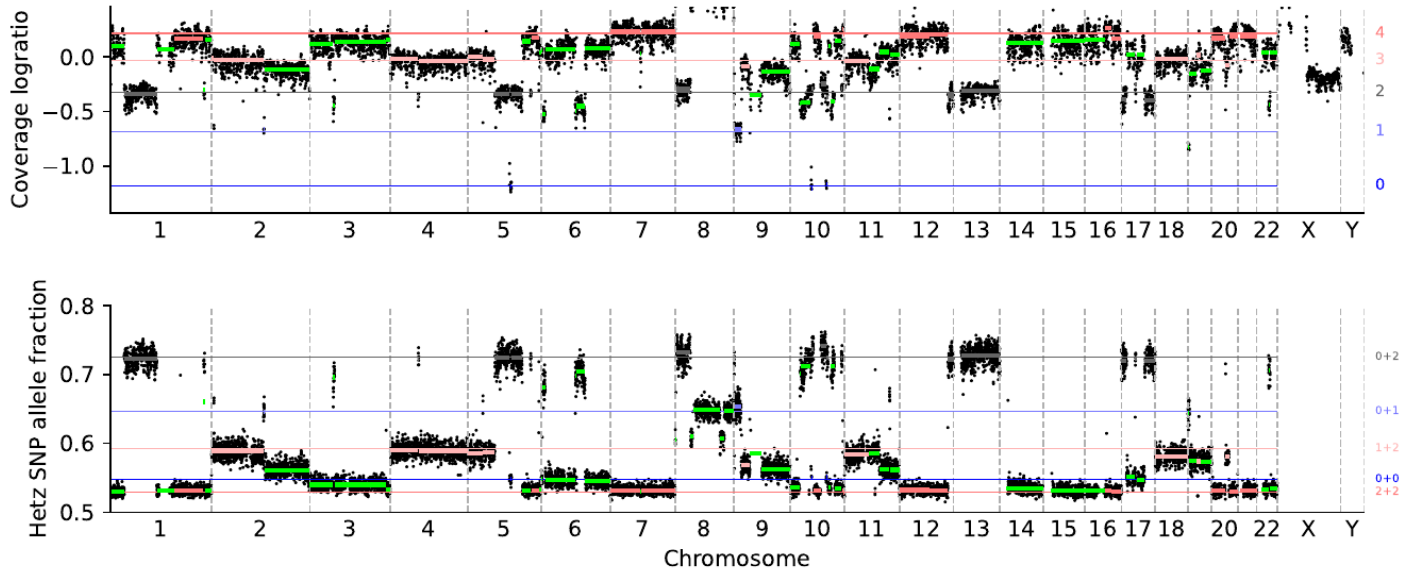
AE-151-Progression-cfDNA

Cancer fraction: 33%, Diploid level: -0.27

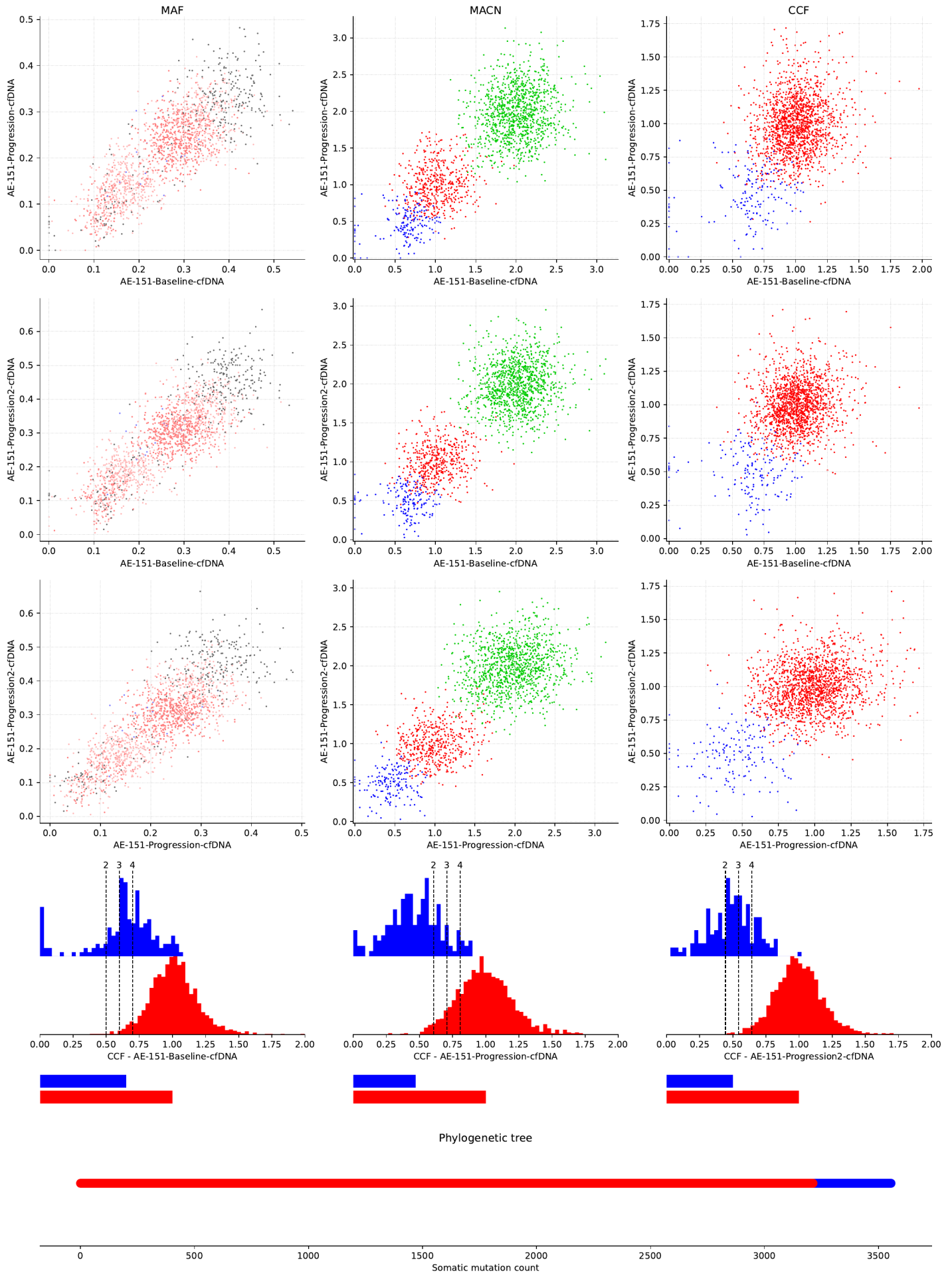


AE-151-Progression2-cfDNA

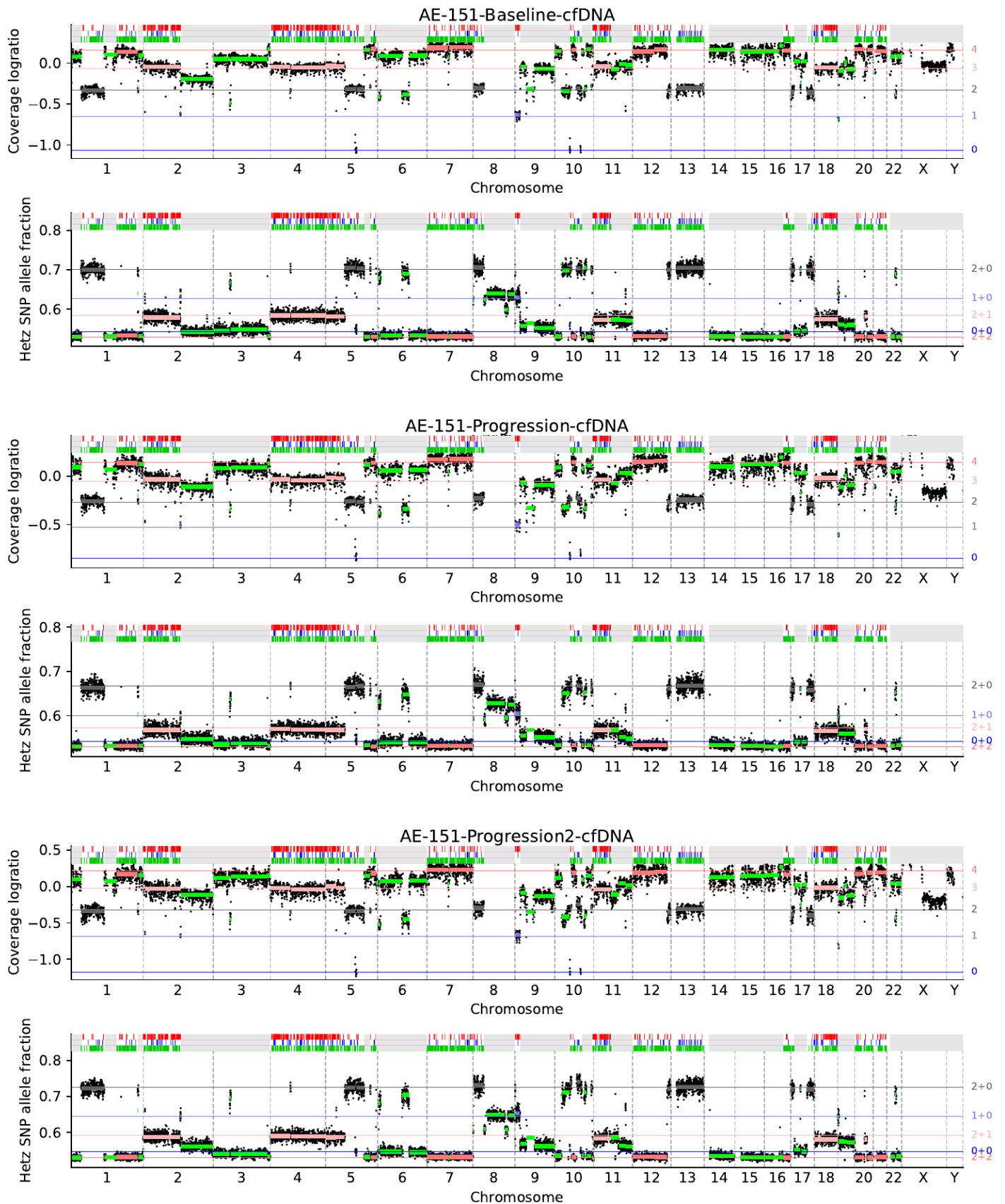
Cancer fraction: 45%, Diploid level: -0.32



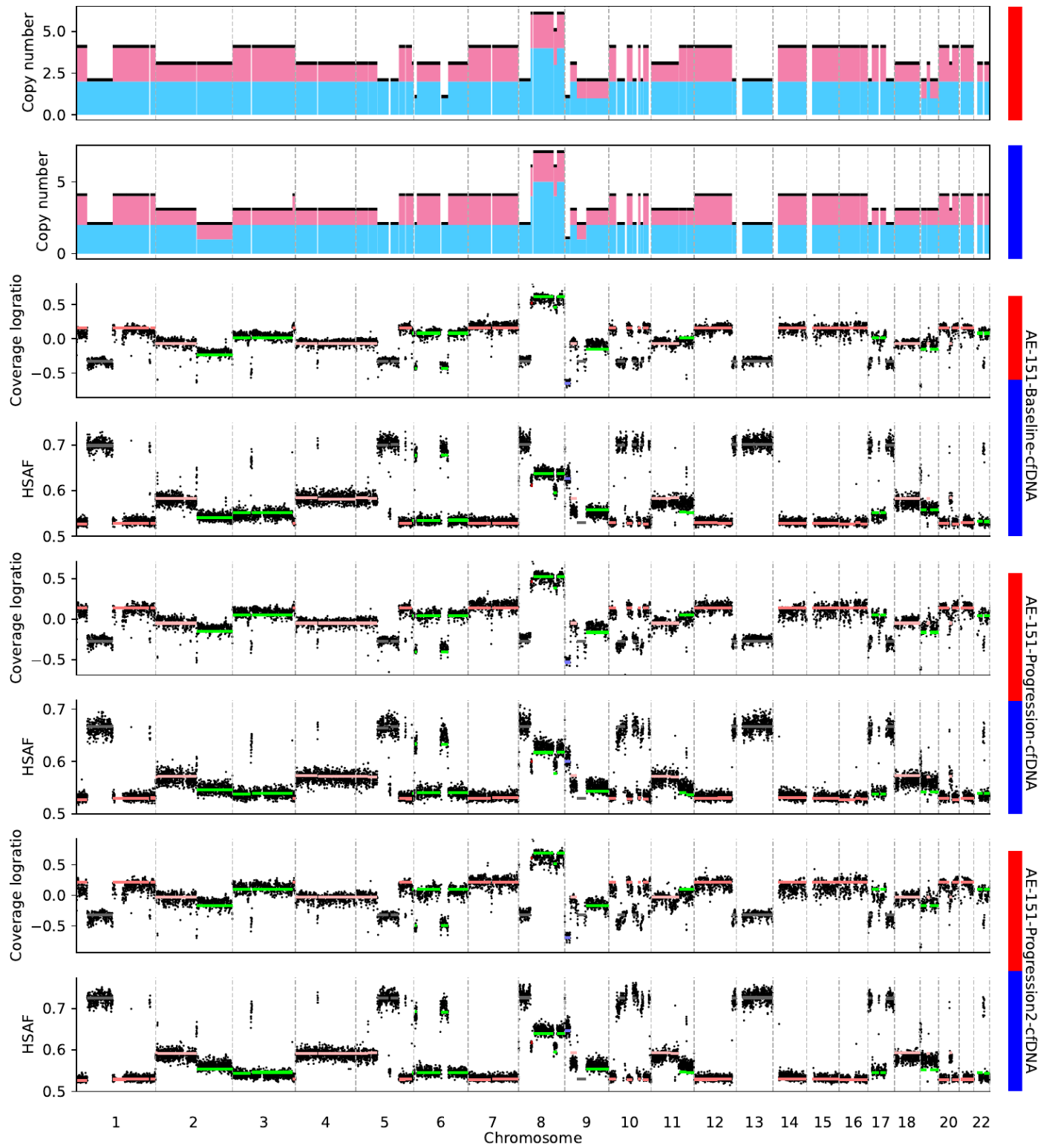
AE-151 mutation clustering



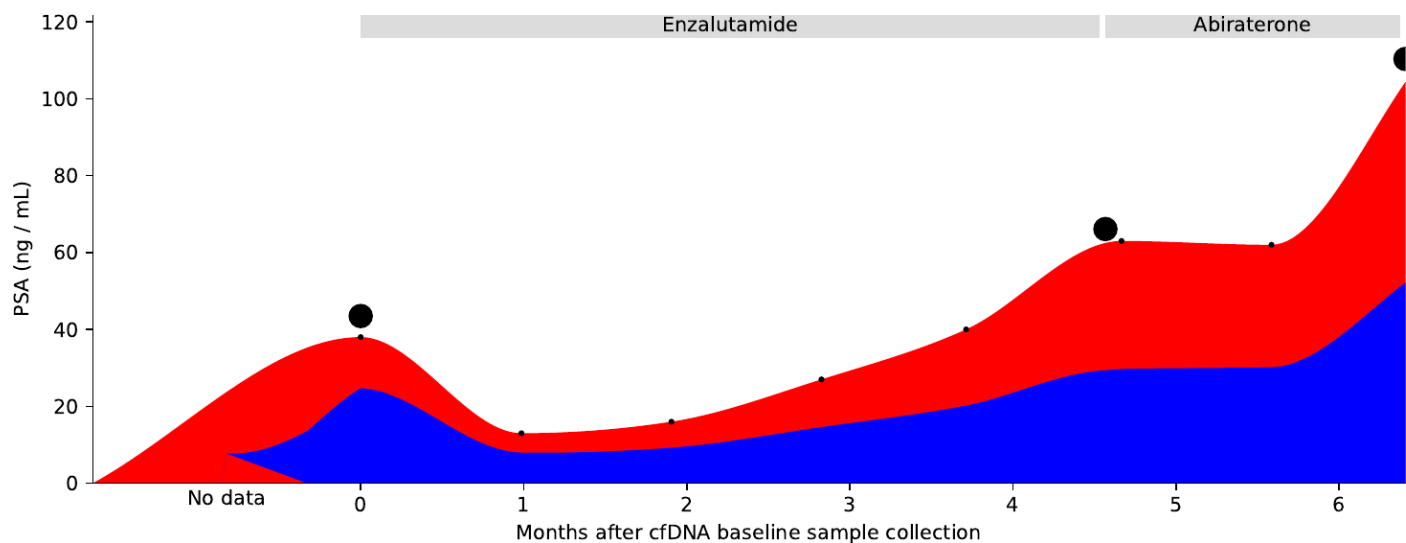
AE-151 mutation cluster positions



AE-151 copy number deconvolution



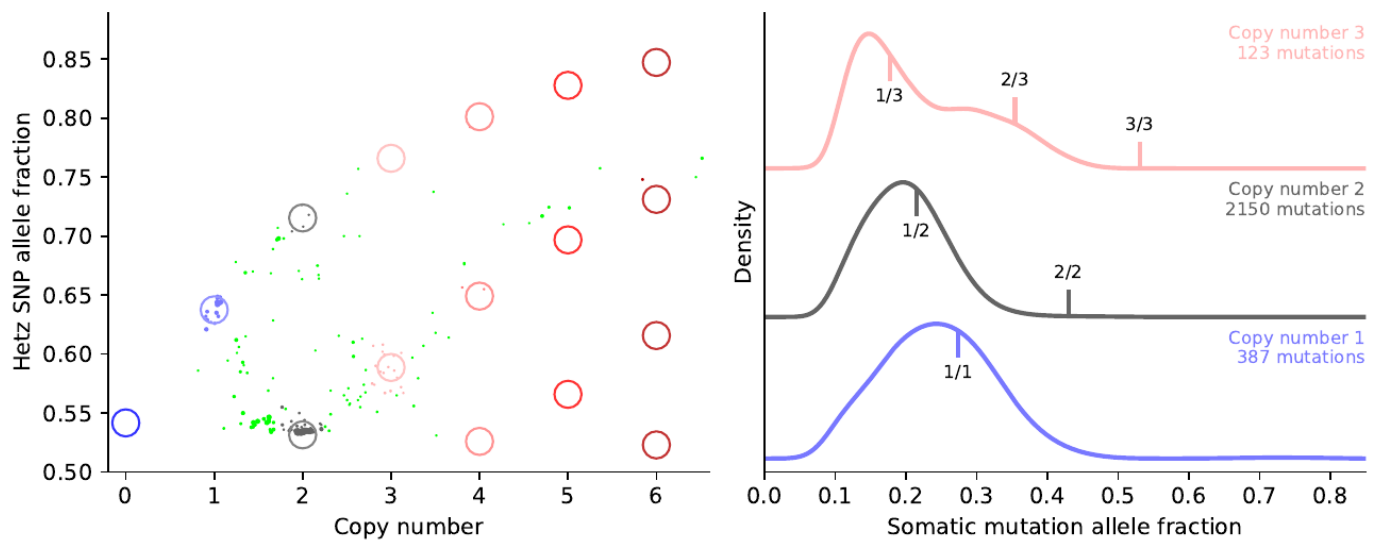
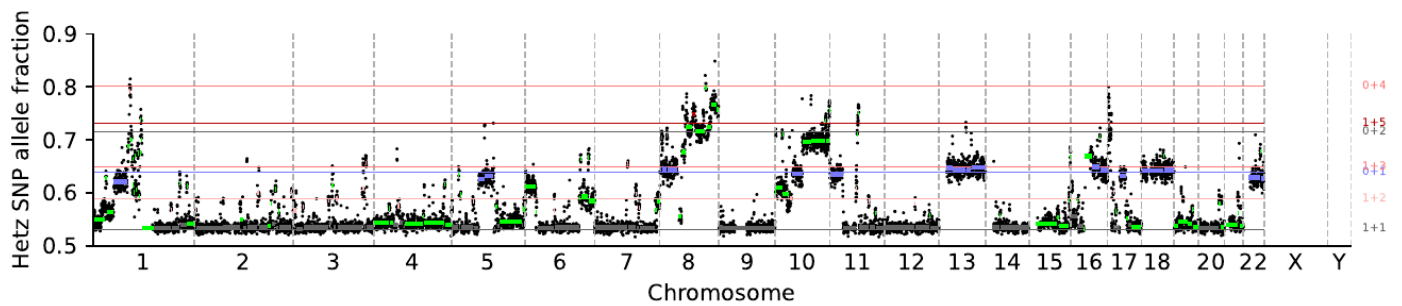
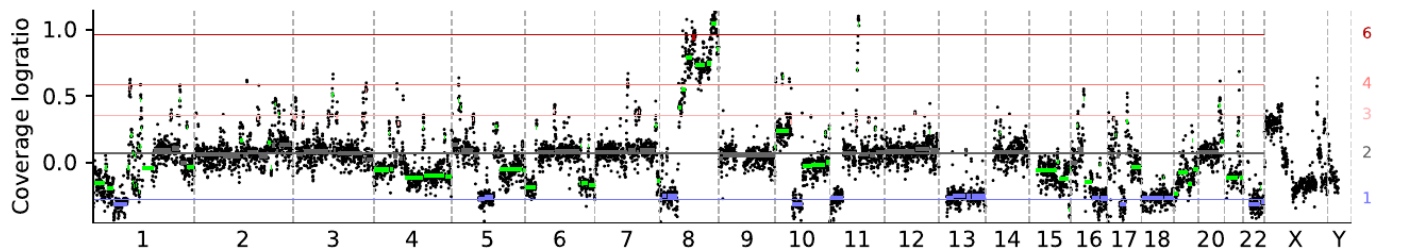
AE-151 clonal expansion timeline



Patient AE-156

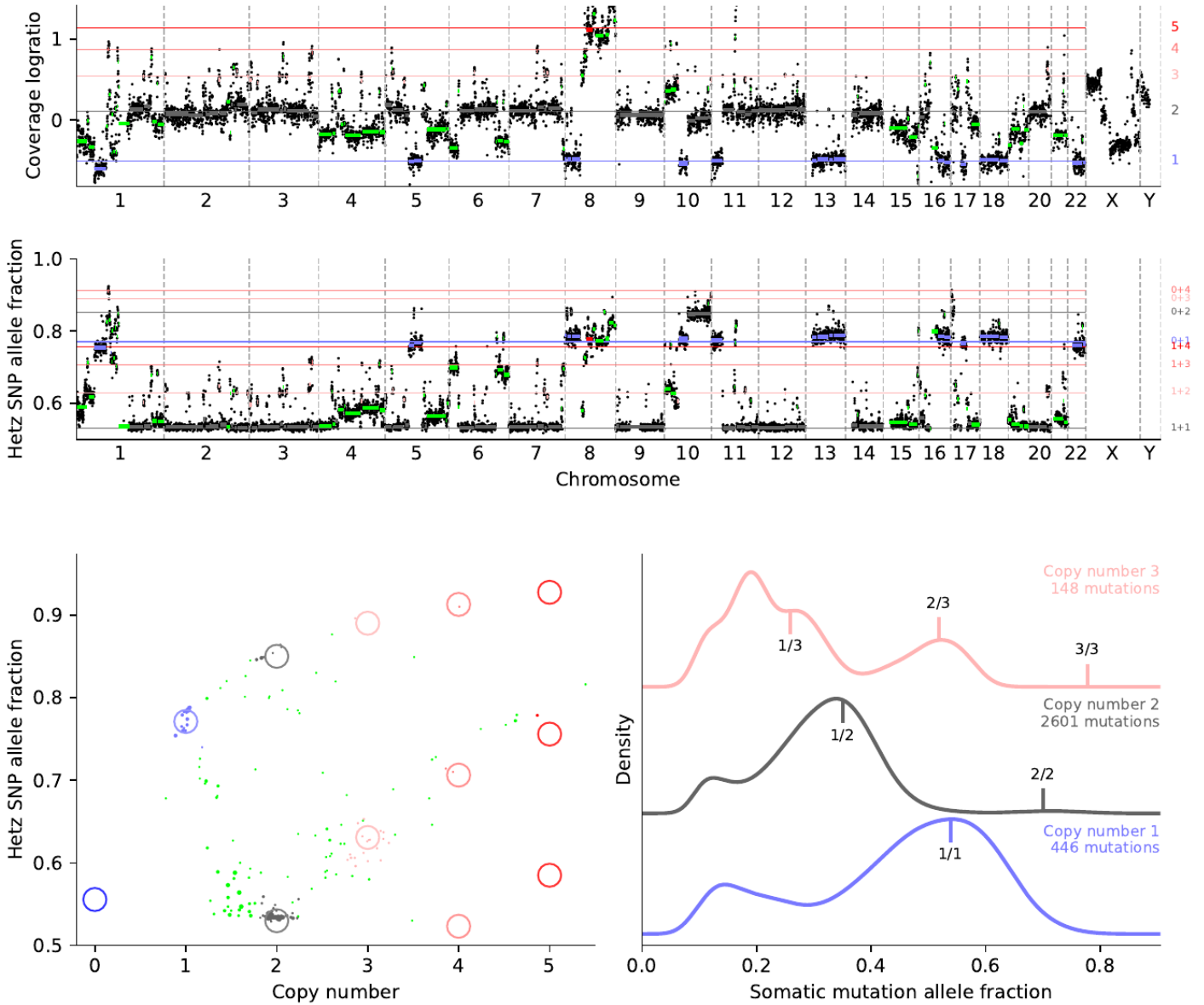
AE-156-Baseline-cfDNA

Cancer fraction: 43%, Diploid level: 0.07

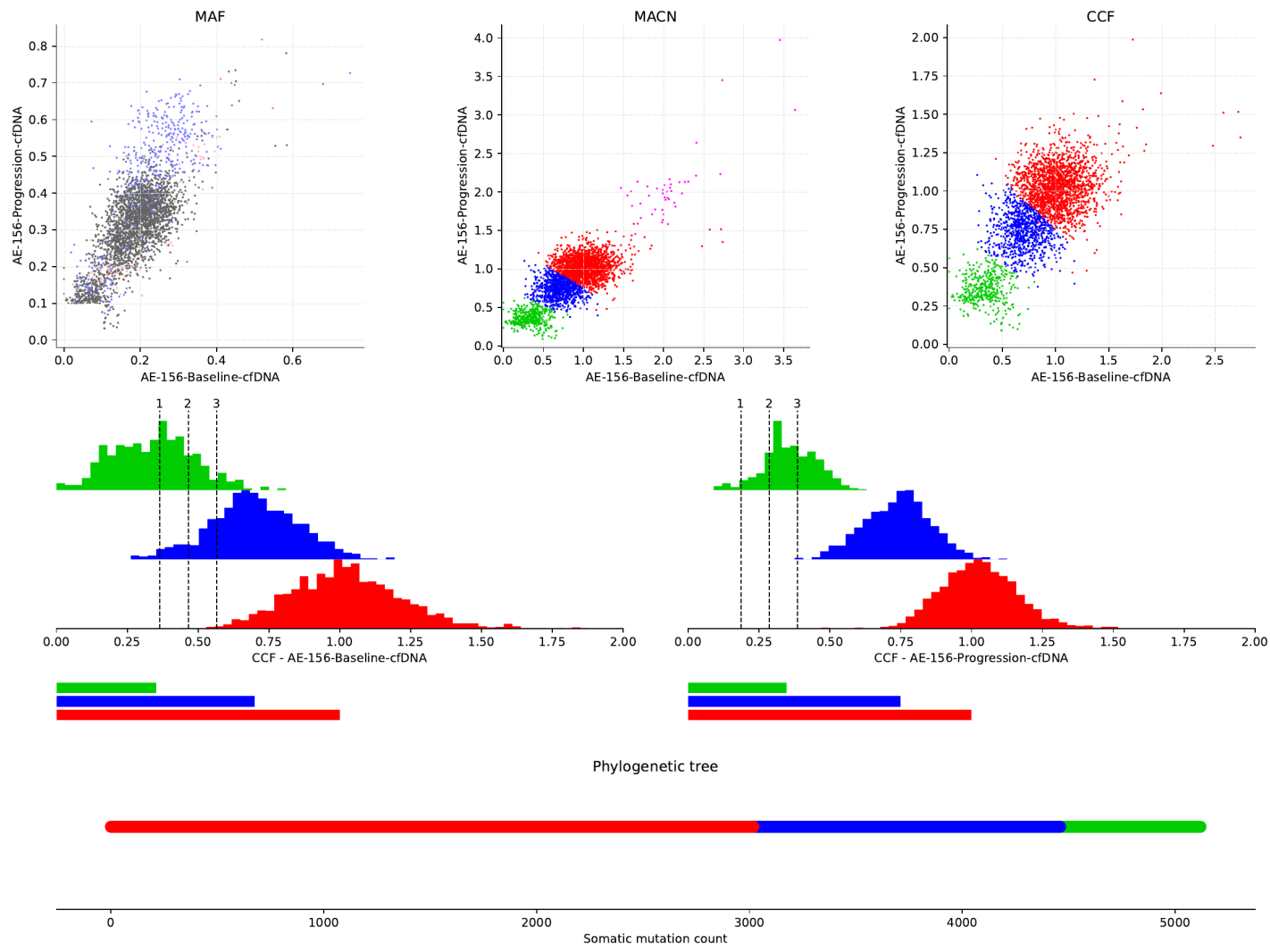


AE-156-Progression-cfDNA

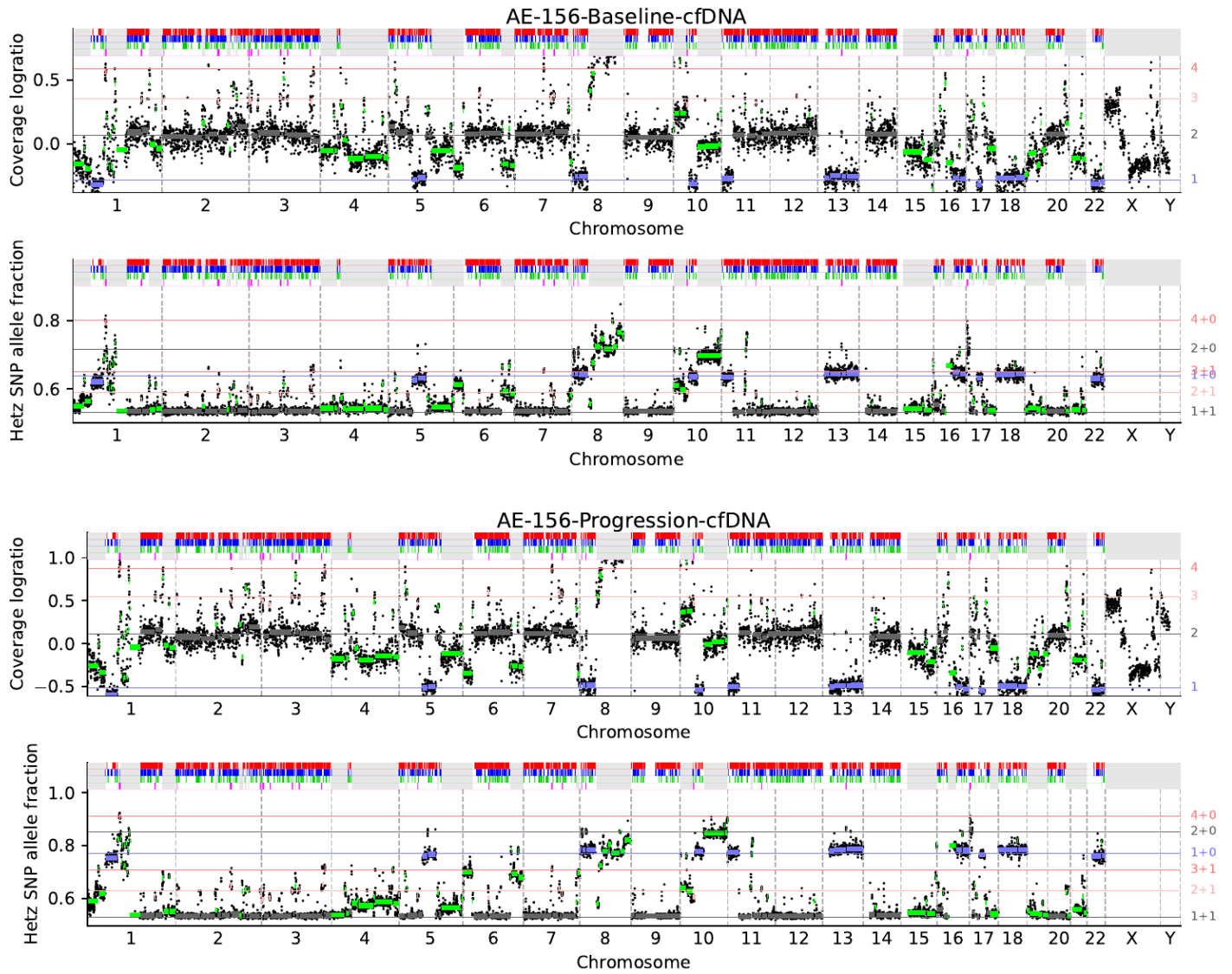
Cancer fraction: 70%, Diploid level: 0.11



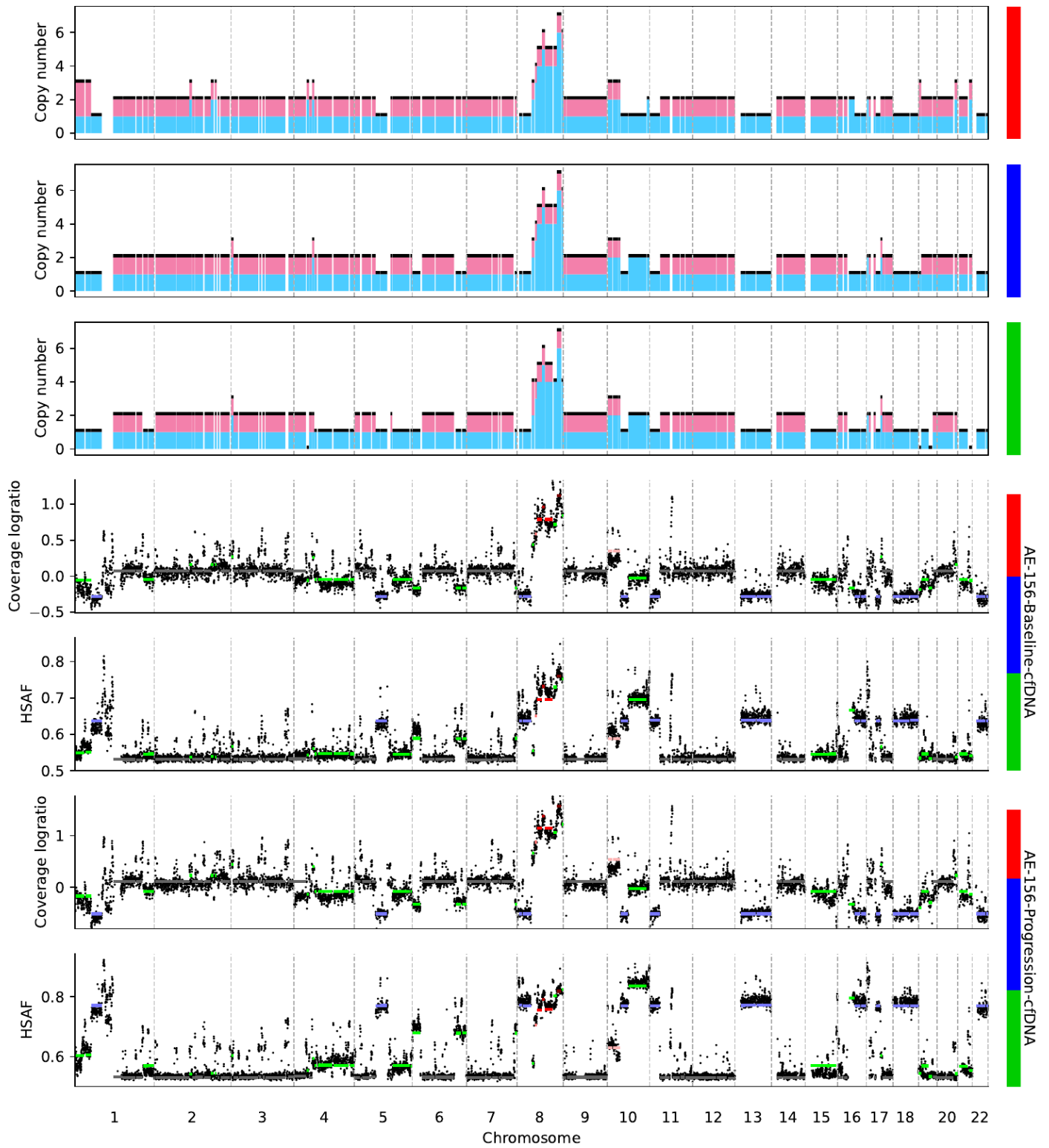
AE-156 mutation clustering



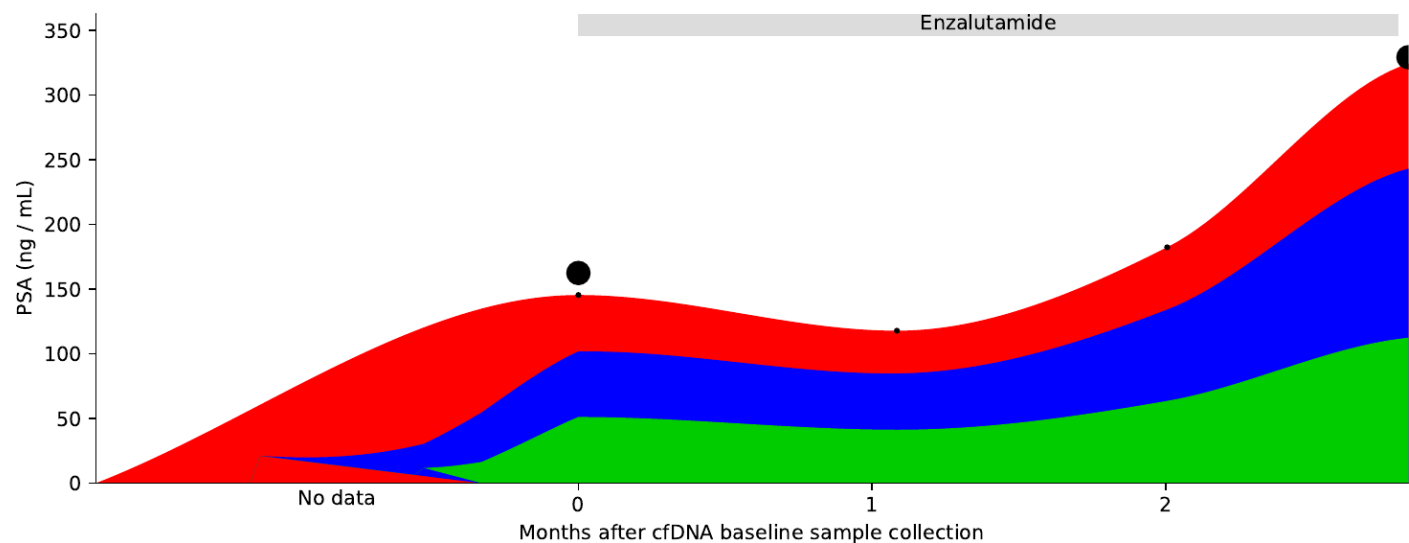
AE-156 mutation cluster positions



AE-156 copy number deconvolution



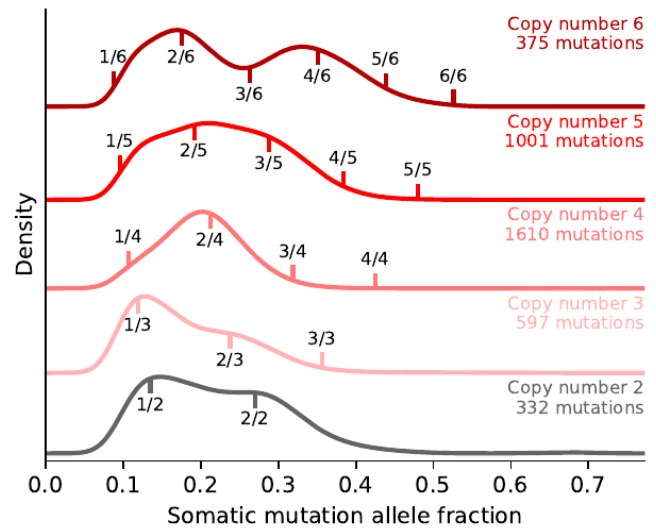
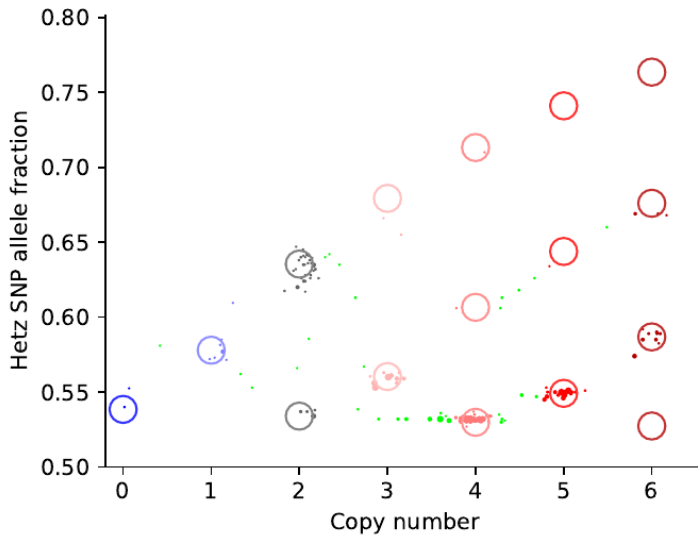
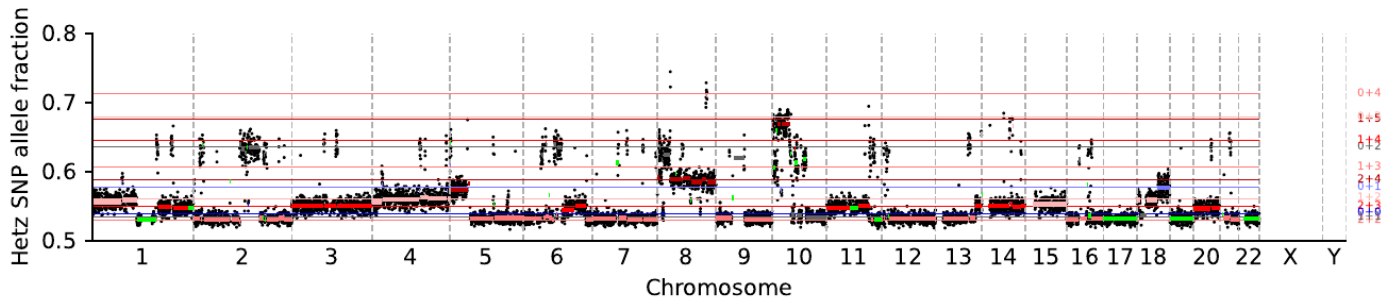
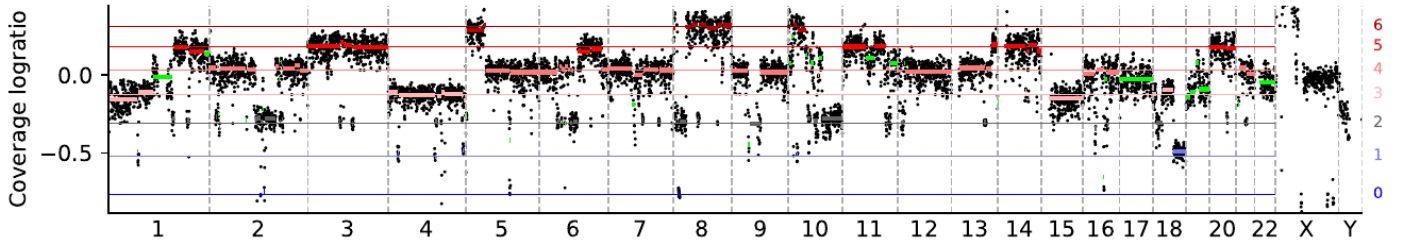
AE-156 clonal expansion timeline



Patient AE-160

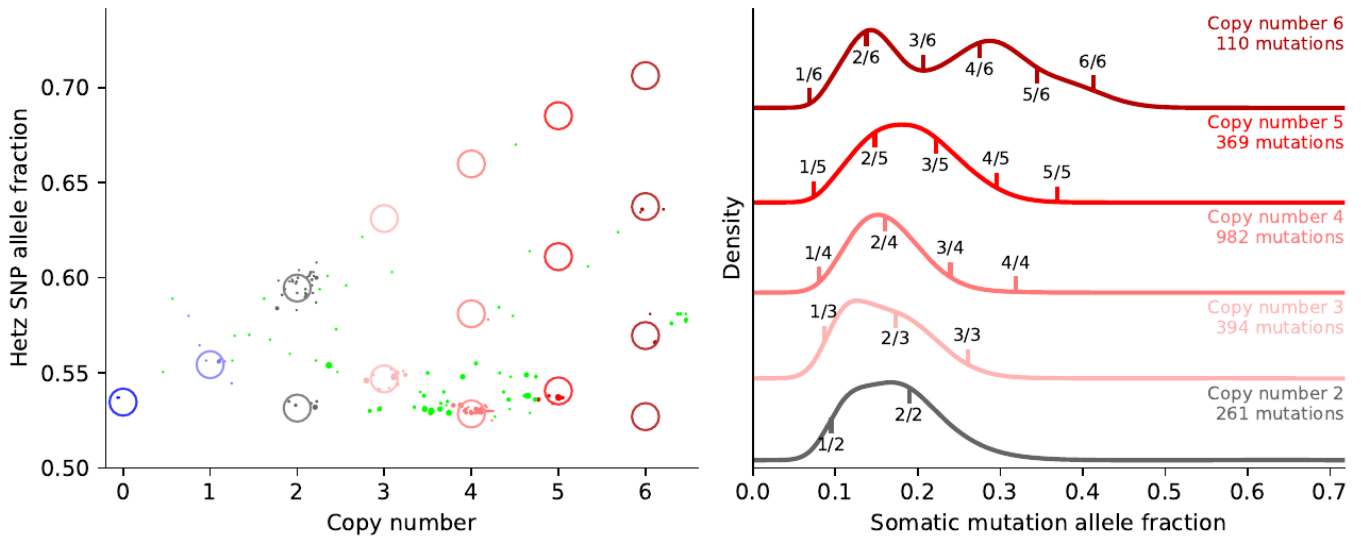
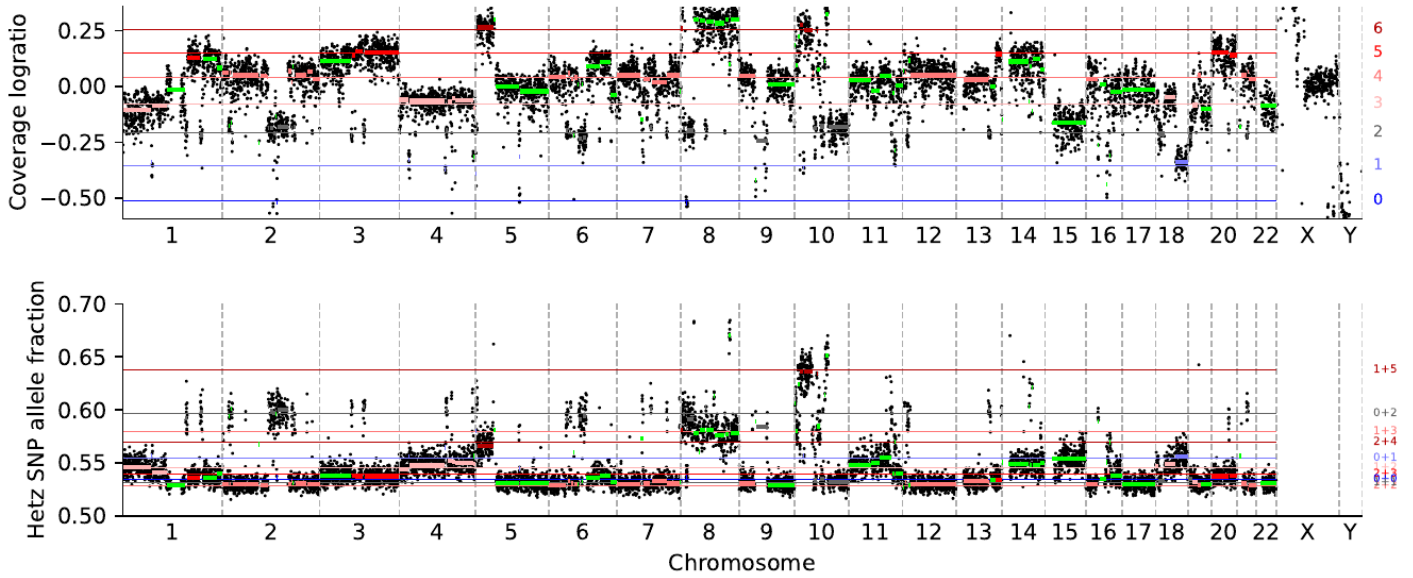
AE-160-Baseline-cfDNA

Cancer fraction: 27%, Diploid level: -0.31



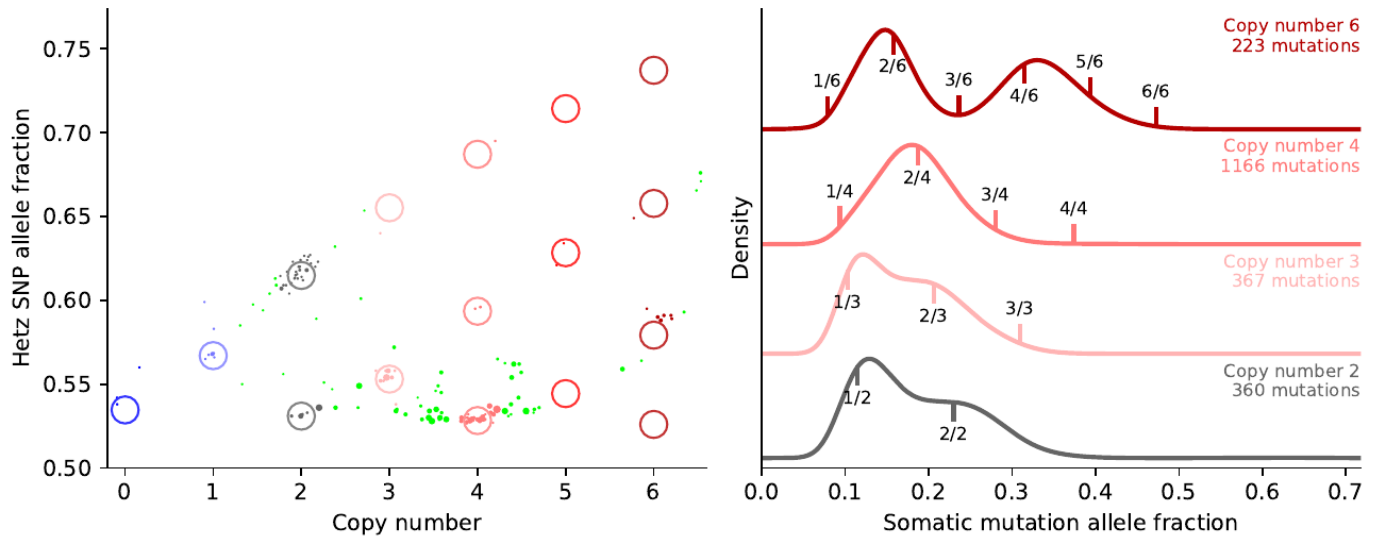
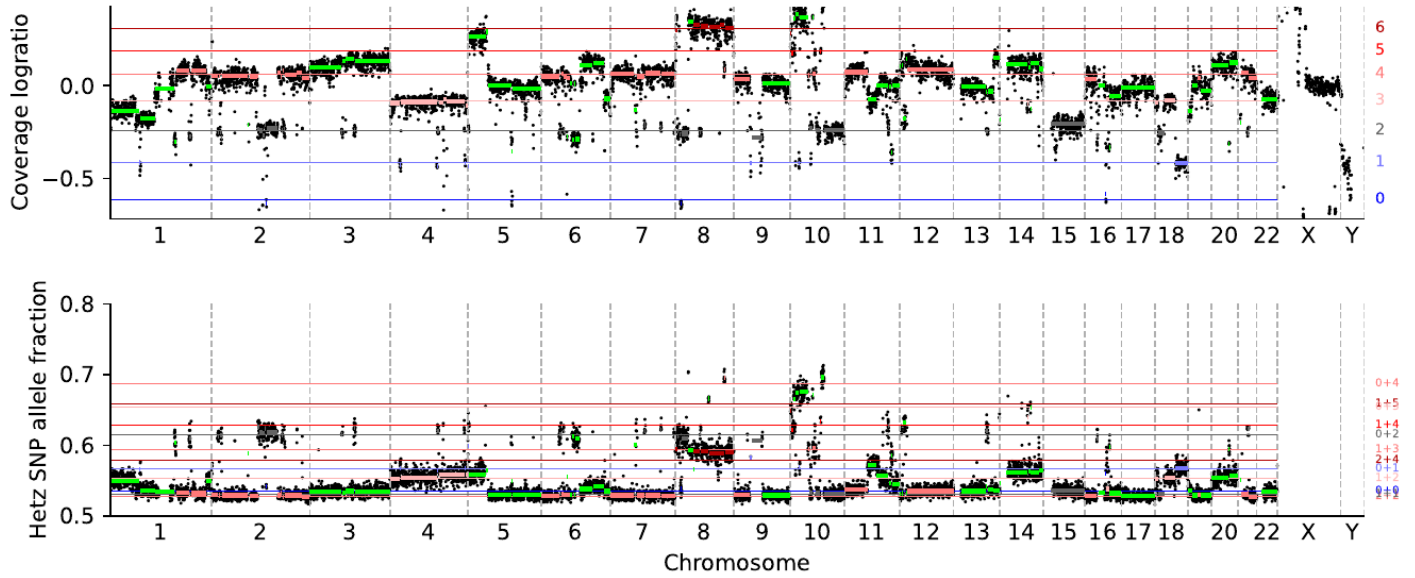
AE-160-Progression-cfDNA

Cancer fraction: 19%, Diploid level: -0.21

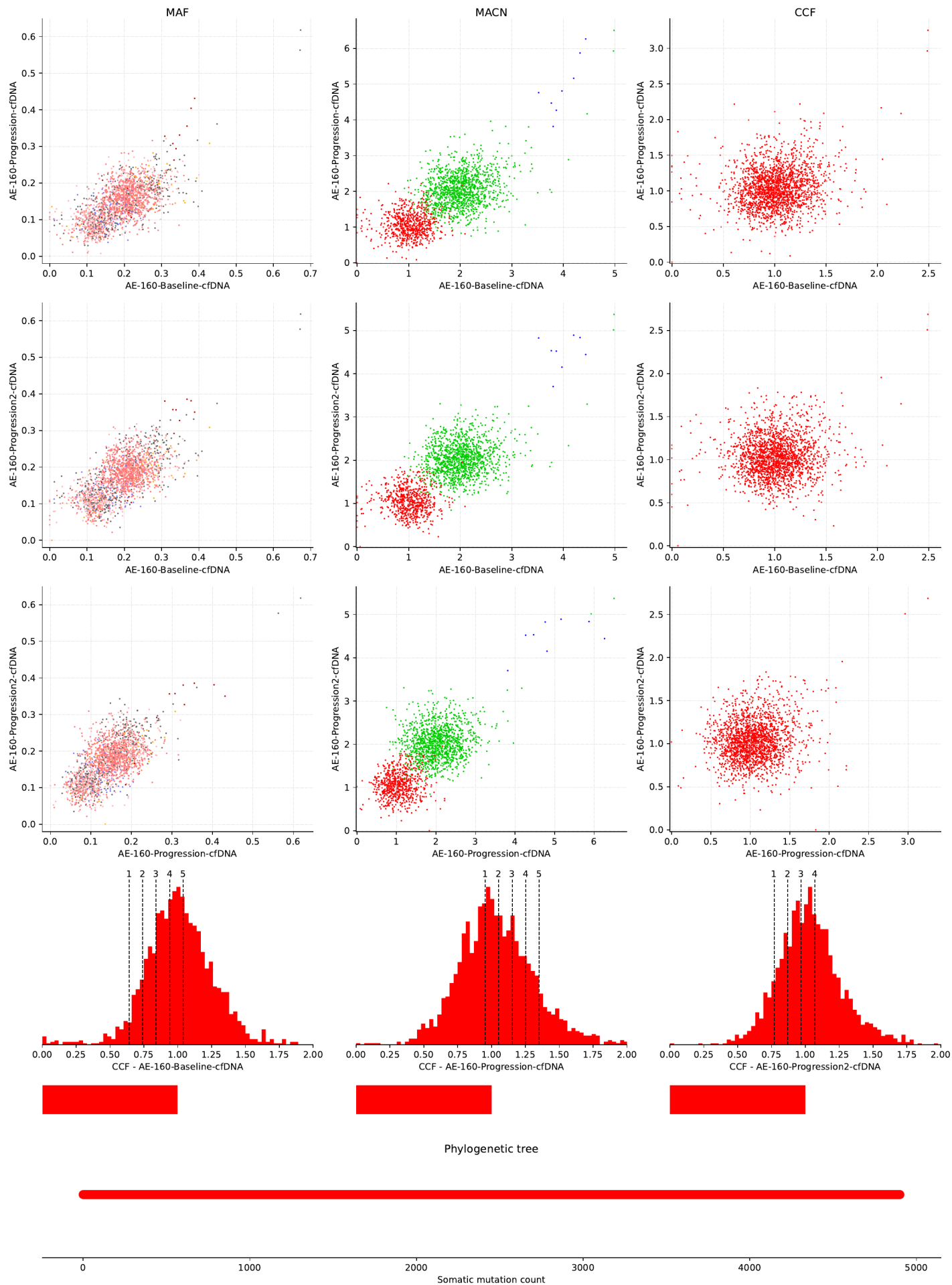


AE-160-Progression2-cfDNA

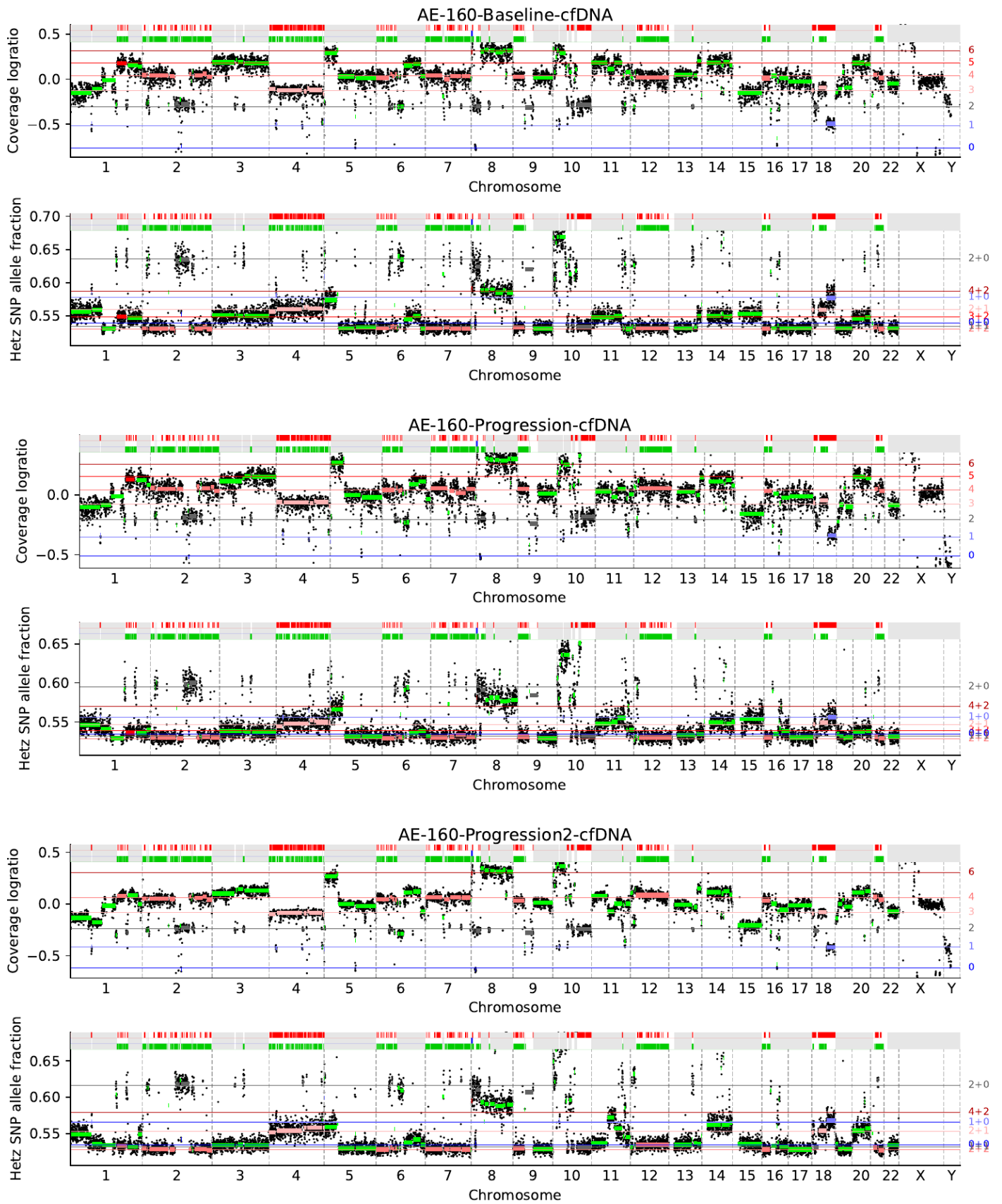
Cancer fraction: 23%, Diploid level: -0.24



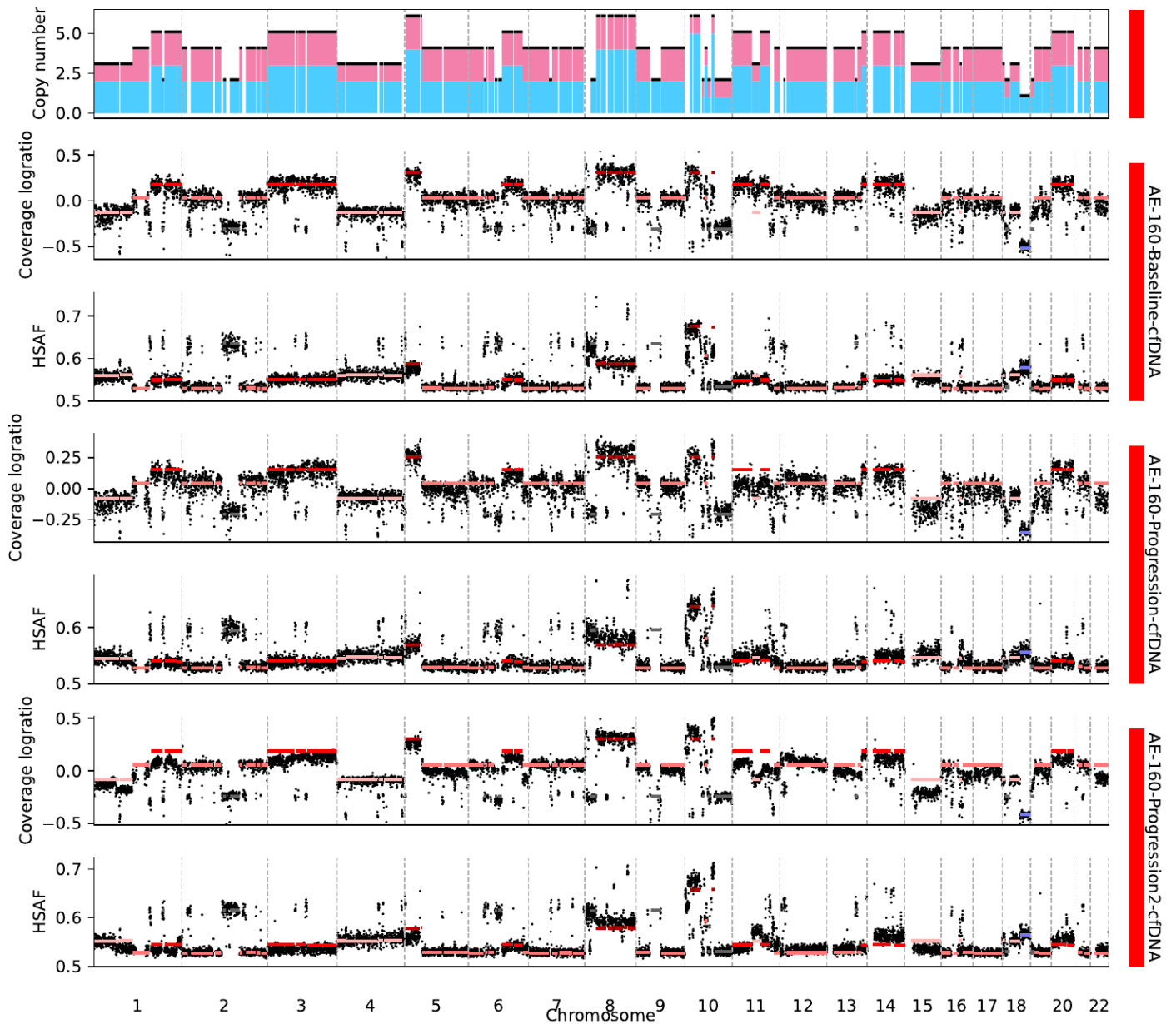
AE-160 mutation clustering



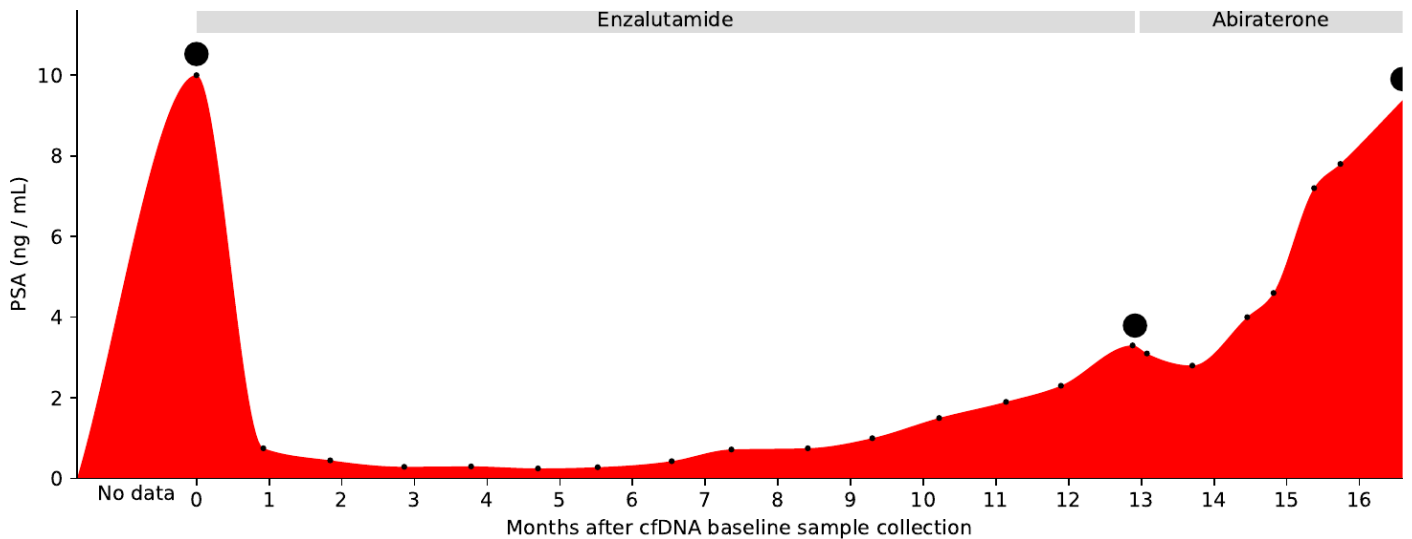
AE-160 mutation cluster positions



AE-160 copy number deconvolution



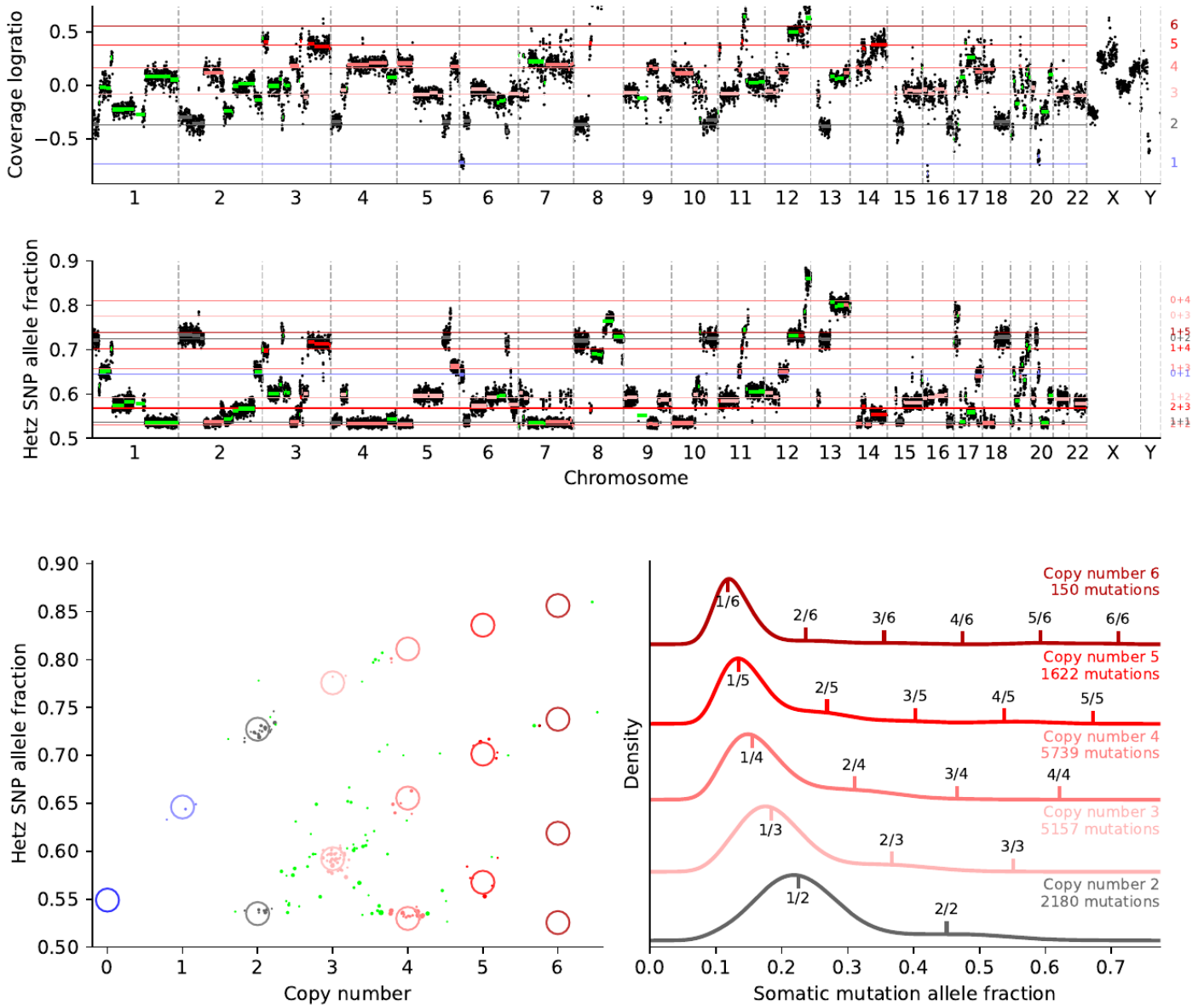
AE-160 clonal expansion timeline



Patient AE-180

AE-180-Baseline-cfDNA

Cancer fraction: 45%, Diploid level: -0.37

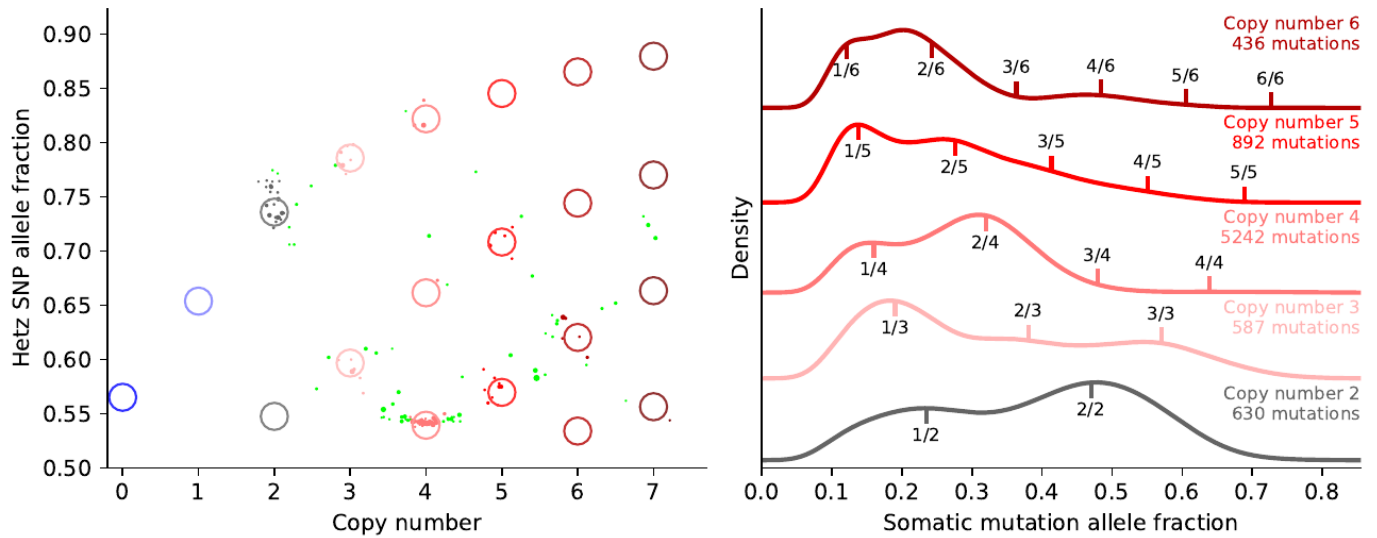
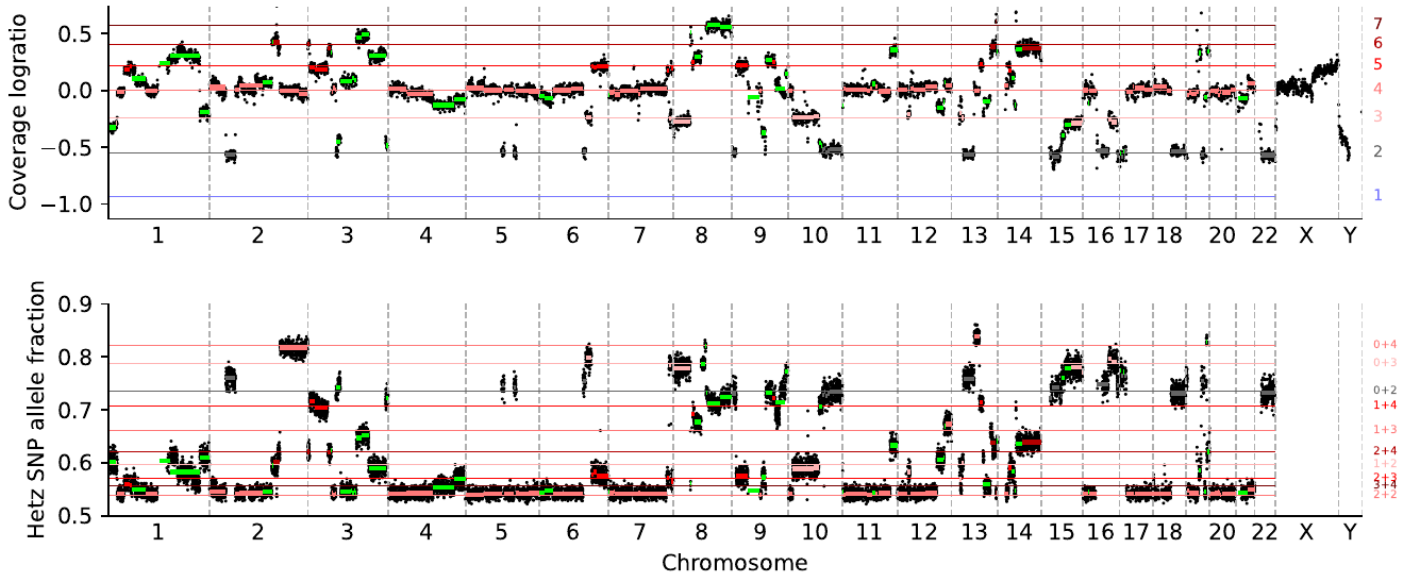


AE-180-Progression-cfDNA

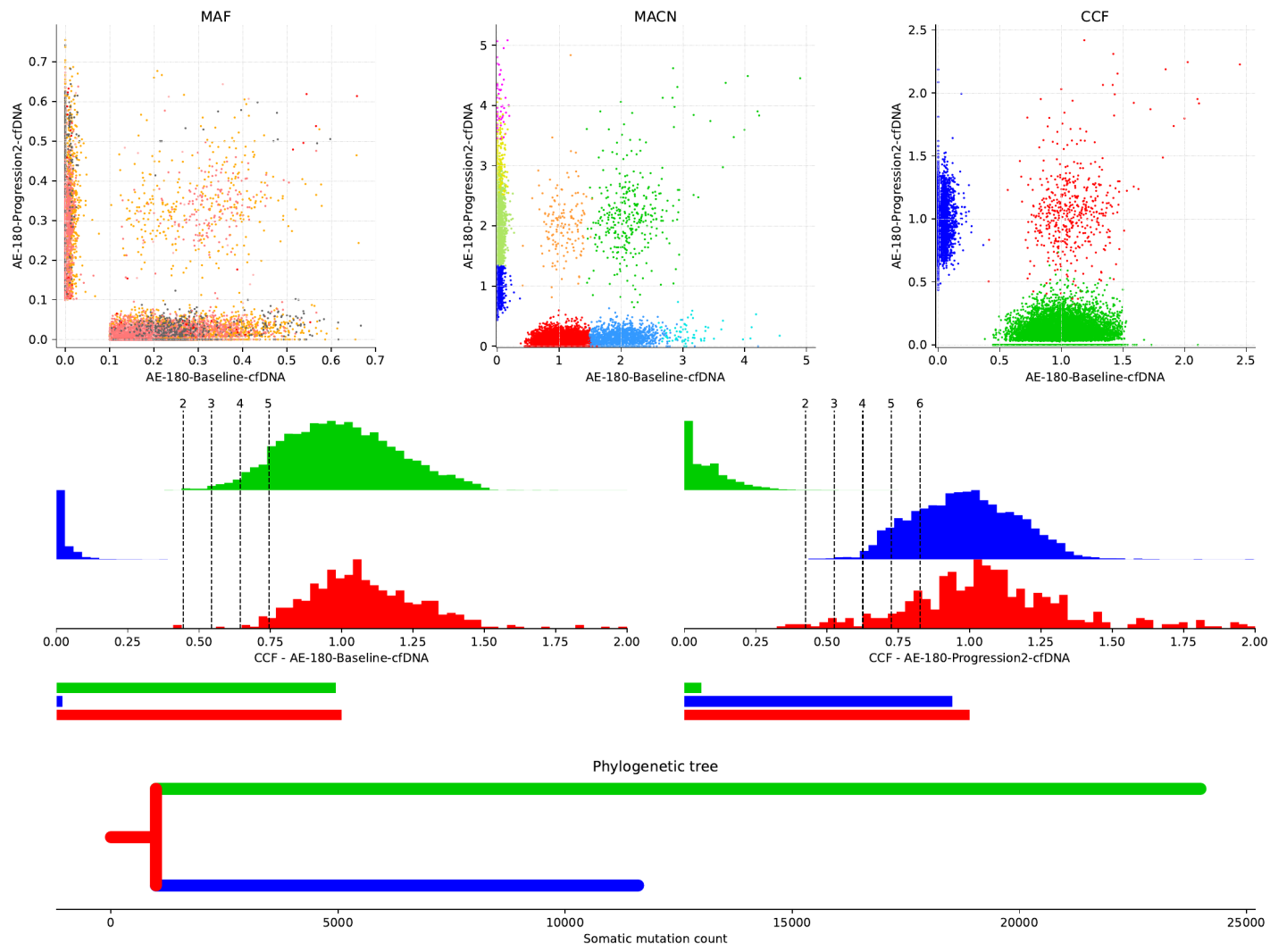
We were unable to infer a truncal copy number model for AE-180-Progression-cfDNA due to an unusually high degree of subclonal heterogeneity.

AE-180-Progression2-cfDNA

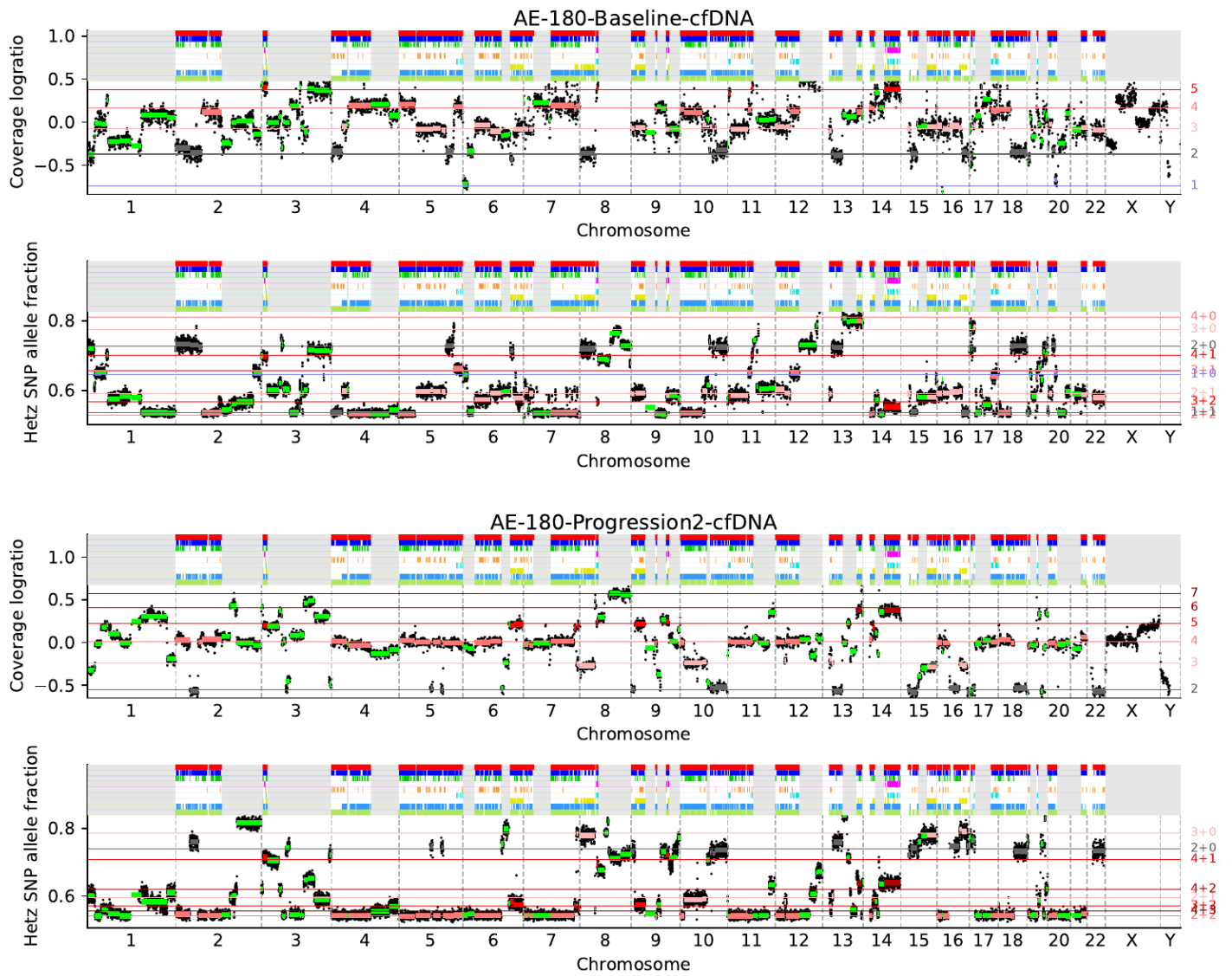
Cancer fraction: 47%, Diploid level: -0.55



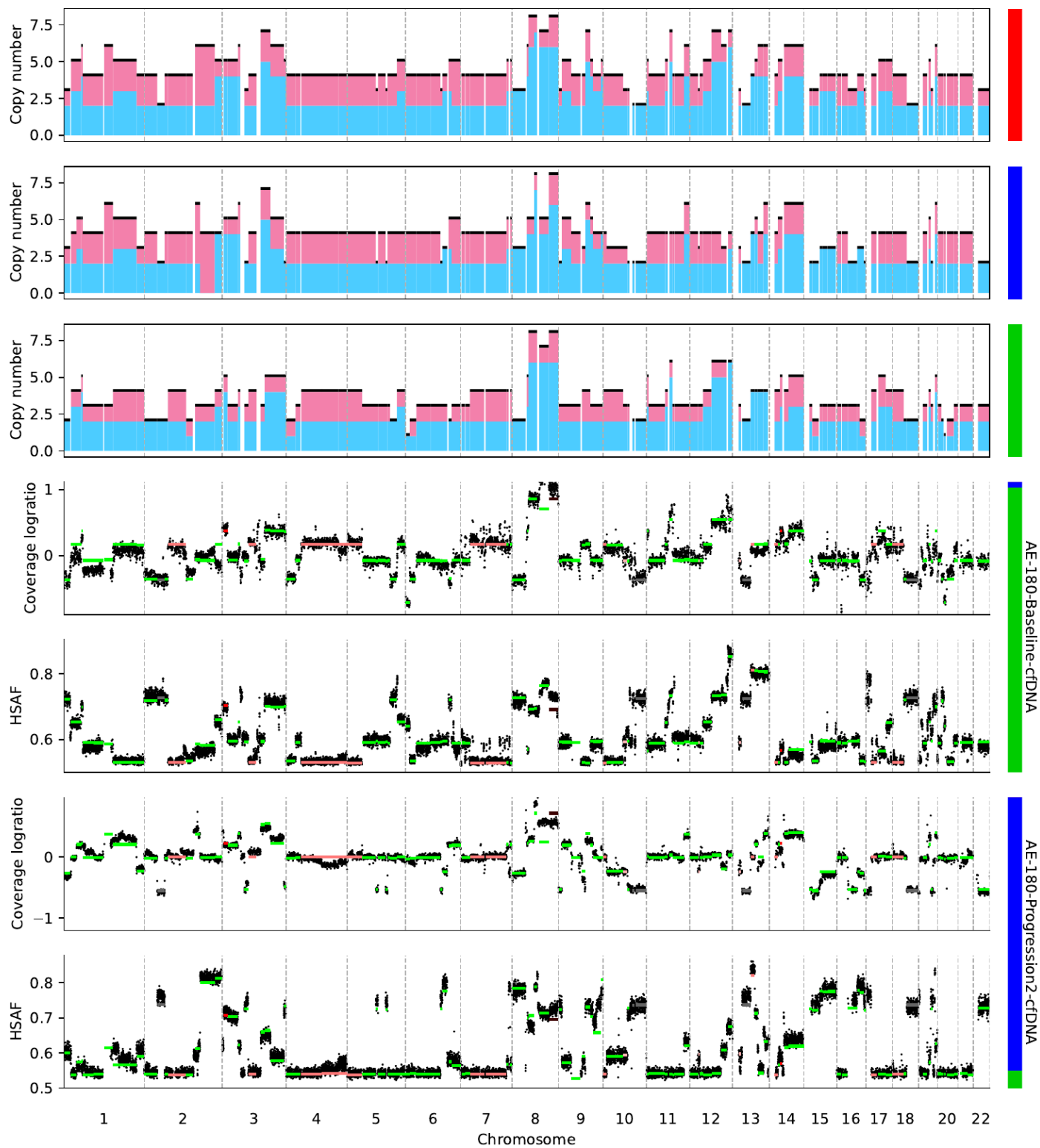
AE-180 mutation clustering



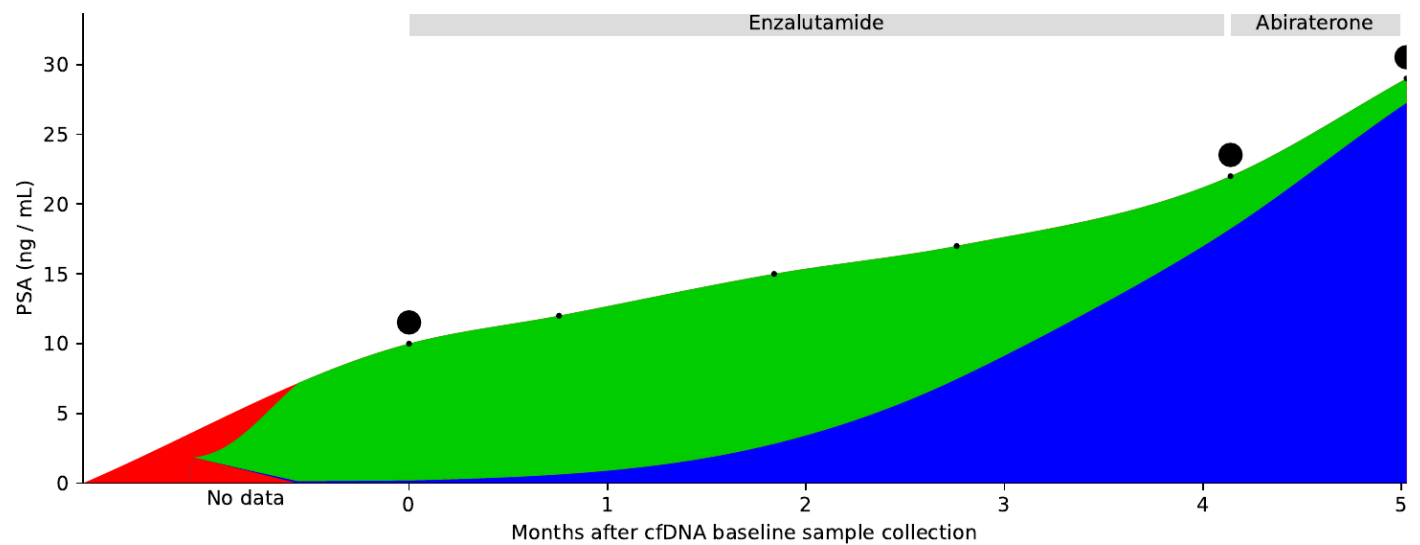
AE-180 mutation cluster positions



AE-180 copy number deconvolution



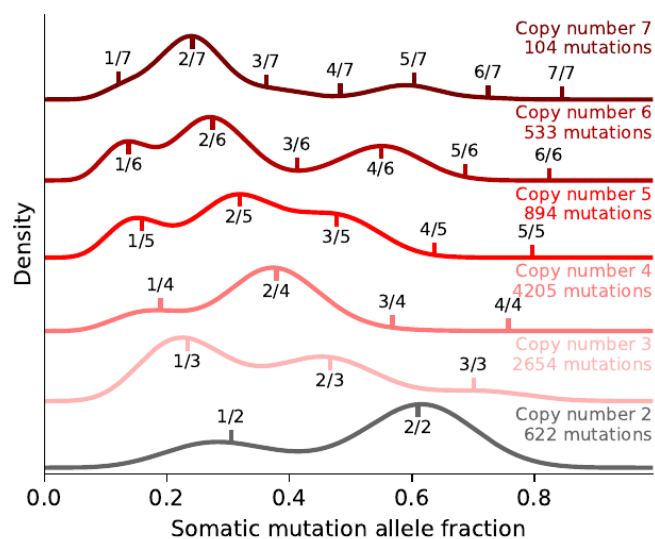
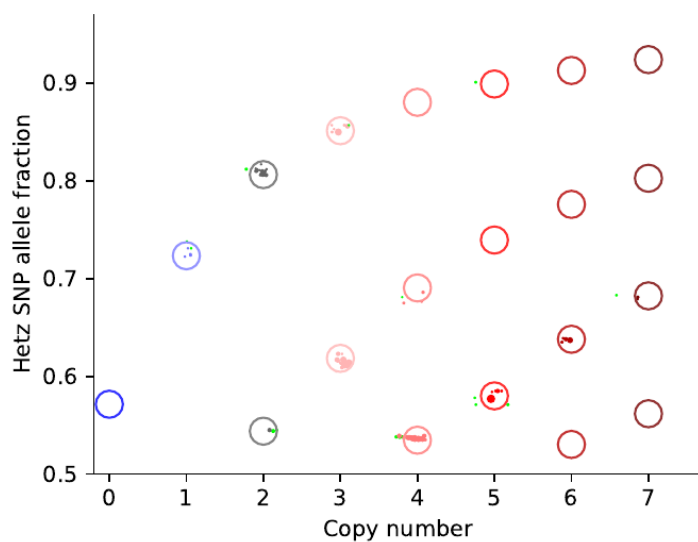
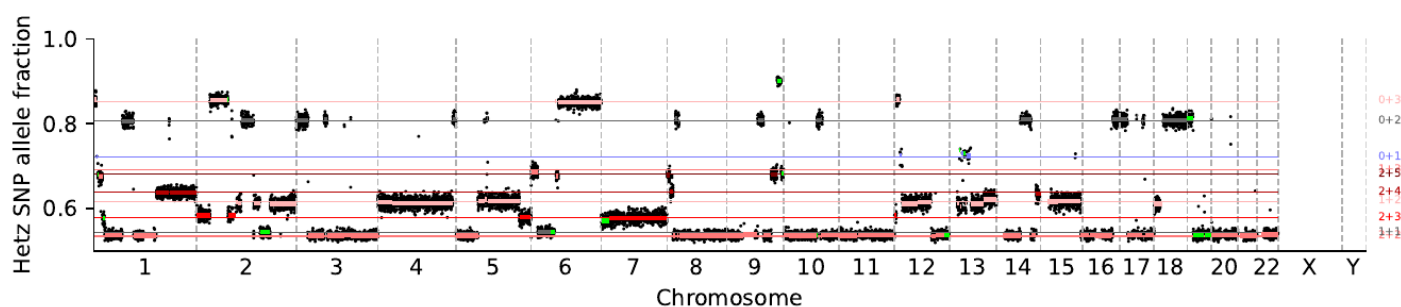
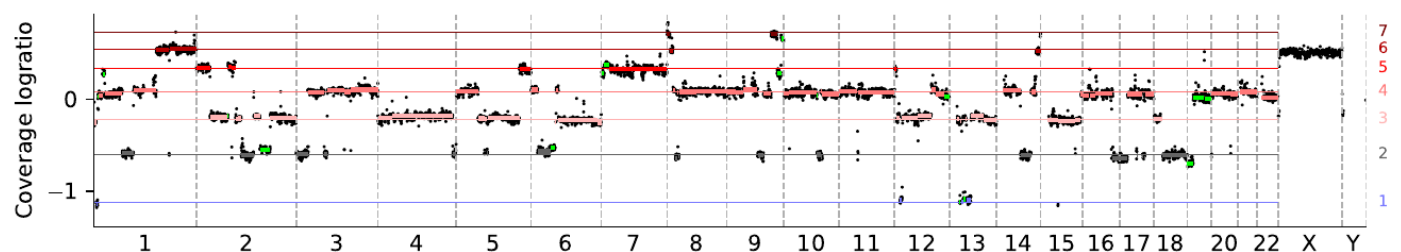
AE-180 clonal expansion timeline



Patient DTB-097

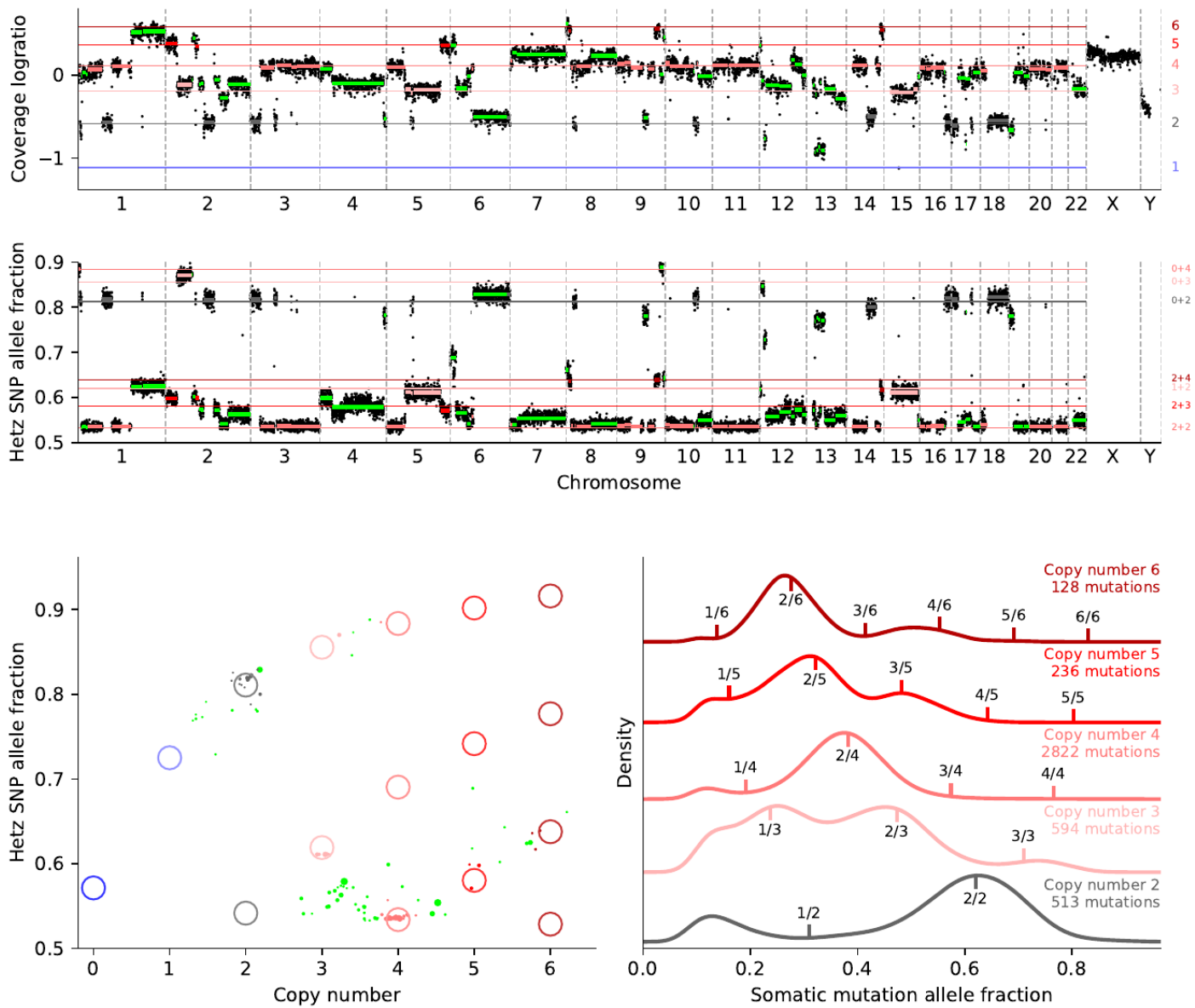
DTB-097-Progression-Tumor

Cancer fraction: 61%, Diploid level: -0.60

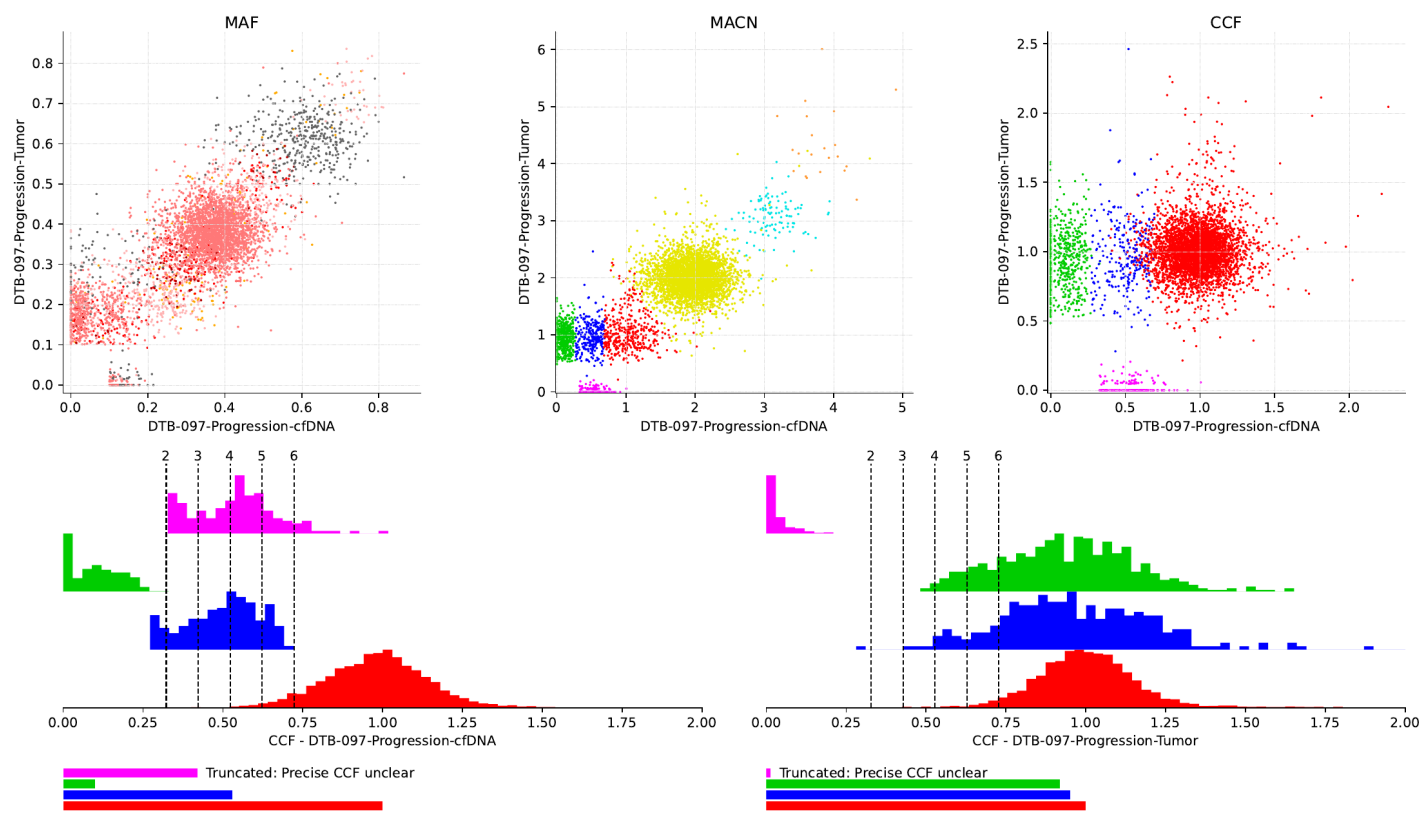


DTB-097-Progression-cfDNA

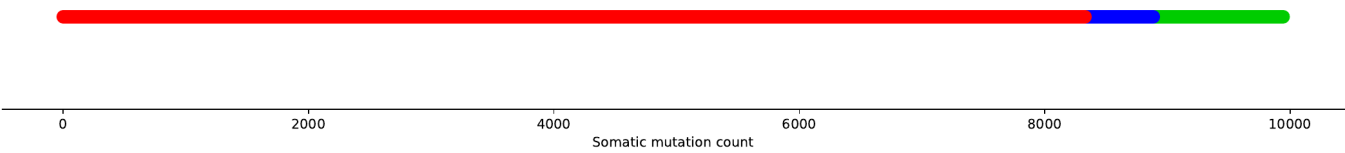
Cancer fraction: 62%, Diploid level: -0.58



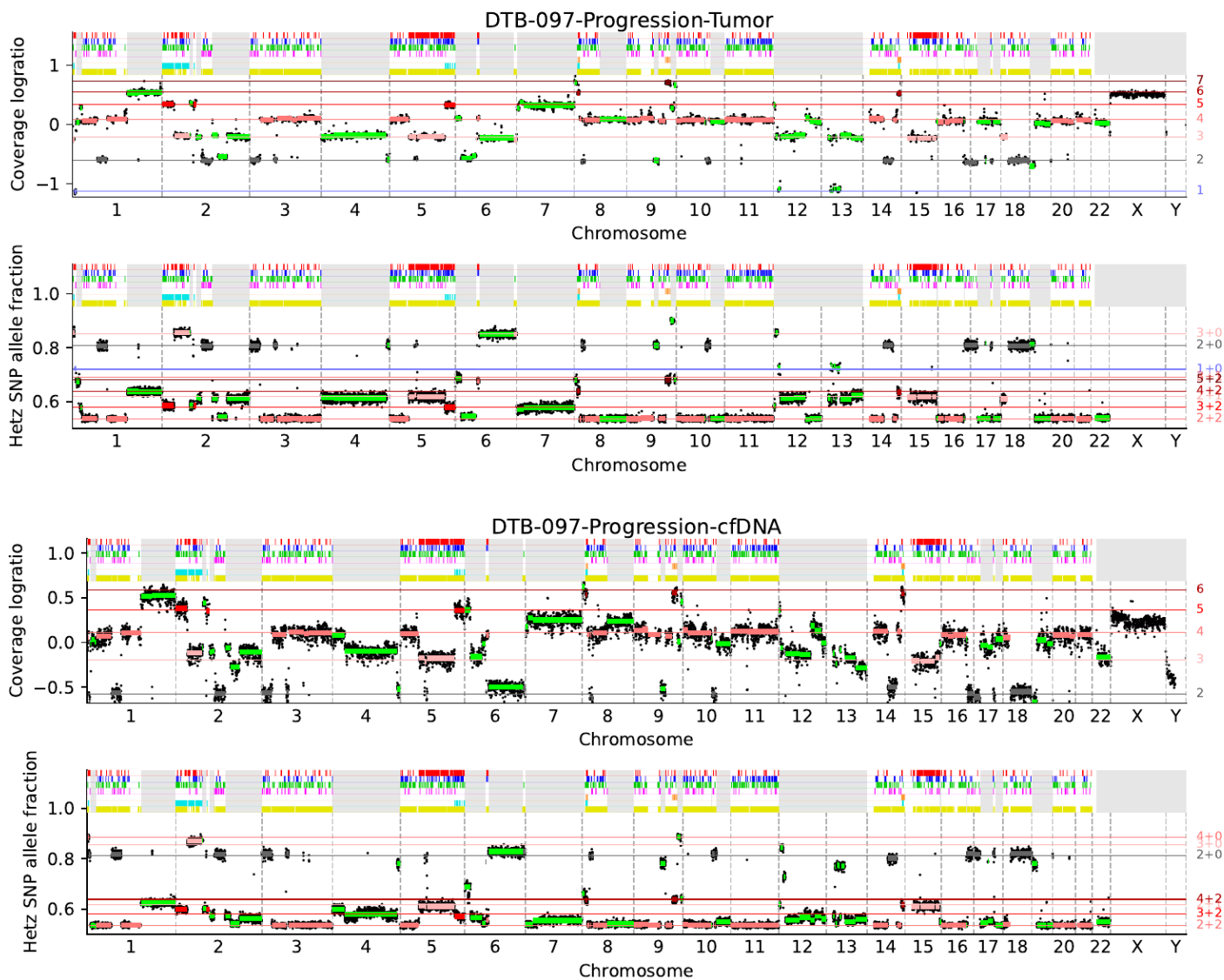
DTB-097 mutation clustering



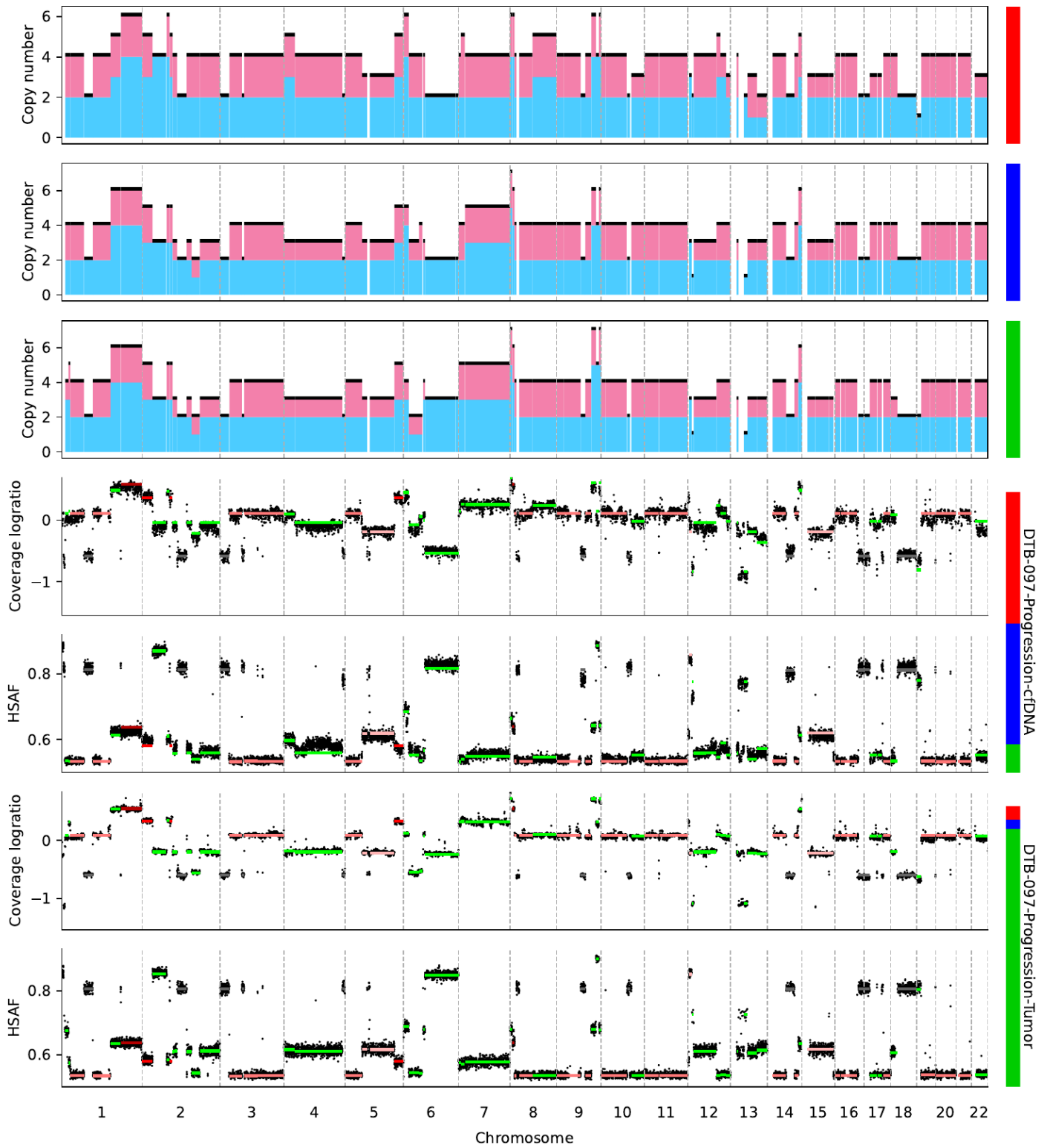
Phylogenetic tree



DTB-097 mutation cluster positions



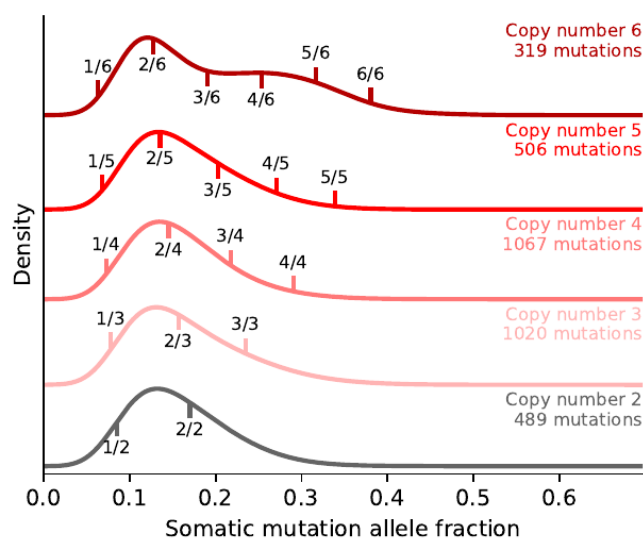
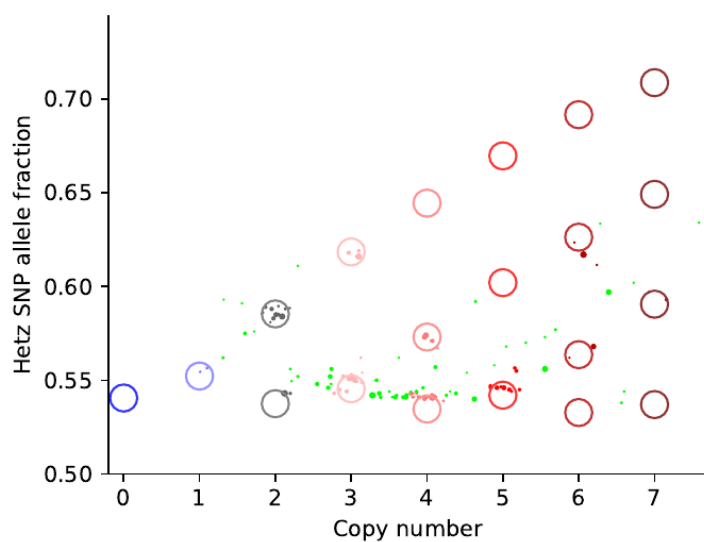
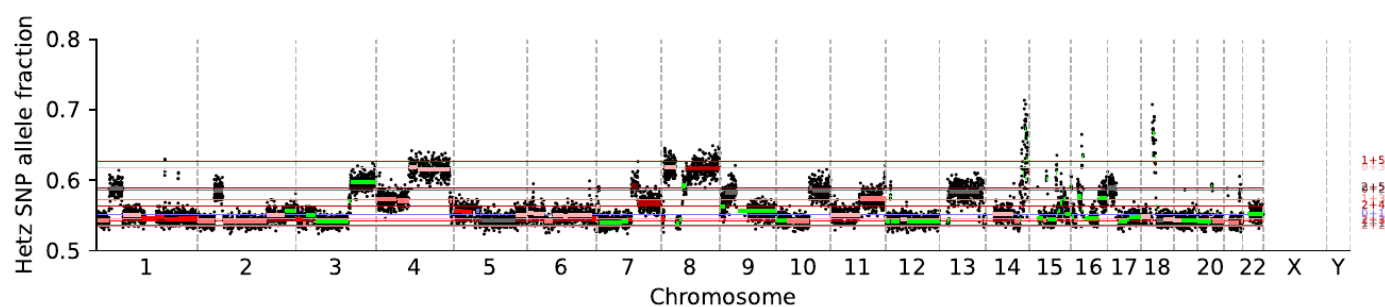
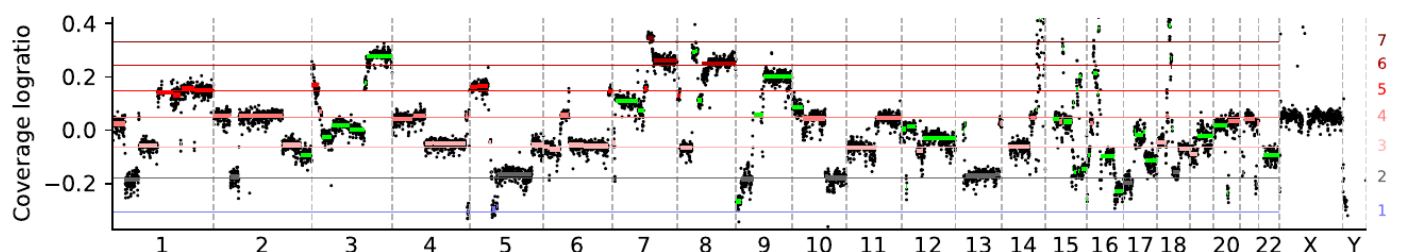
DTB-097 copy number deconvolution



Patient DTB-102

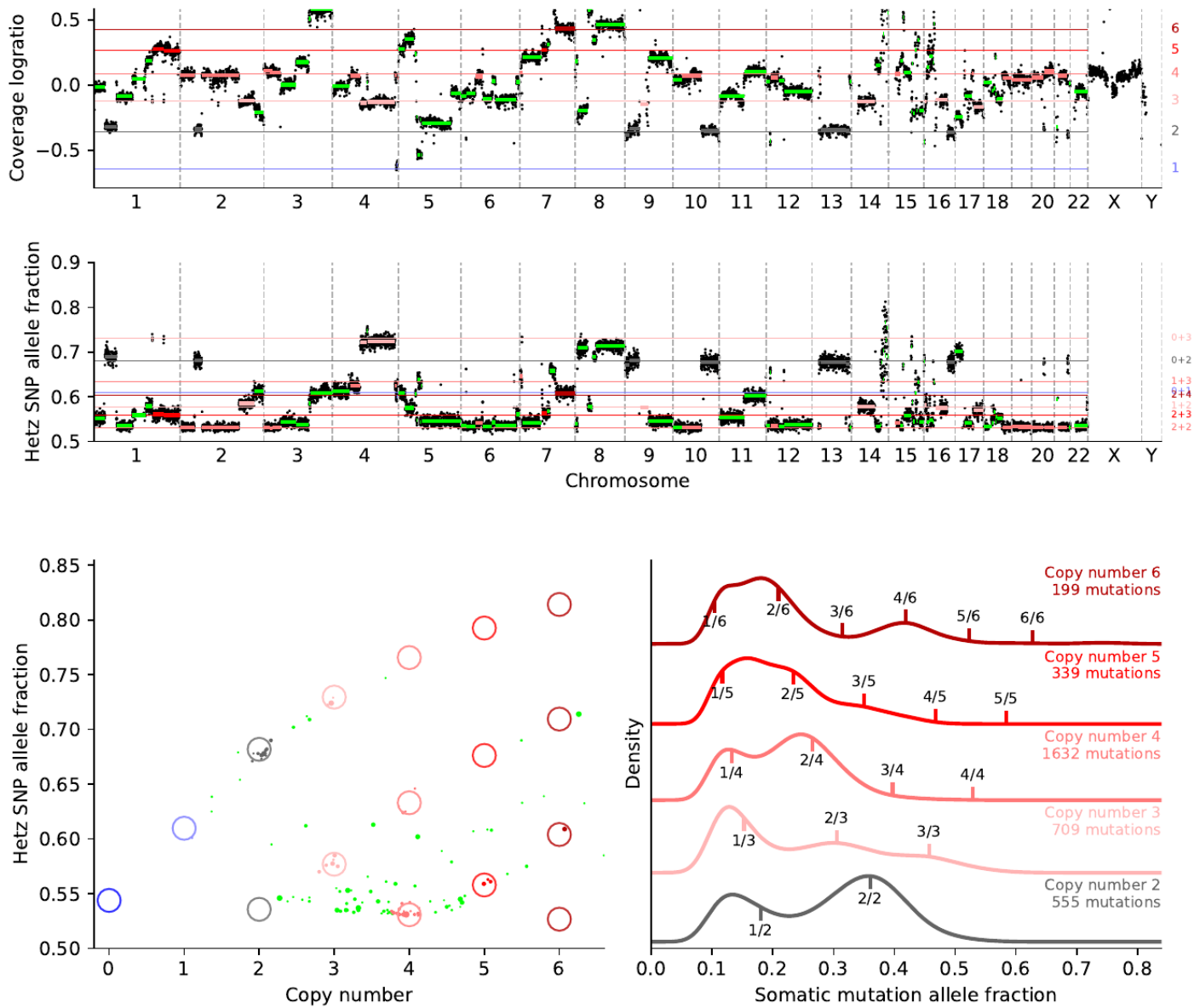
DTB-102-Progression-Tumor

Cancer fraction: 17%, Diploid level: -0.18

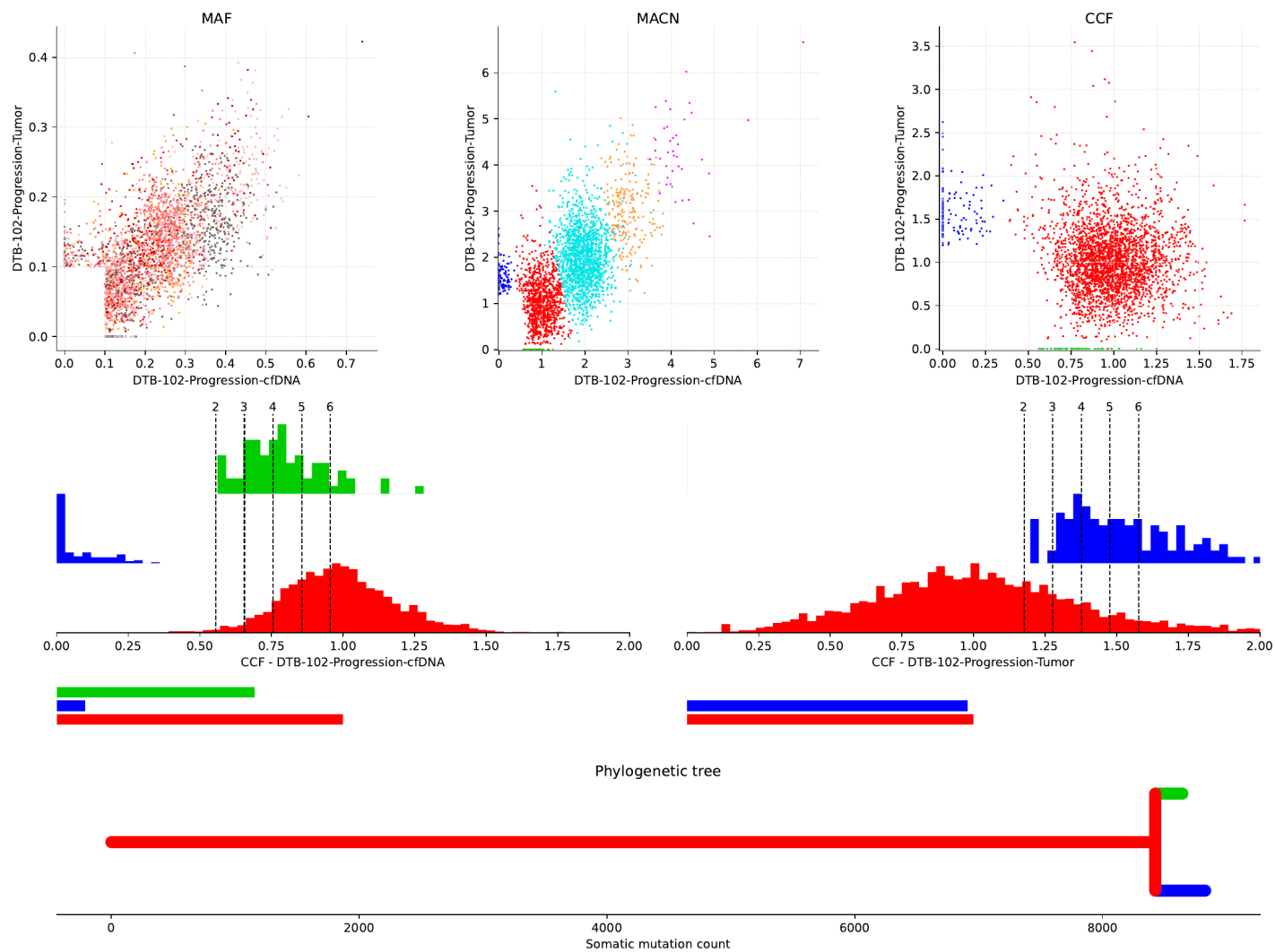


DTB-102-Progression-cfDNA

Cancer fraction: 36%, Diploid level: -0.36

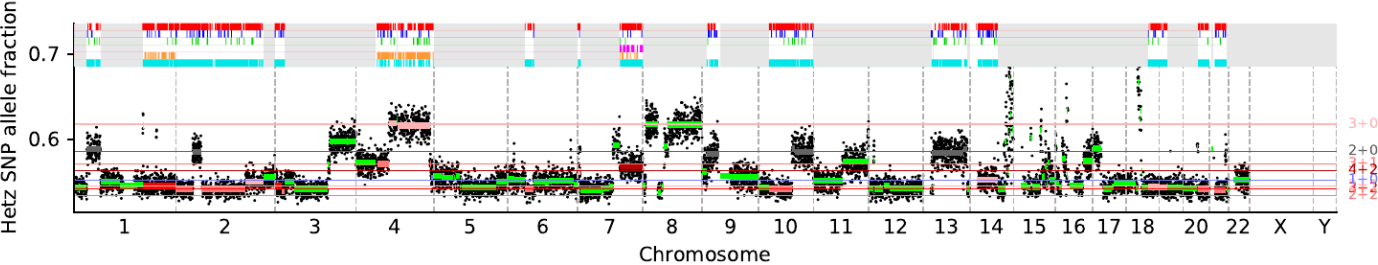
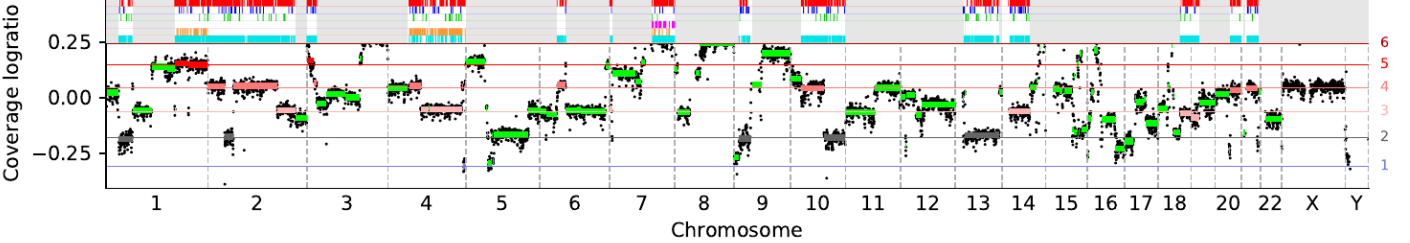


DTB-102 mutation clustering

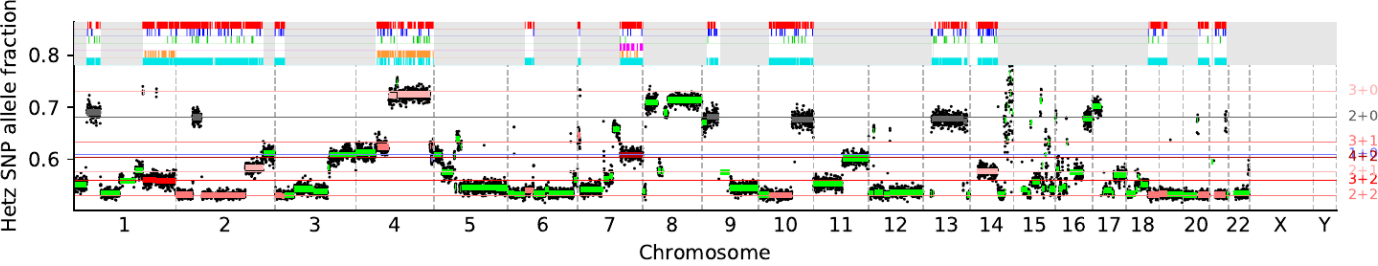
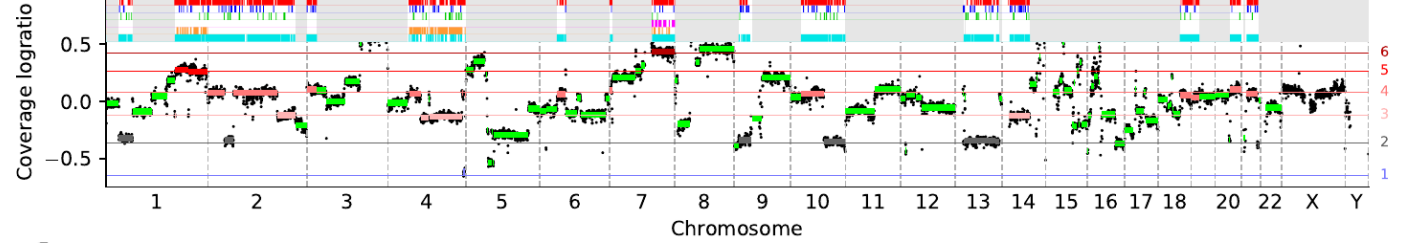


DTB-102 mutation cluster positions

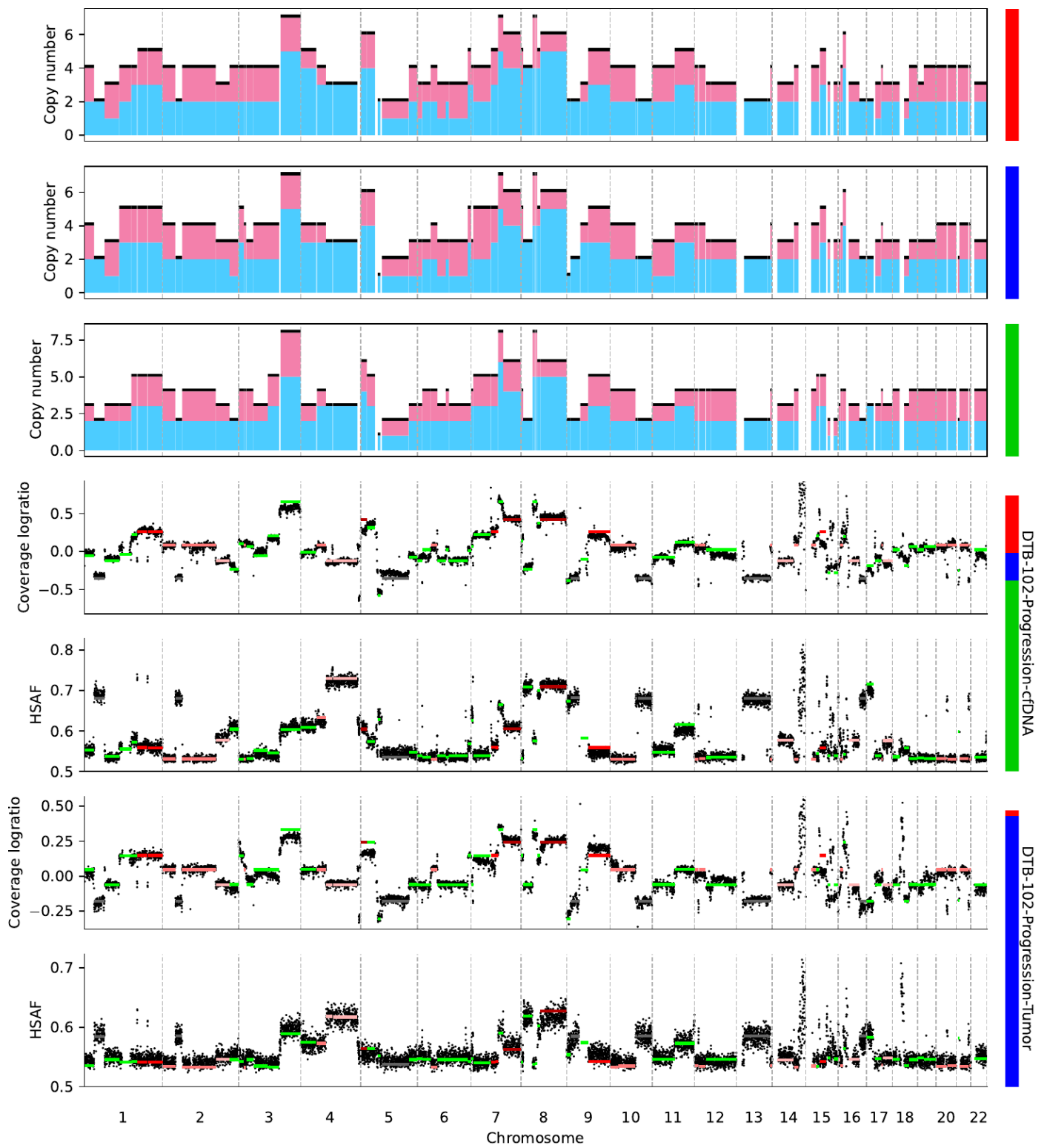
DTB-102-Progression-Tumor



DTB-102-Progression-cfDNA



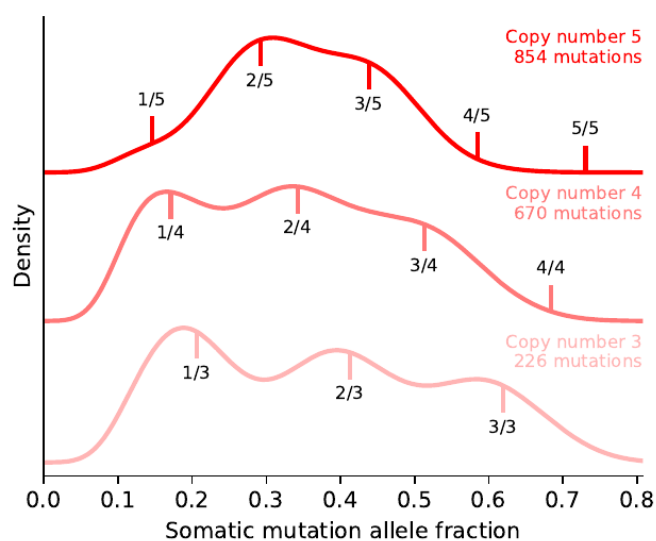
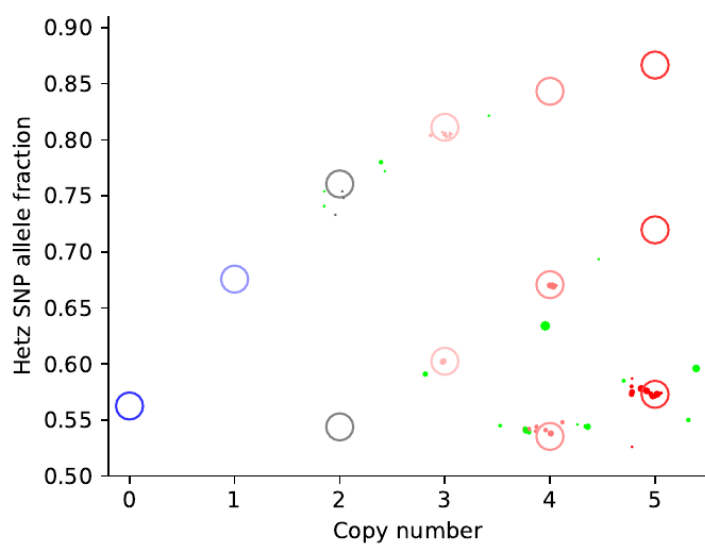
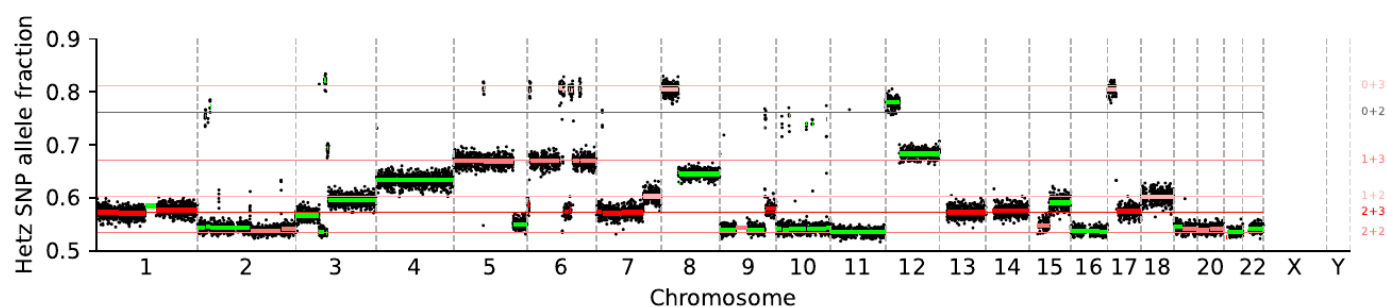
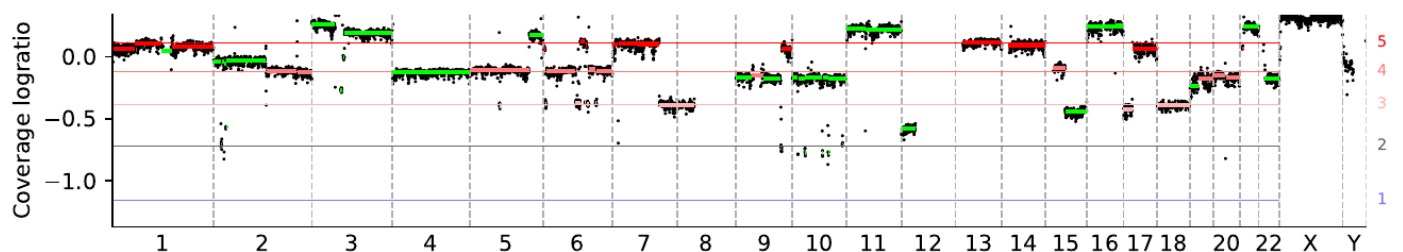
DTB-102 copy number deconvolution



Patient DTB-119

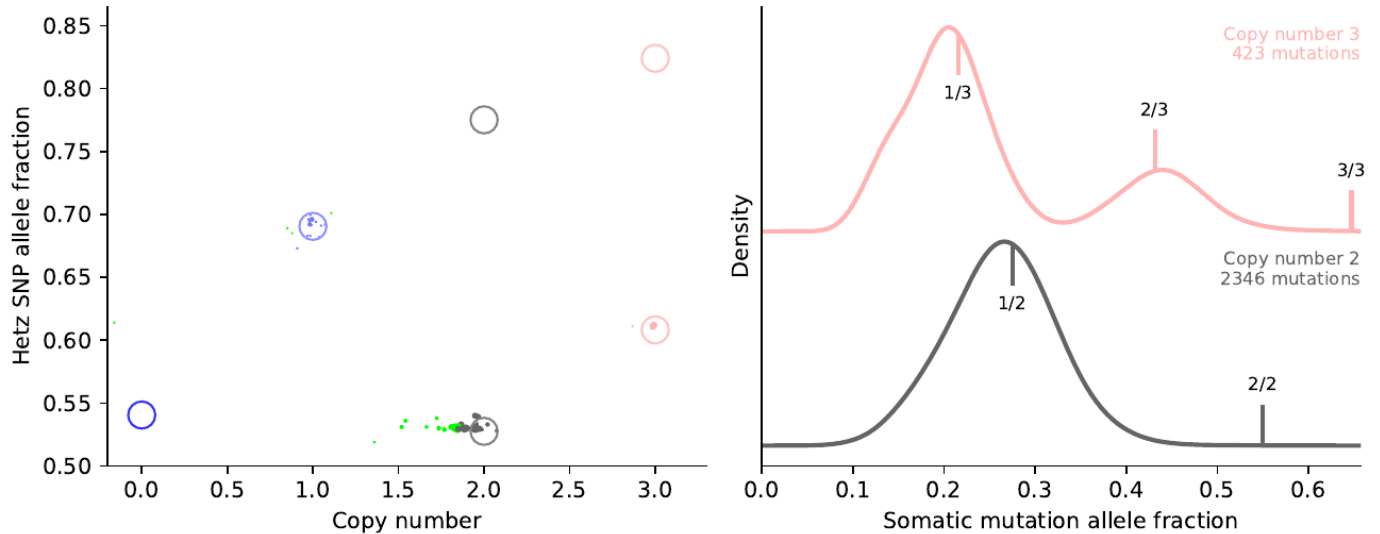
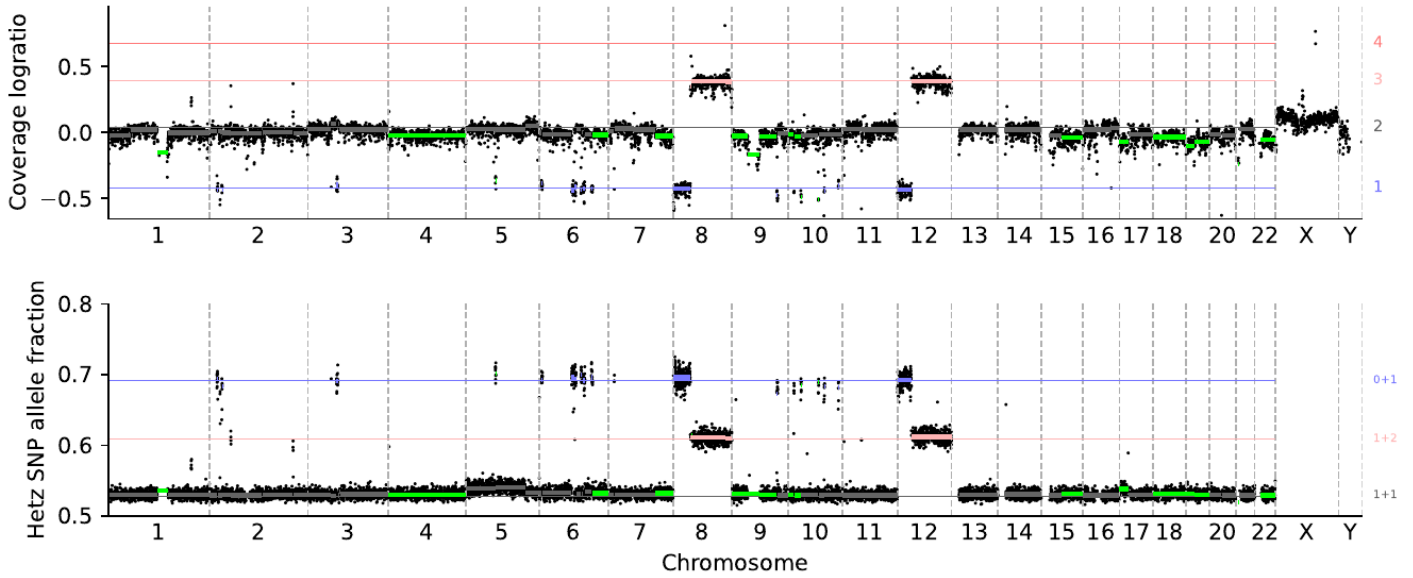
DTB-119-Progression-Tumor

Cancer fraction: 52%, Diploid level: -0.72



DTB-119-Progression-cfDNA

Cancer fraction: 55%, Diploid level: 0.04



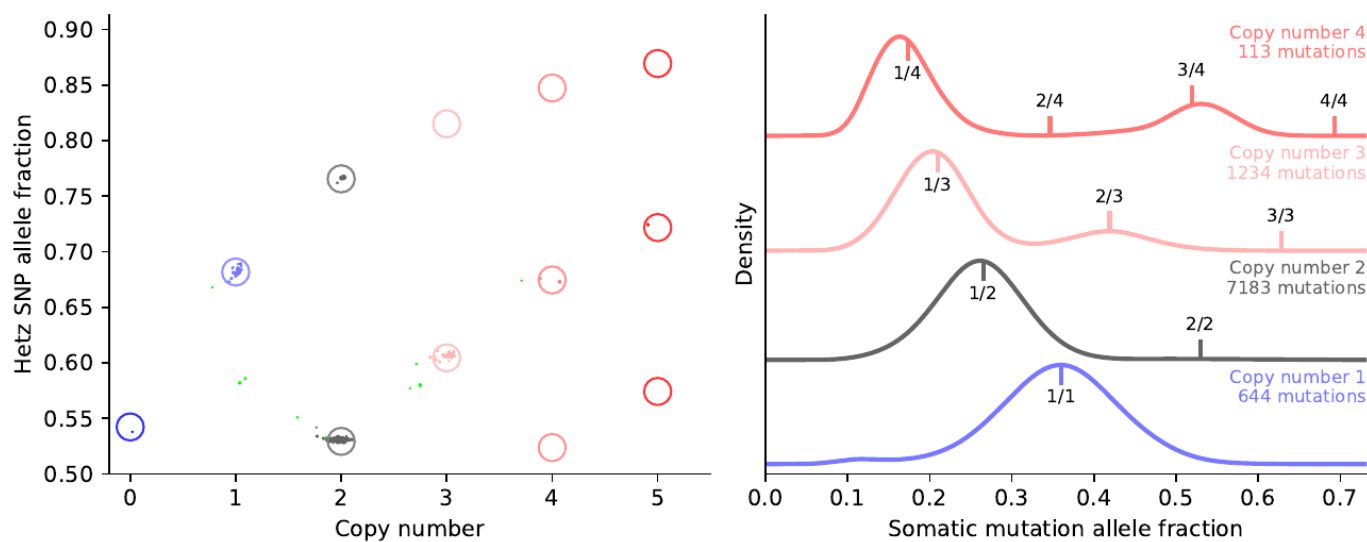
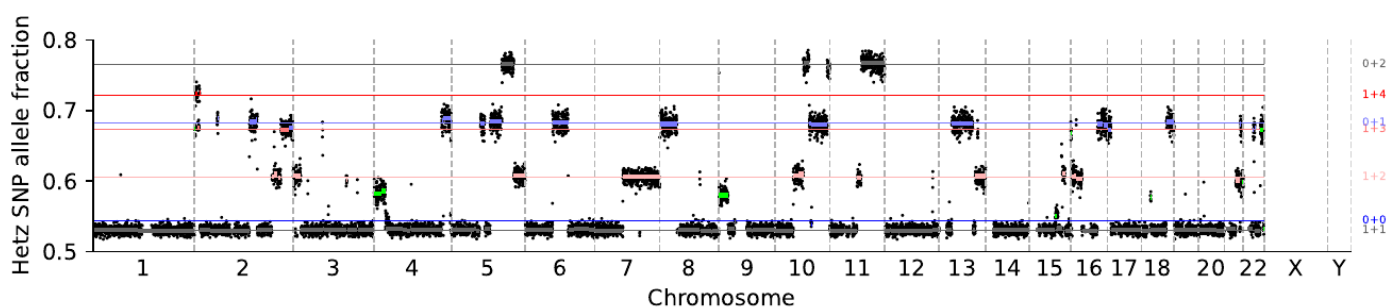
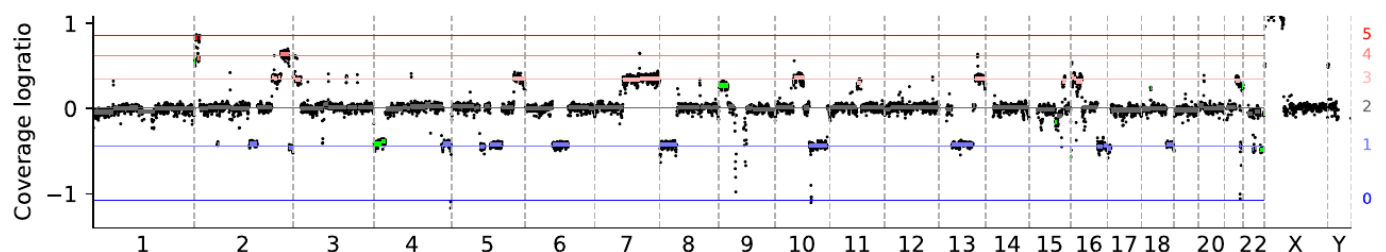
DTB-119 mutation clustering

Patient DTB-119 had samples that were admixtures of cancer cell subpopulations with different whole genome duplication statuses (i.e. subclonal heterogeneity with respect to presence of whole genome duplication). Please see Supplementary Information section "Detection of subclonal whole genome duplication events" for a detailed subclonal reconstruction.

Patient DTB-129

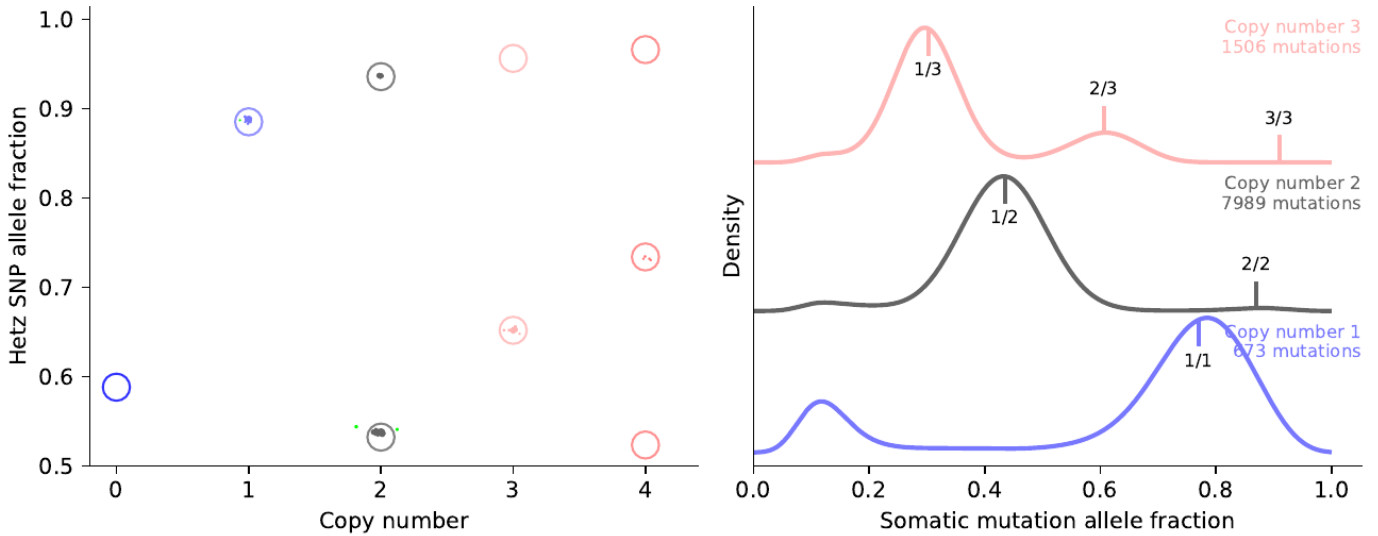
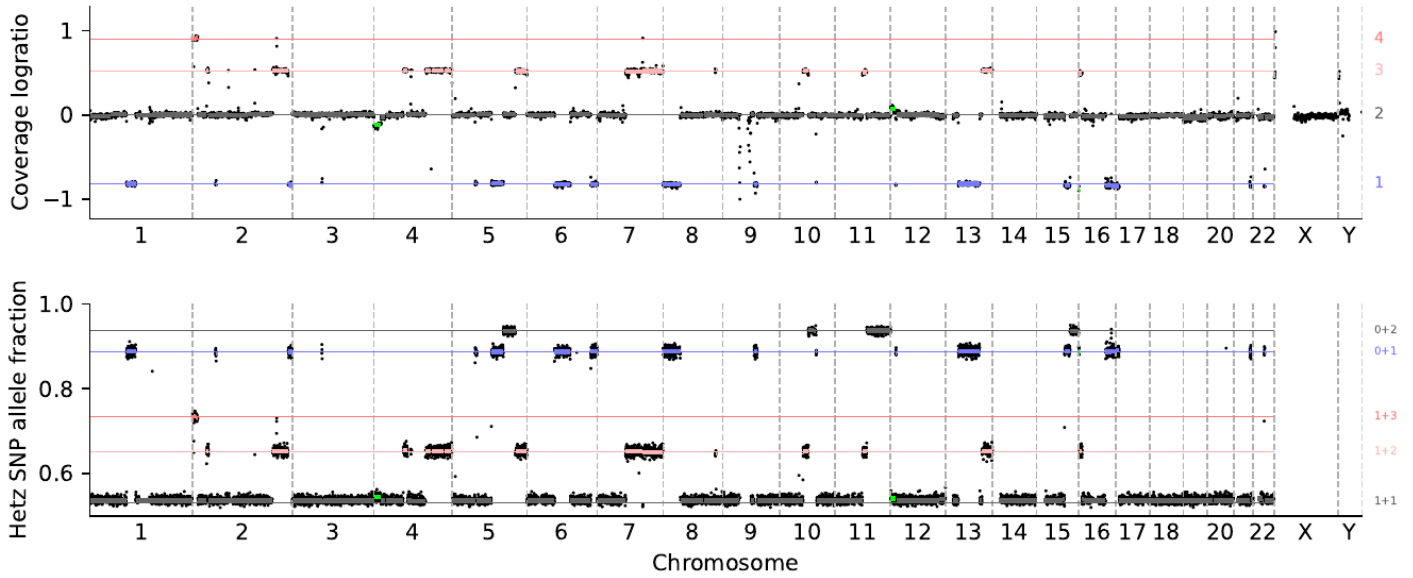
DTB-129-1st-Progression-cfDNA

Cancer fraction: 53%, Diploid level: 0.01

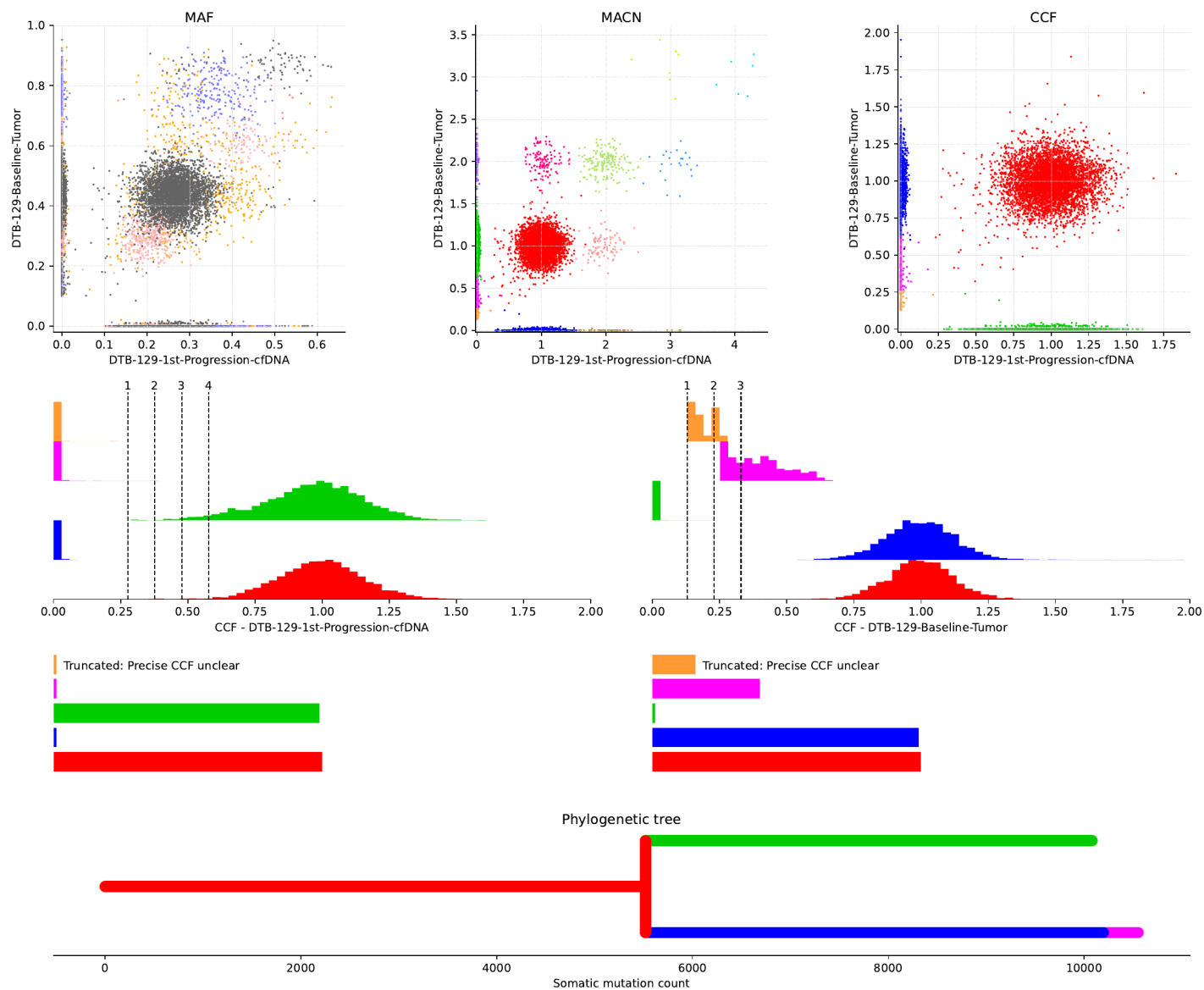


DTB-129-Baseline-Tumor

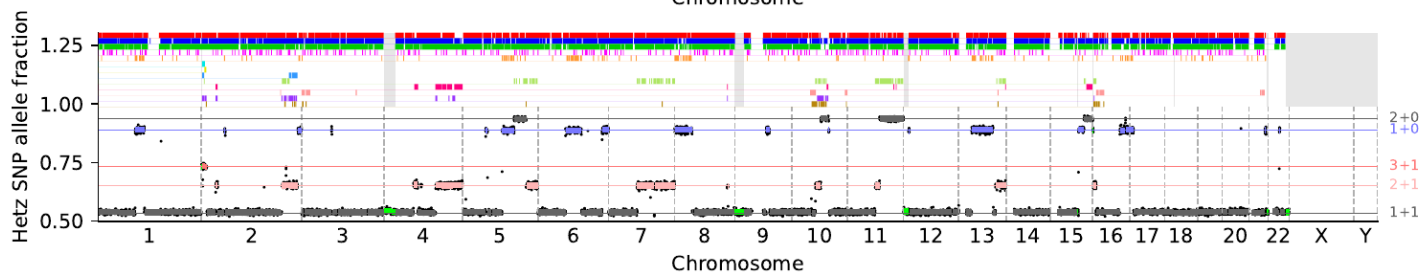
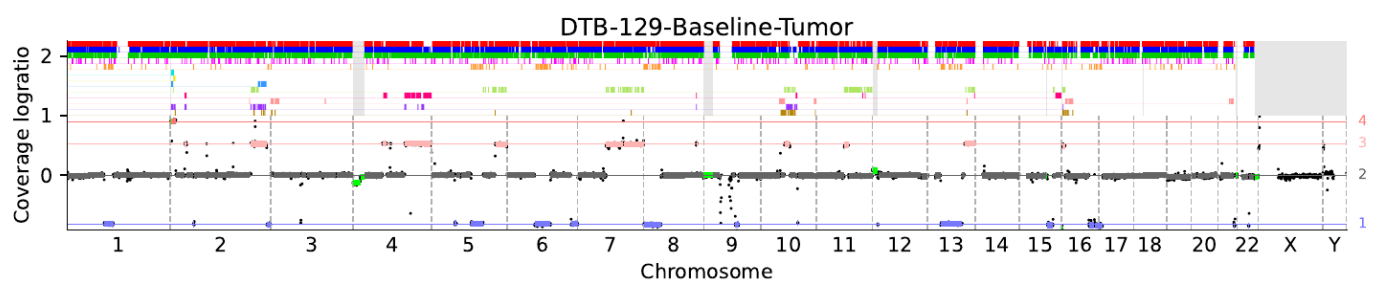
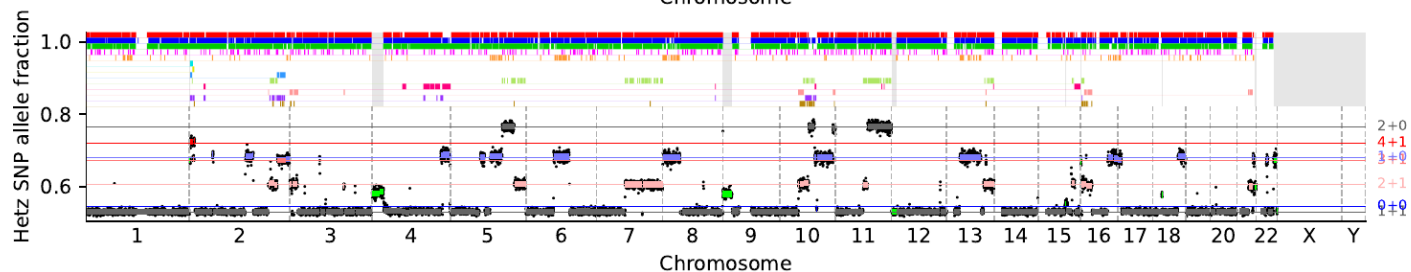
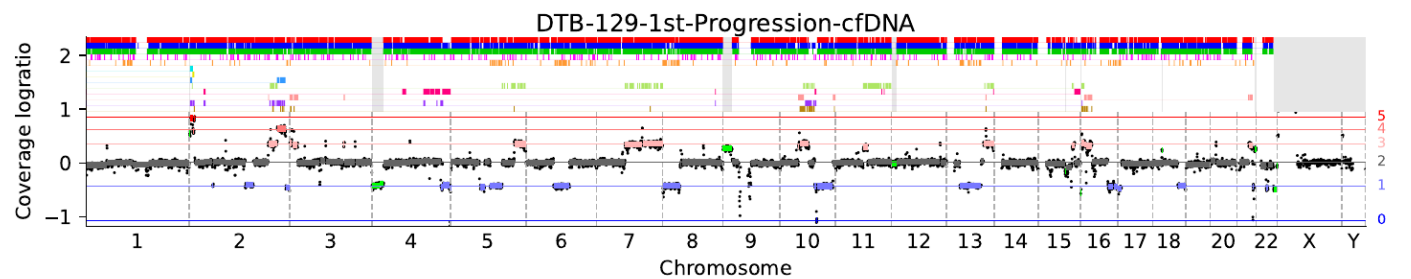
Cancer fraction: 87%, Diploid level: 0.00



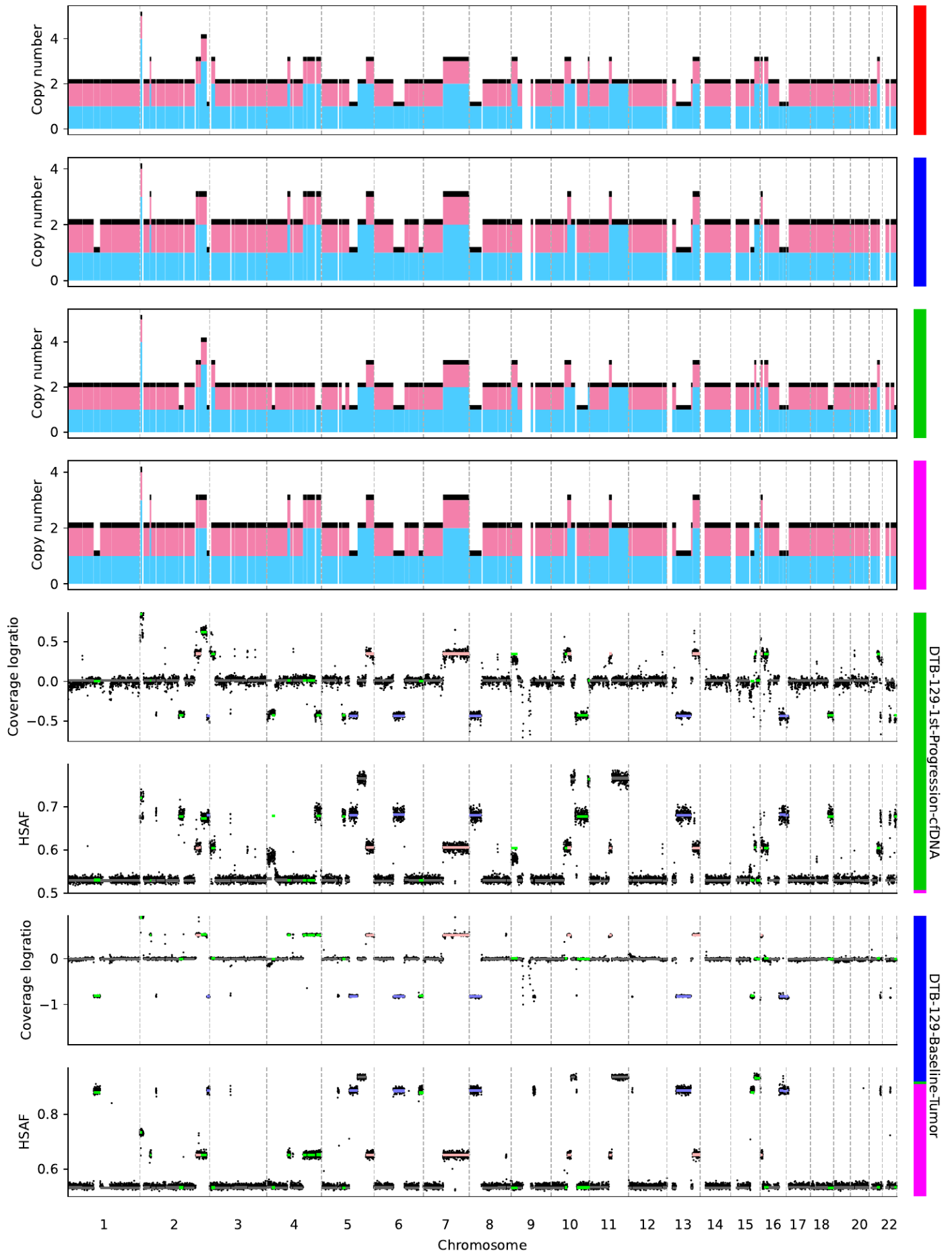
DTB-129 mutation clustering



DTB-129 mutation cluster positions



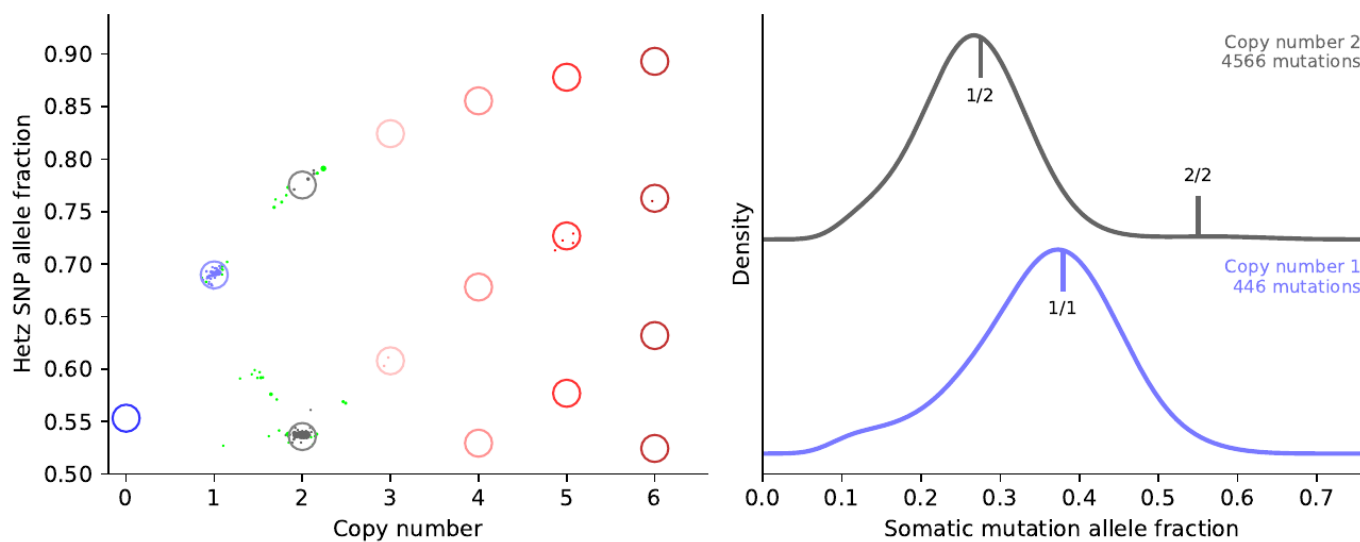
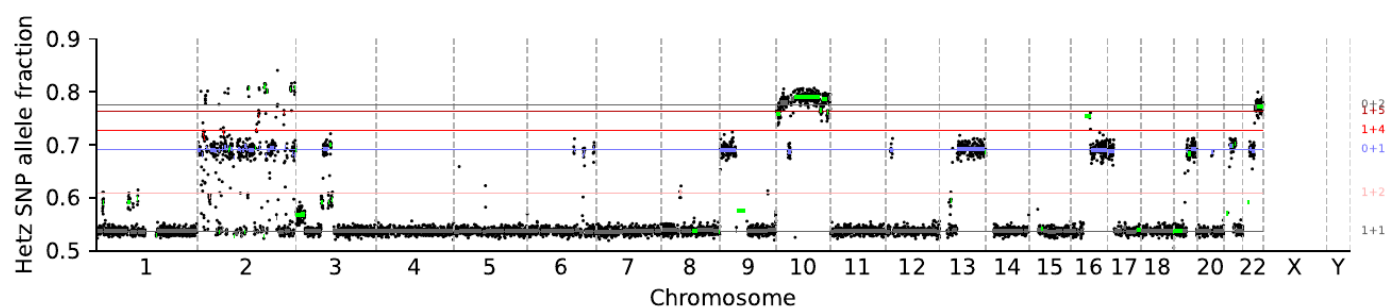
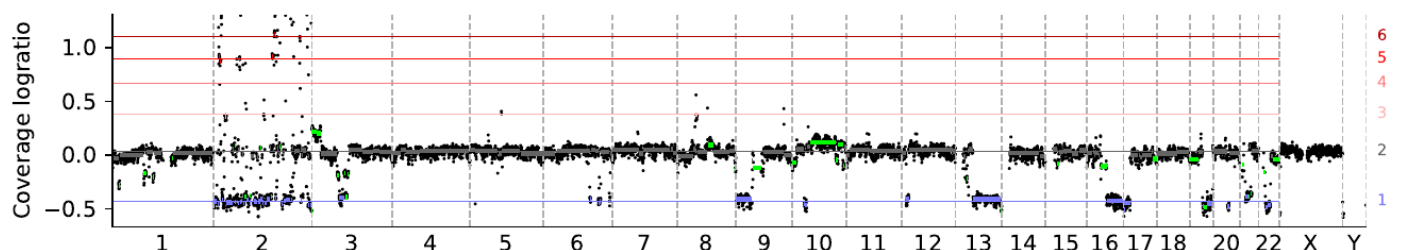
DTB-129 copy number deconvolution



Patient DTB-149

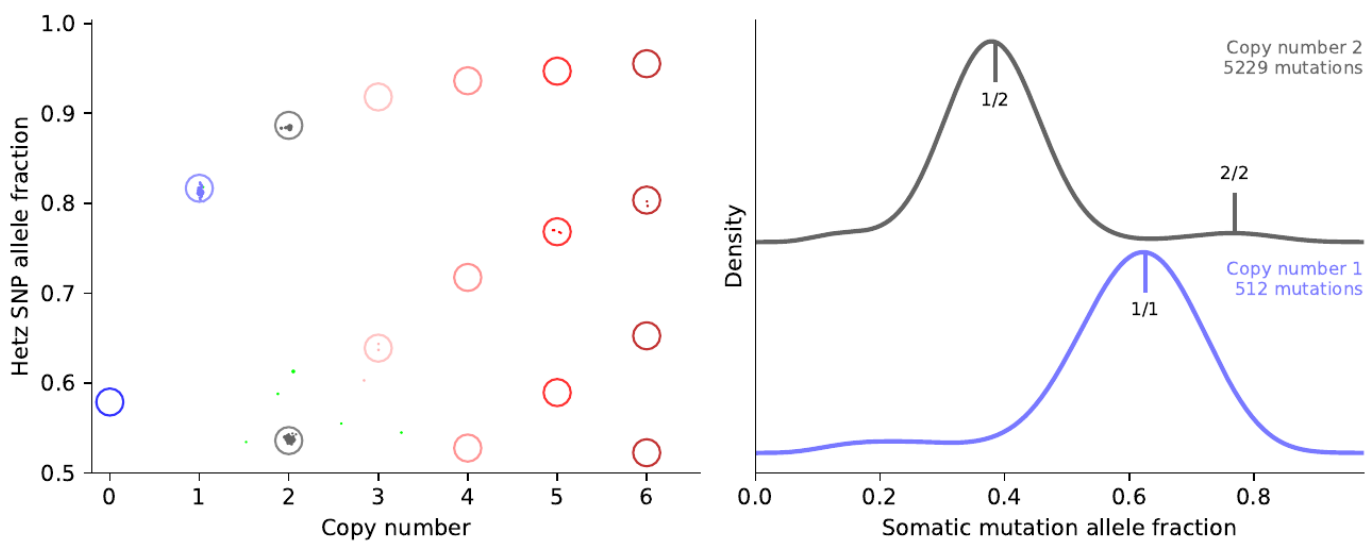
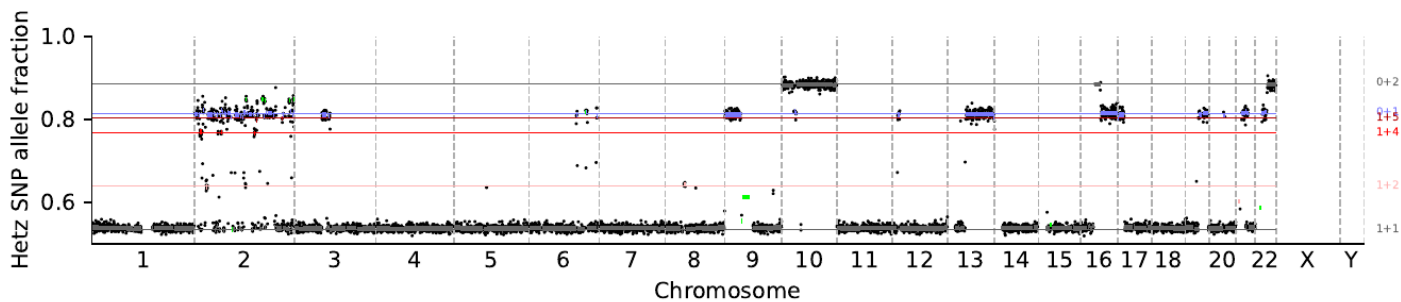
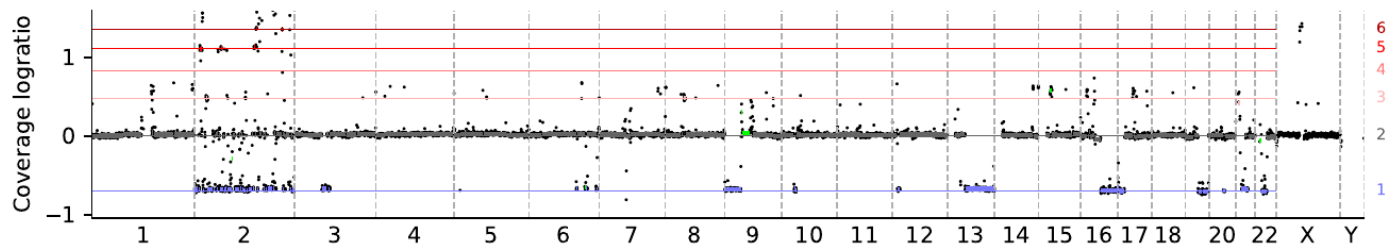
DTB-149-1st-Progression-cfDNA

Cancer fraction: 55%, Diploid level: 0.03



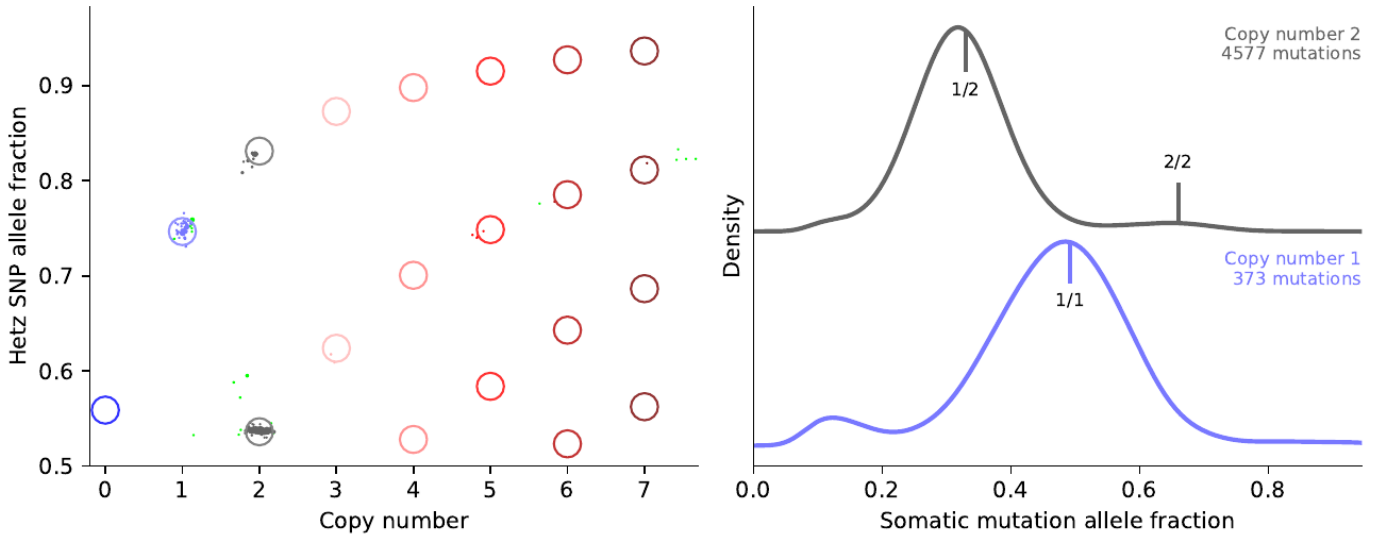
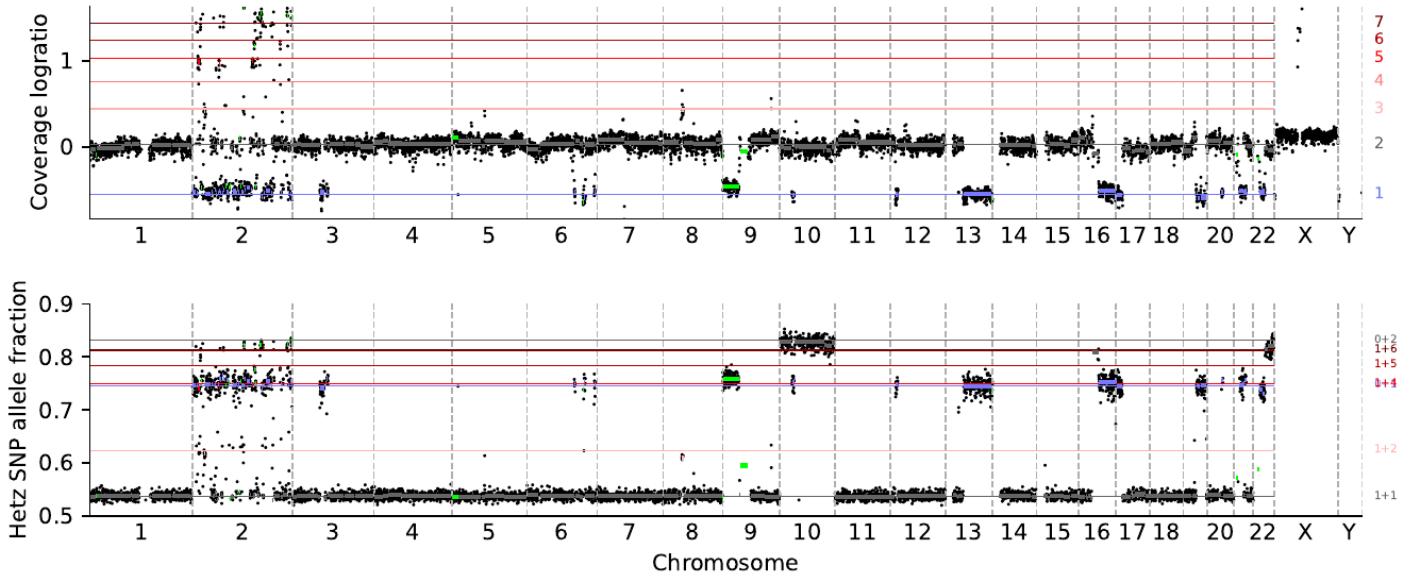
DTB-149-Baseline-Tumor

Cancer fraction: 77%, Diploid level: 0.01

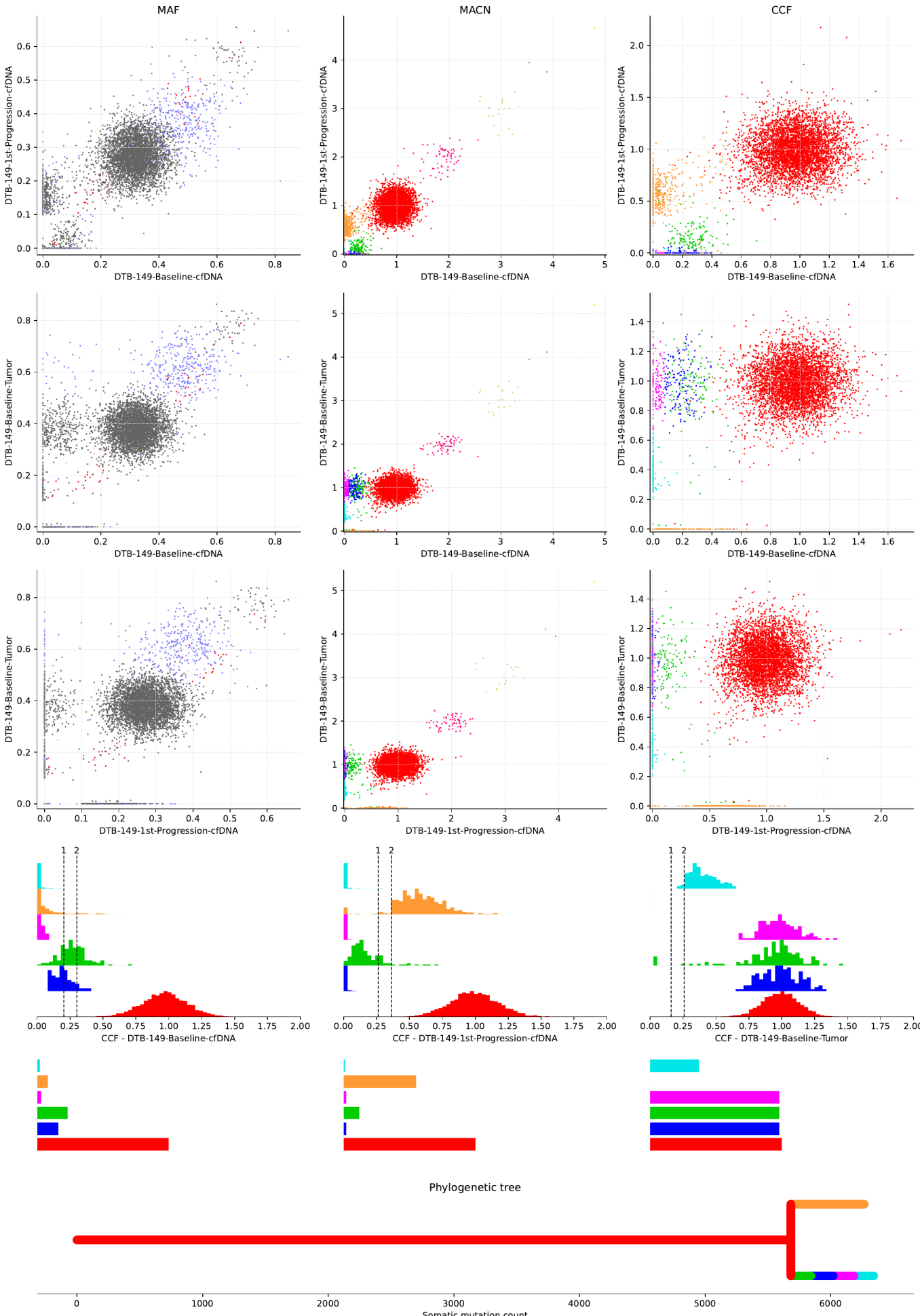


DTB-149-Baseline-cfDNA

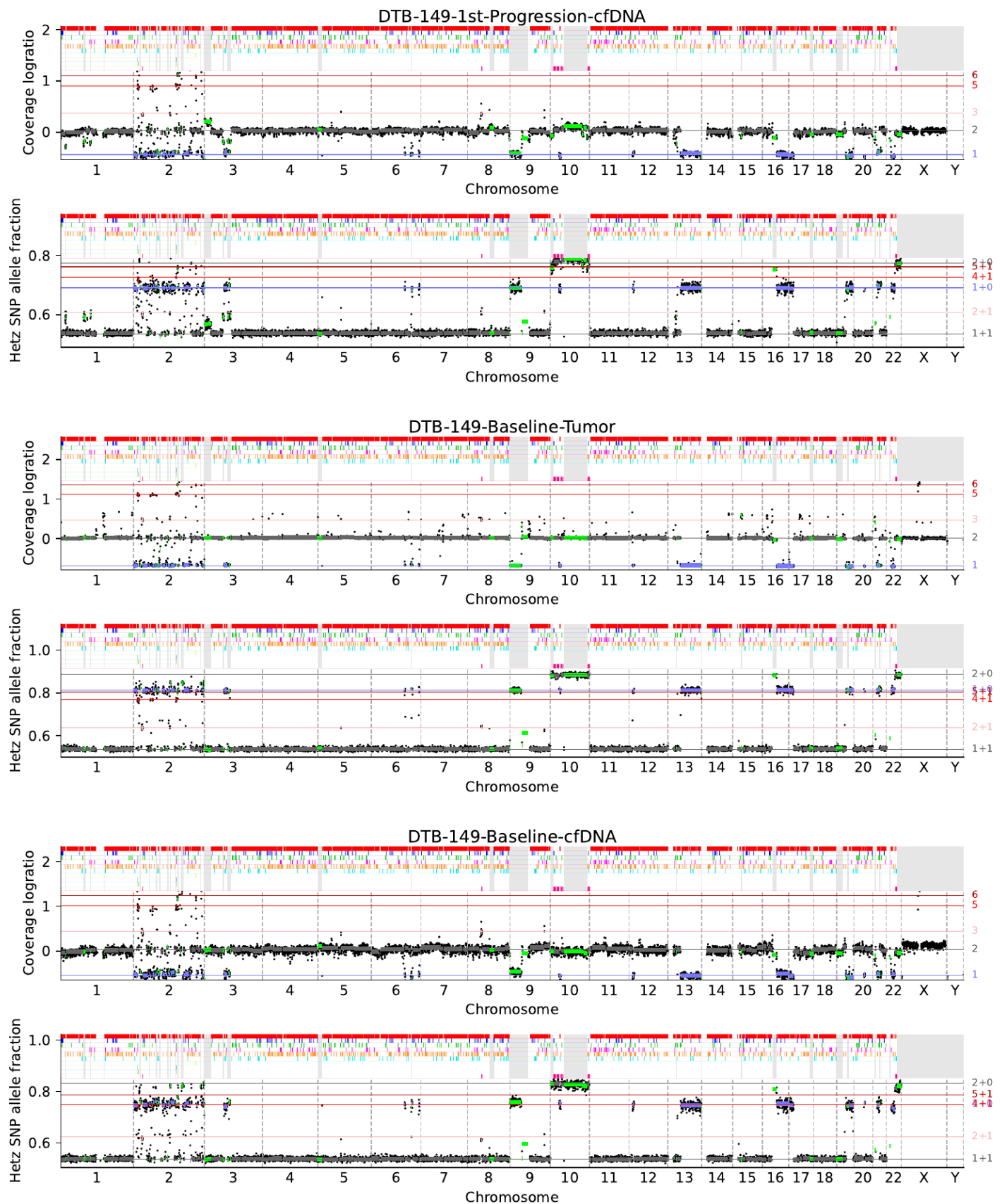
Cancer fraction: 66%, Diploid level: 0.03



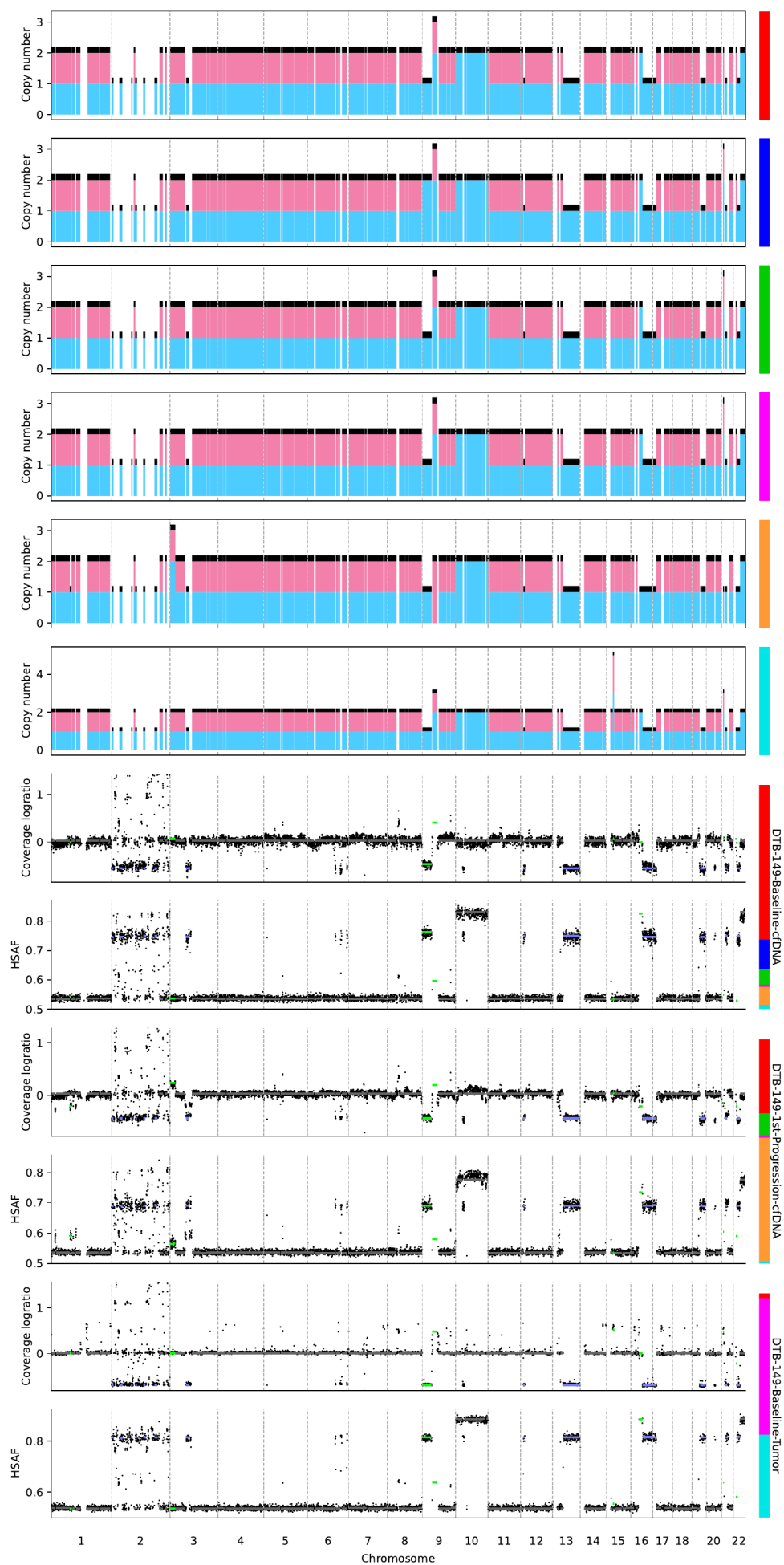
DTB-149 mutation clustering



DTB-149 mutation cluster positions



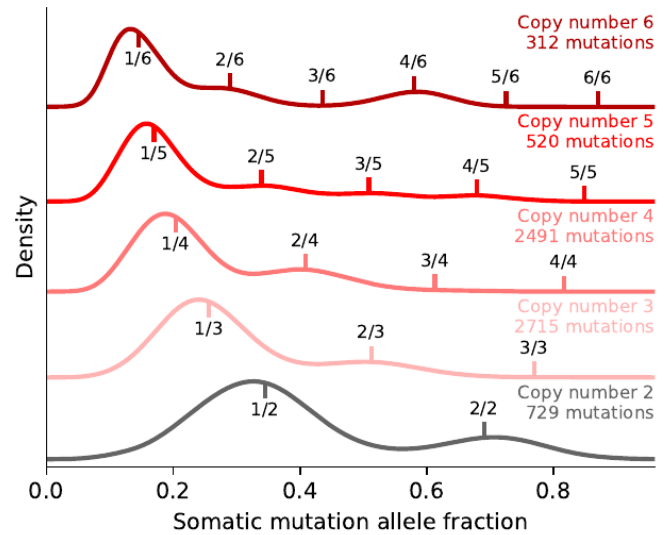
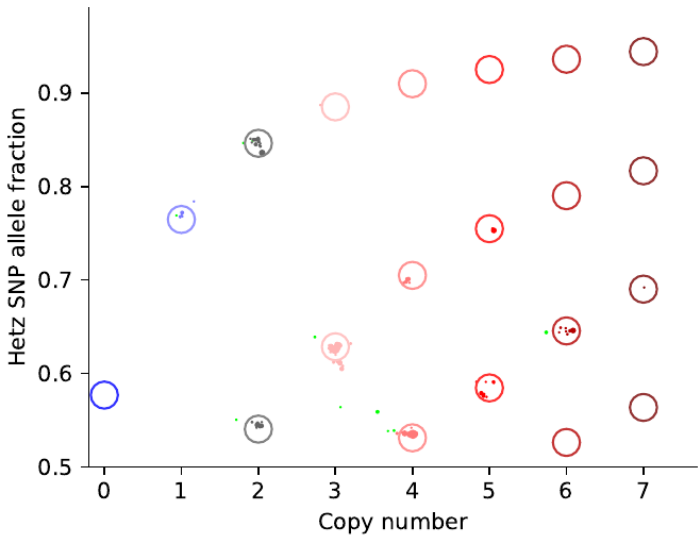
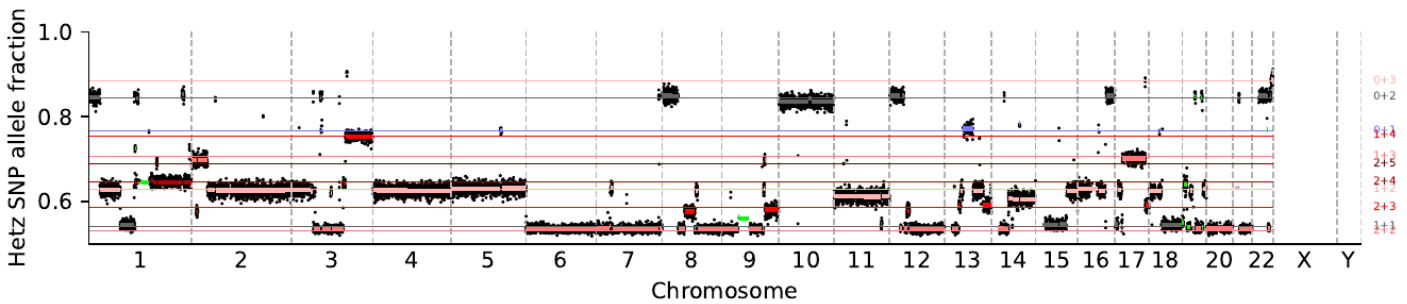
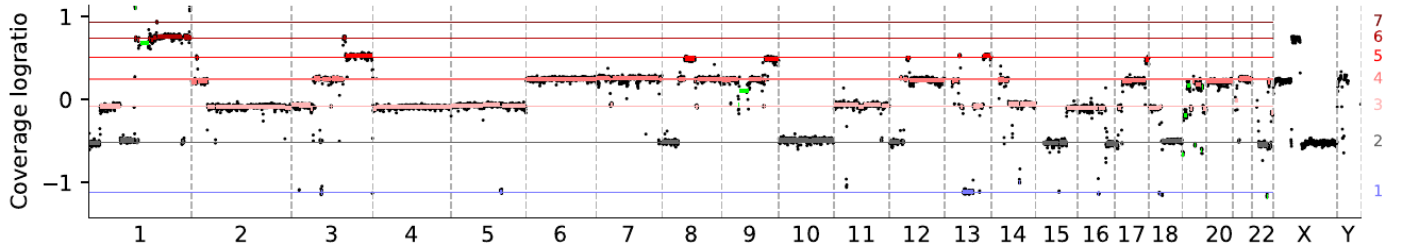
DTB-149 copy number deconvolution



Patient DTB-156

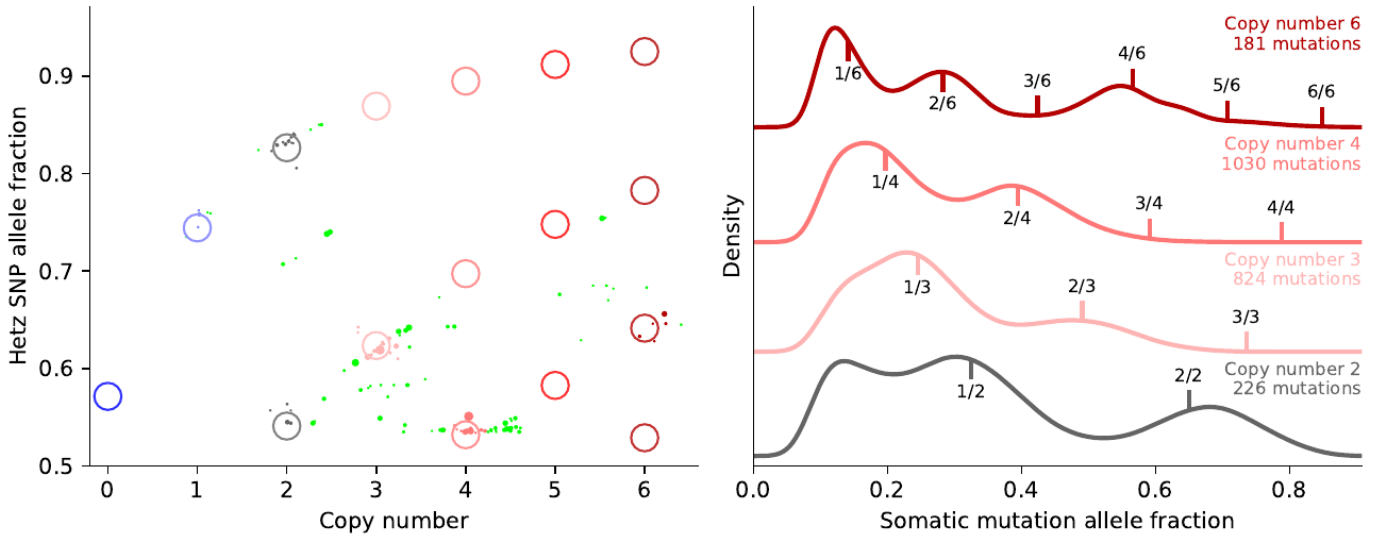
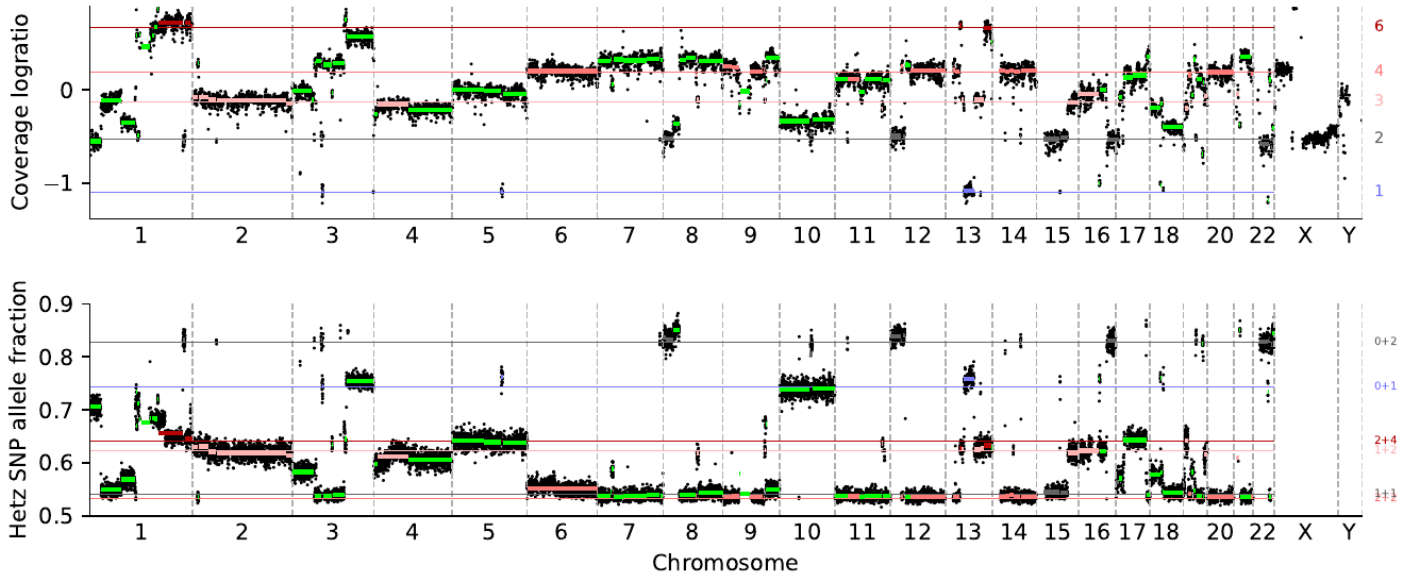
DTB-156-Baseline-Tumor

Cancer fraction: 69%, Diploid level: -0.51

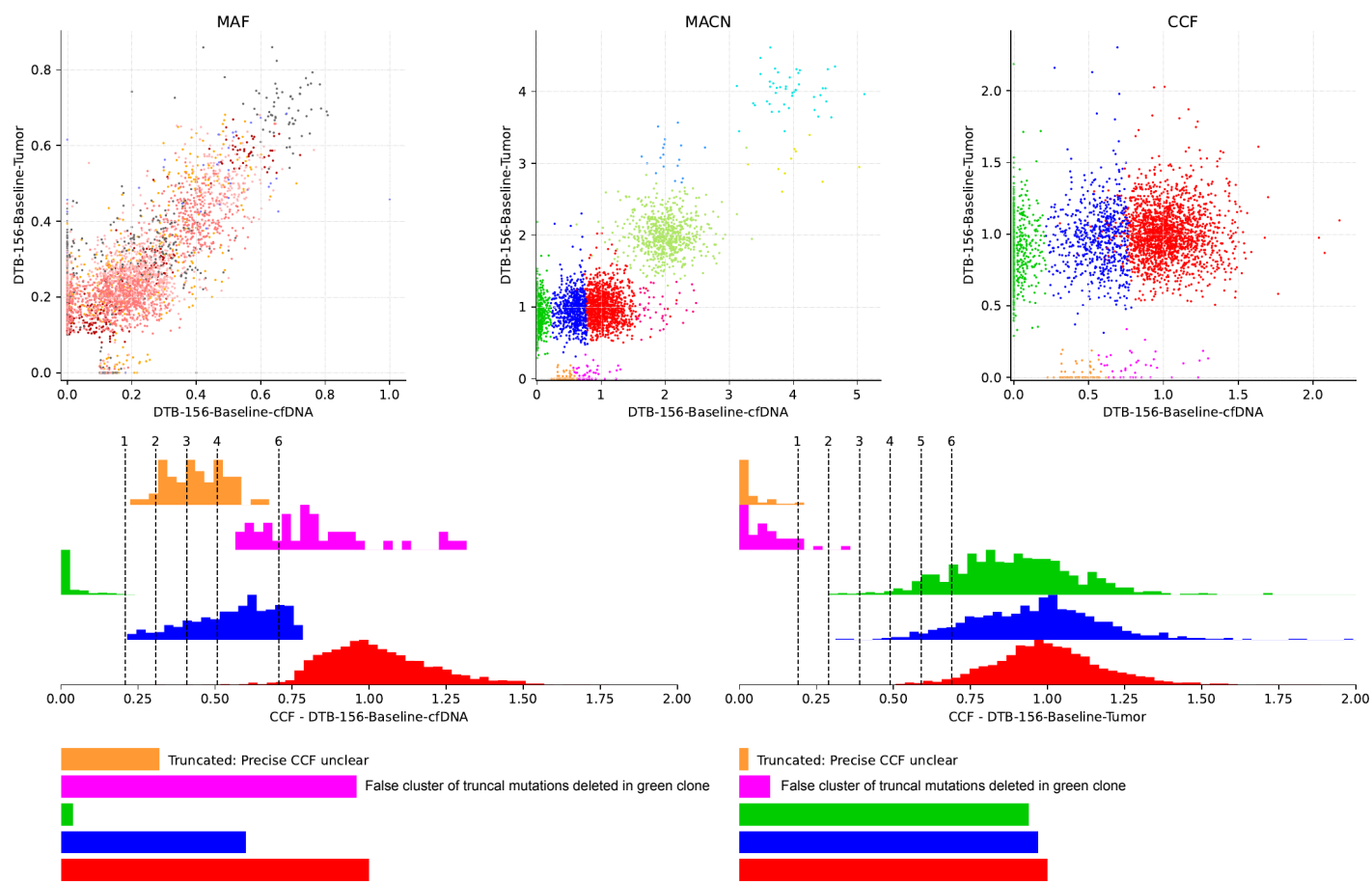


DTB-156-Baseline-cfDNA

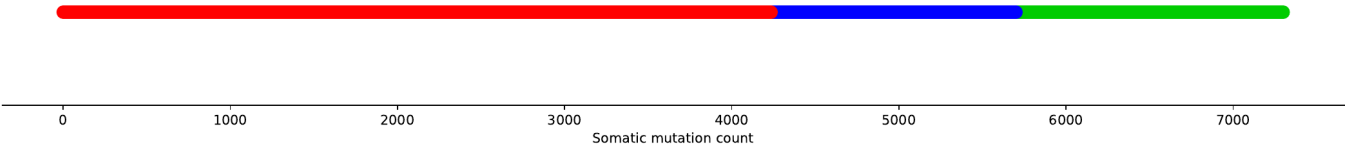
Cancer fraction: 65%, Diploid level: -0.53



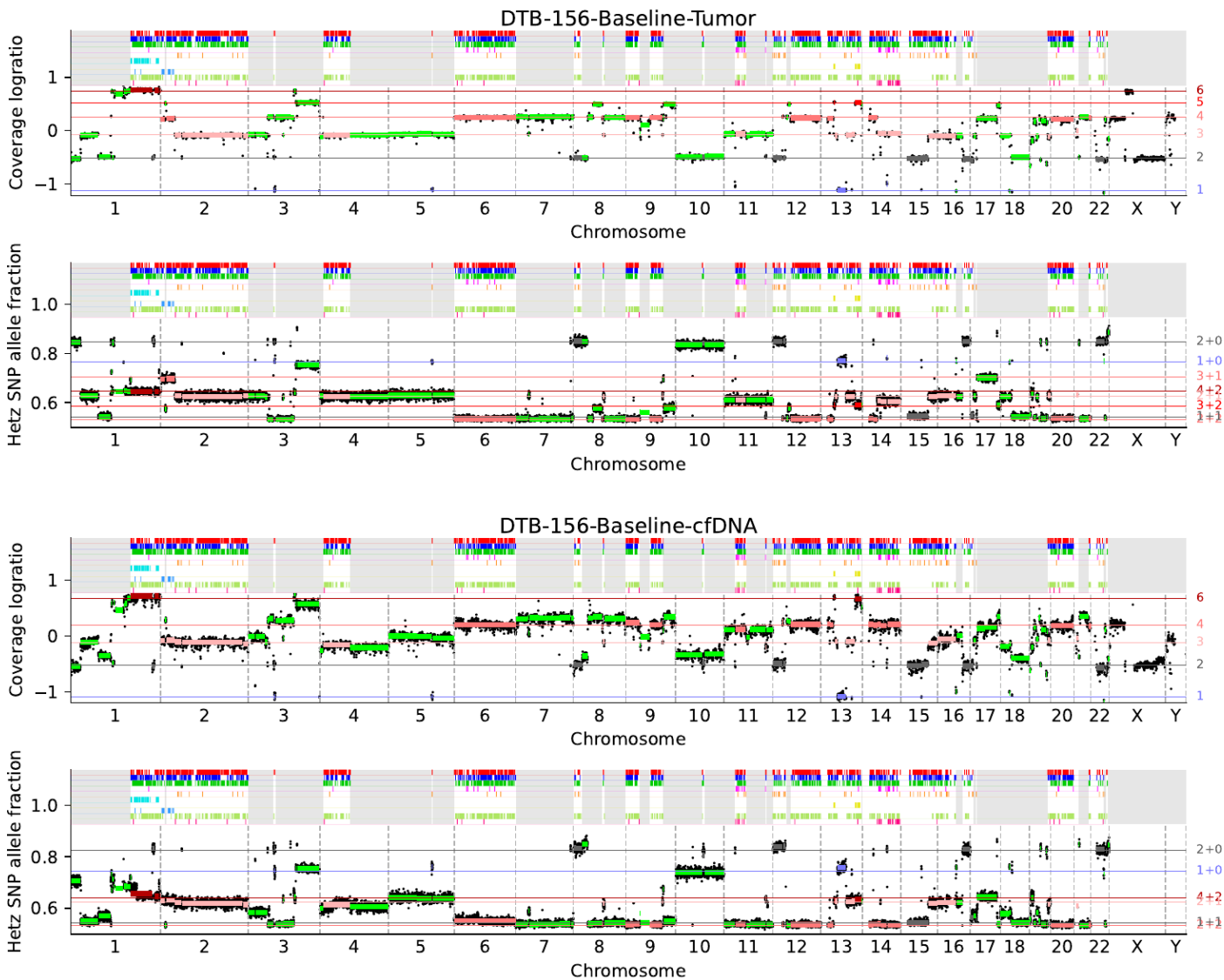
DTB-156 mutation clustering



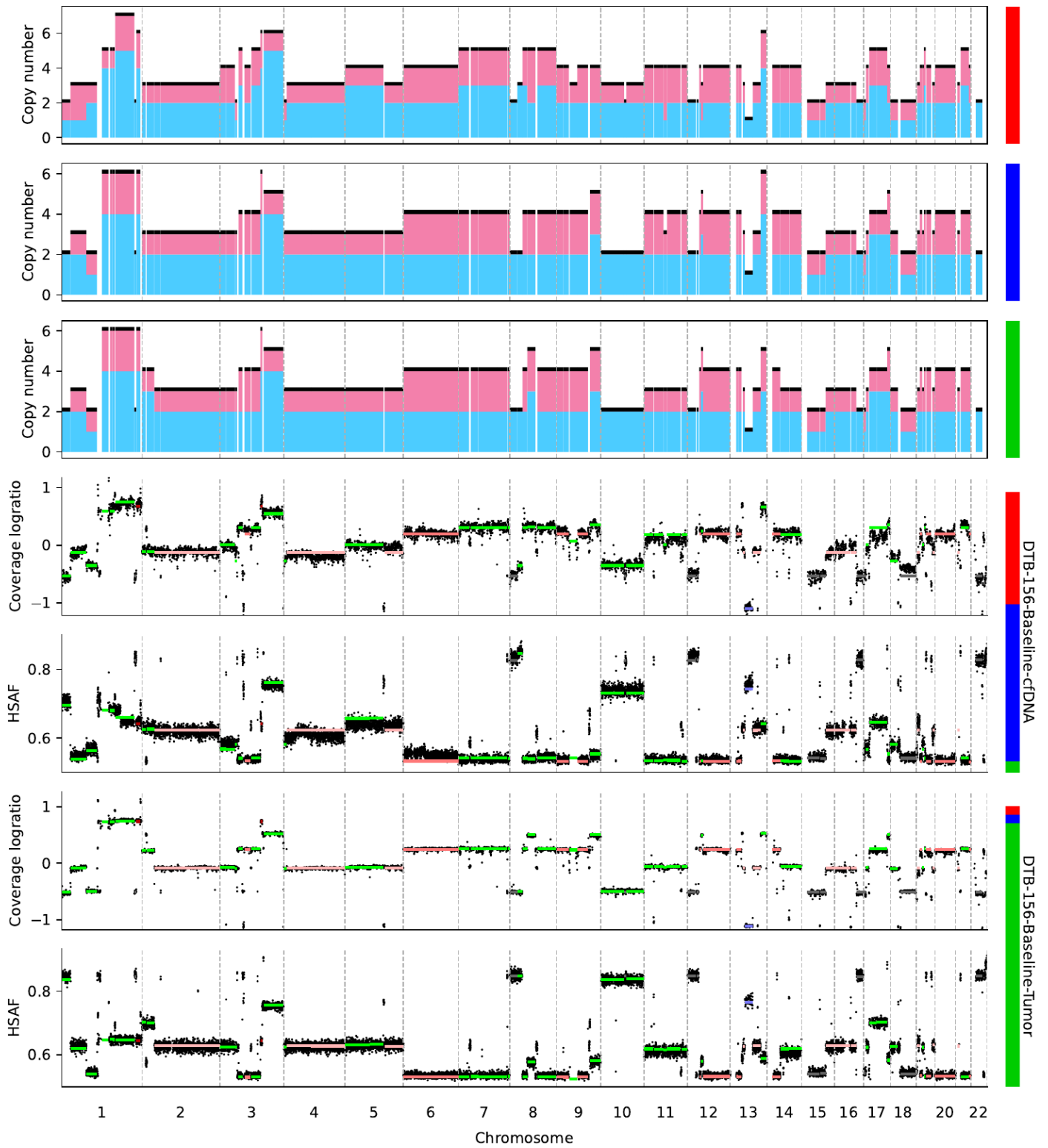
Phylogenetic tree



DTB-156 mutation cluster positions



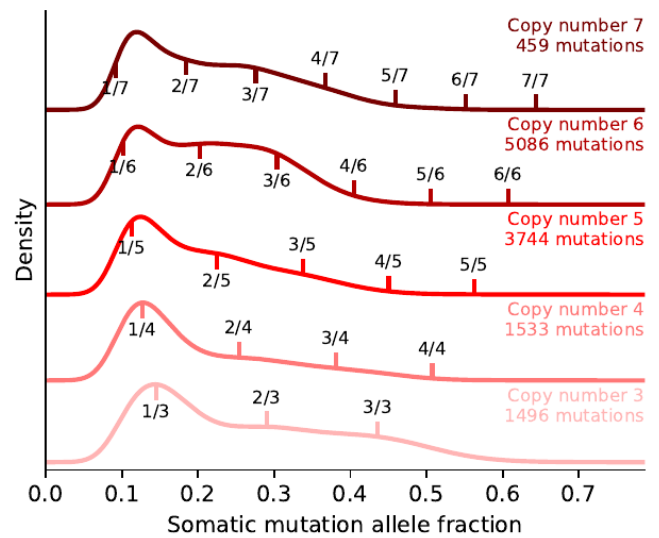
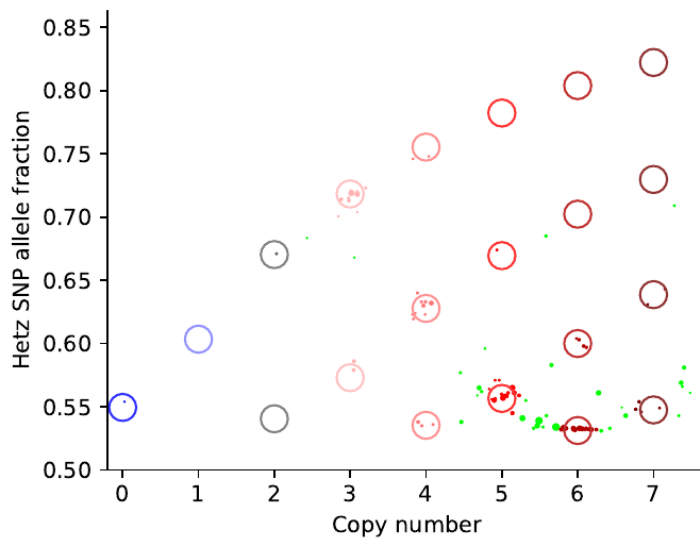
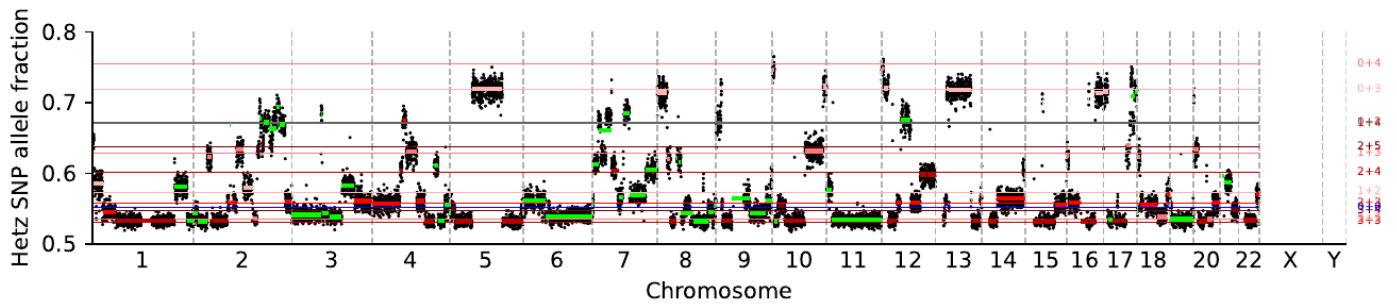
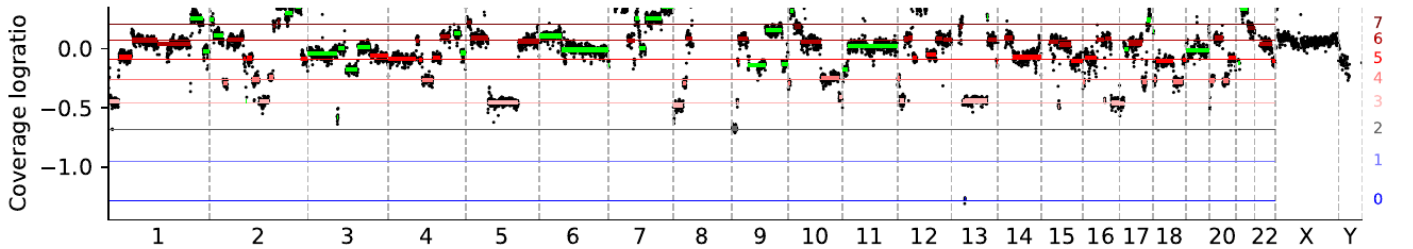
DTB-156 copy number deconvolution



Patient DTB-174

DTB-174-Baseline-cfDNA

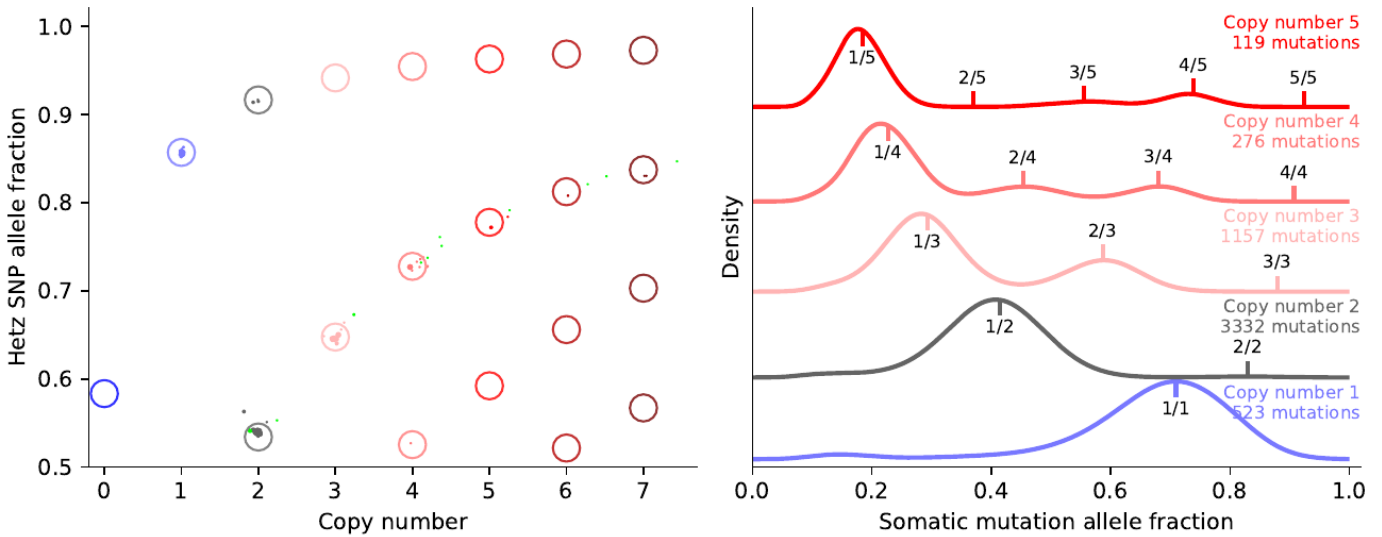
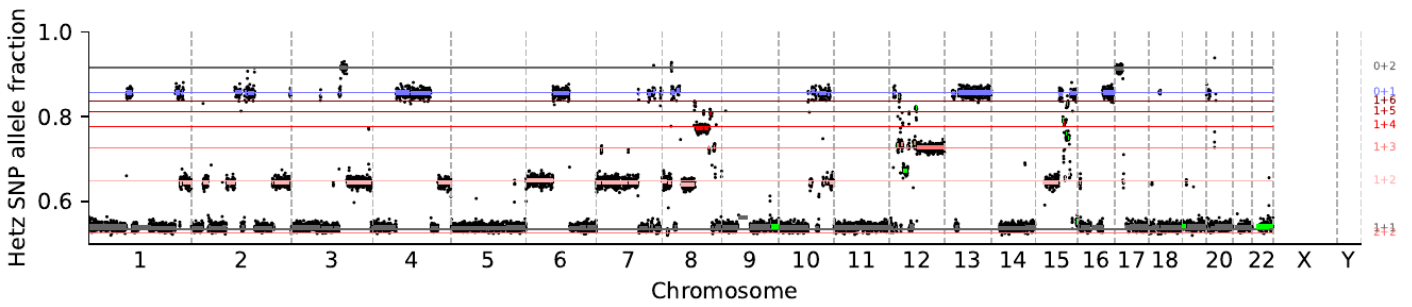
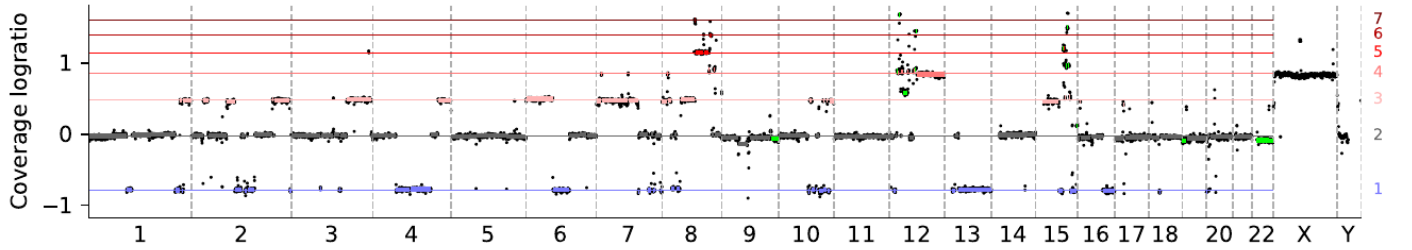
Cancer fraction: 34%, Diploid level: -0.68



Patient DTB-175

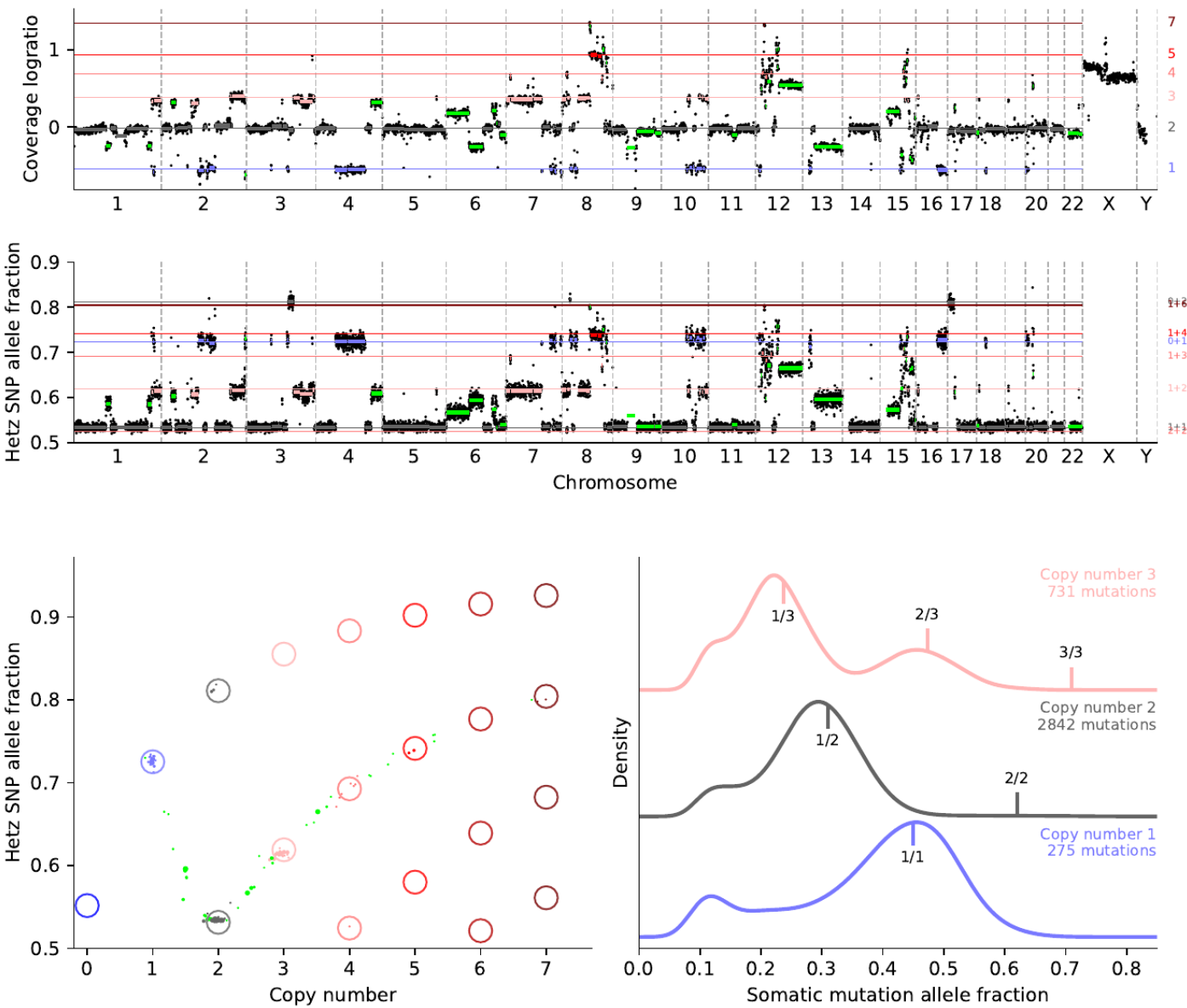
DTB-175-Baseline-Tumor

Cancer fraction: 83%, Diploid level: -0.02

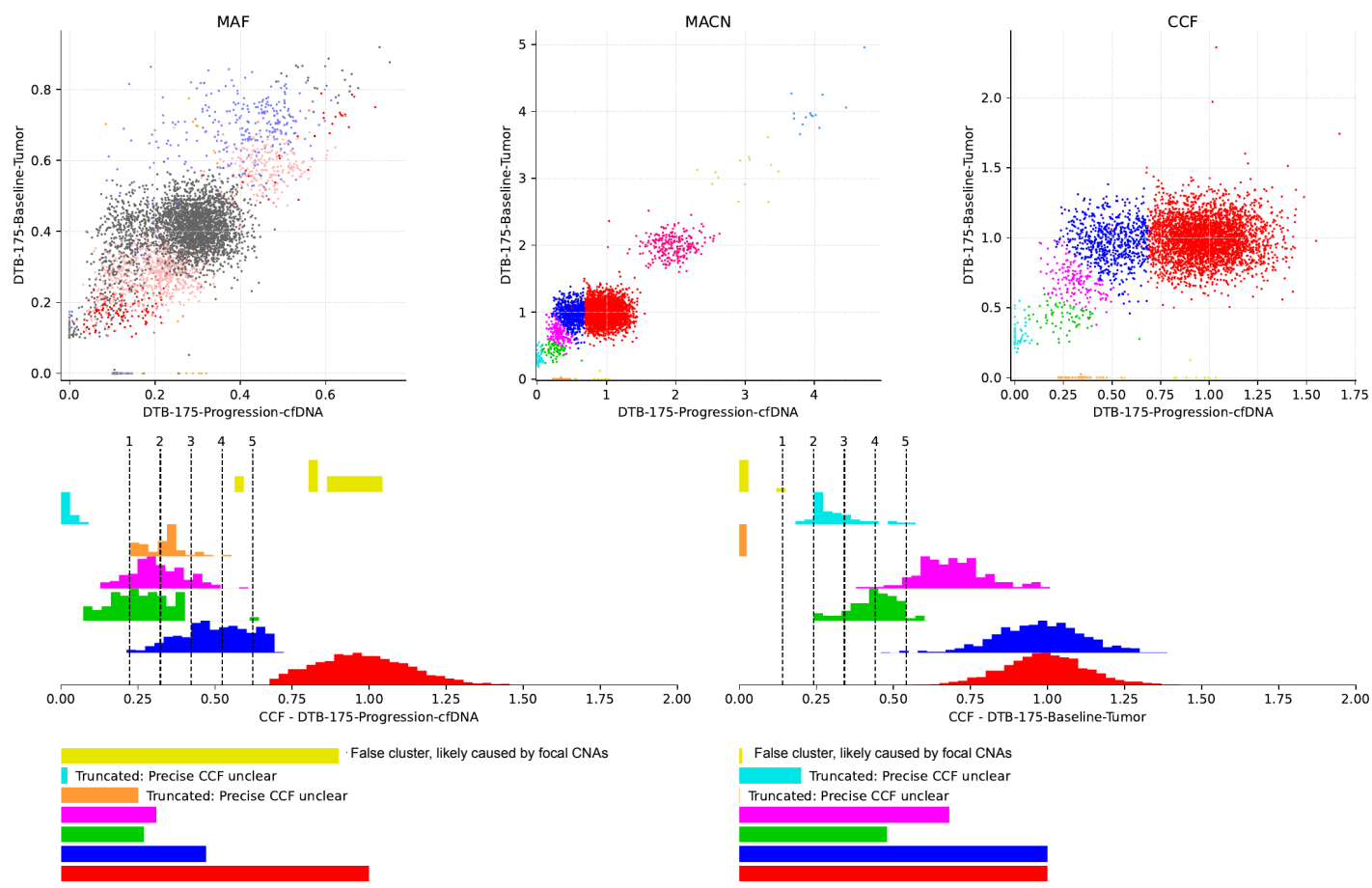


DTB-175-Progression-cfDNA

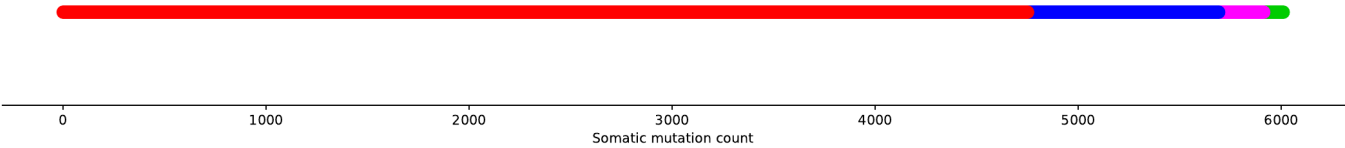
Cancer fraction: 62%, Diploid level: -0.01



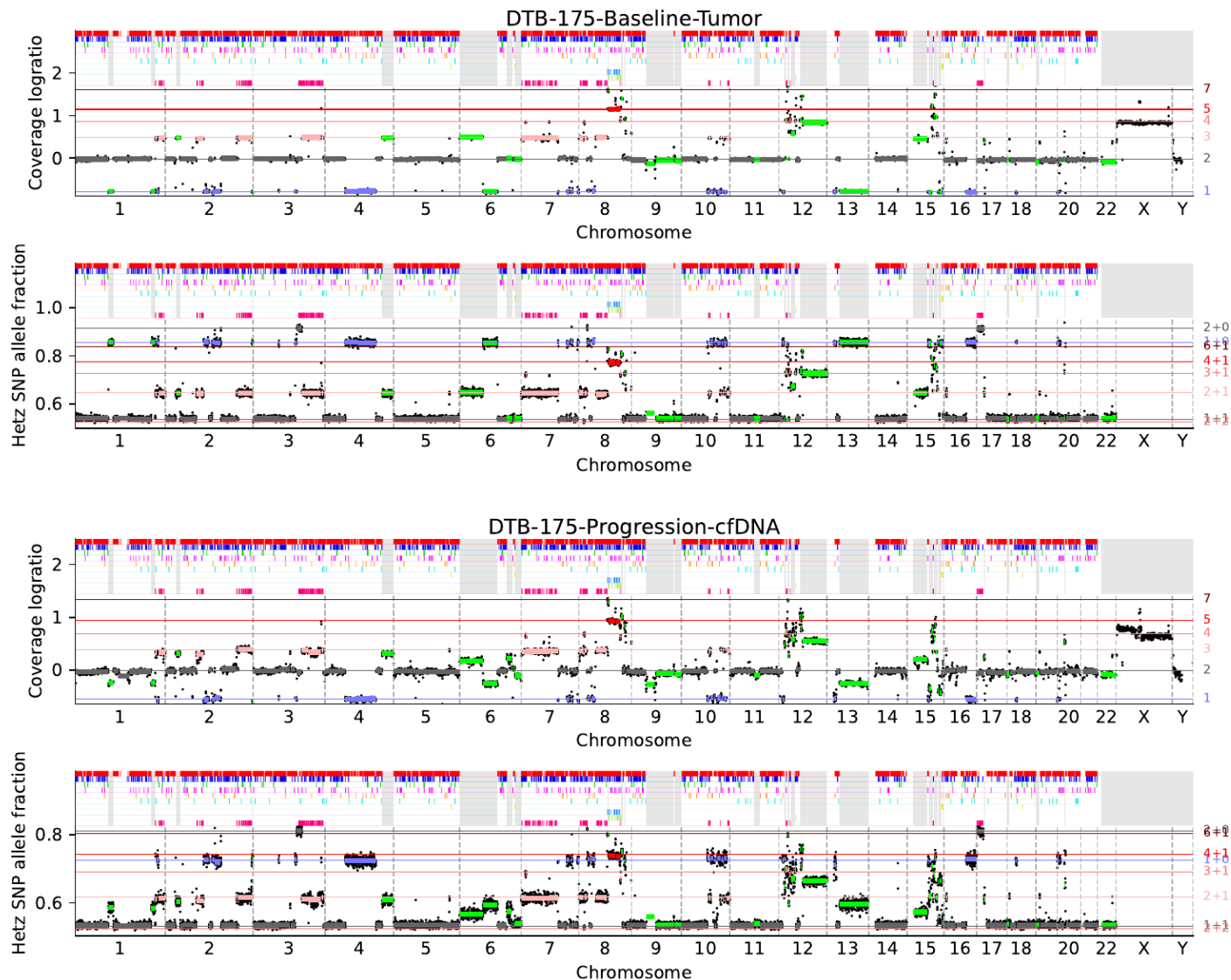
DTB-175 mutation clustering



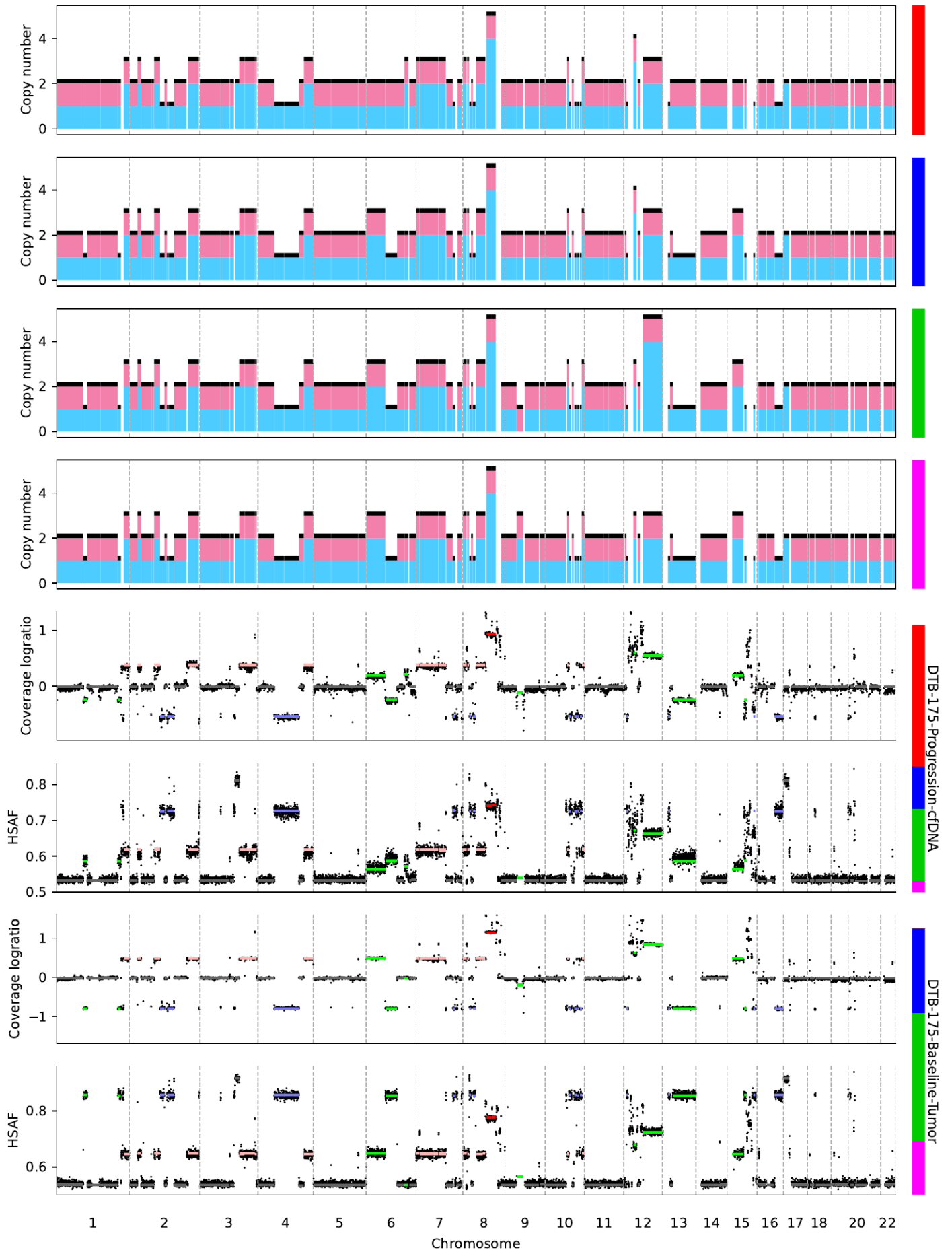
Phylogenetic tree



DTB-175 mutation cluster positions



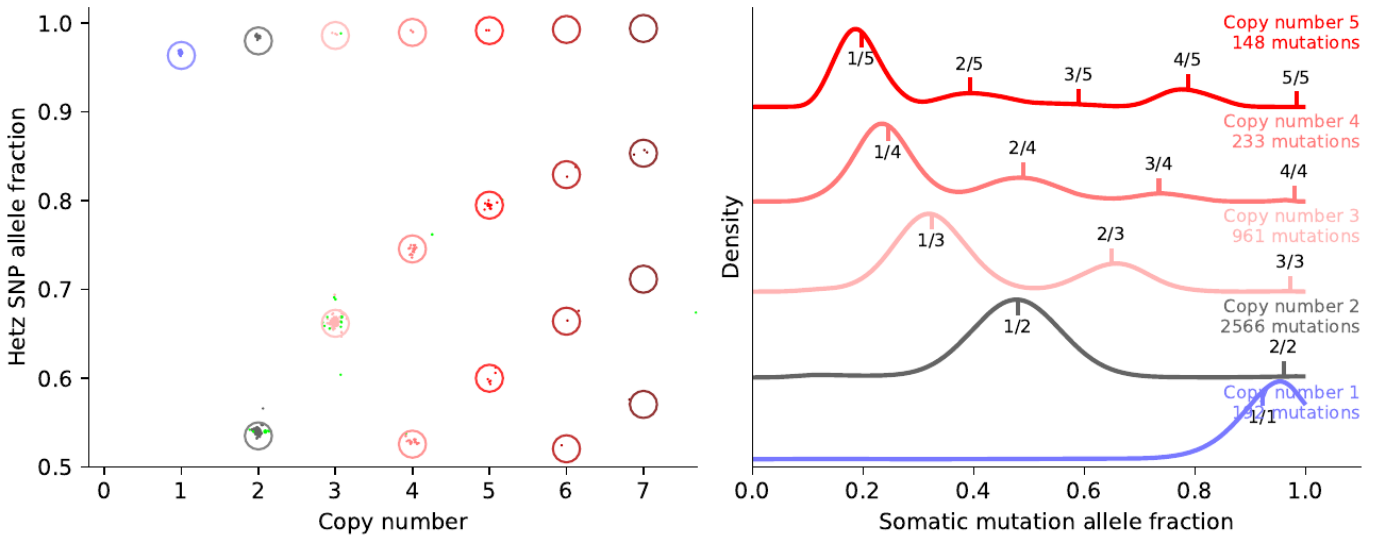
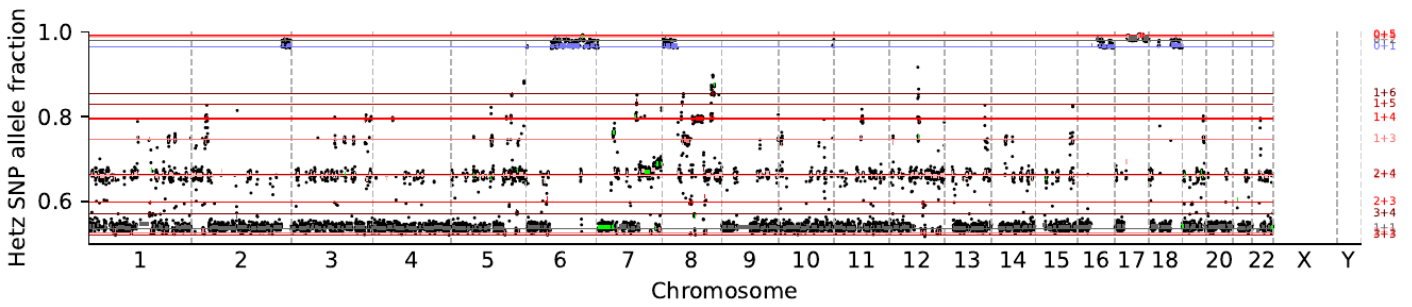
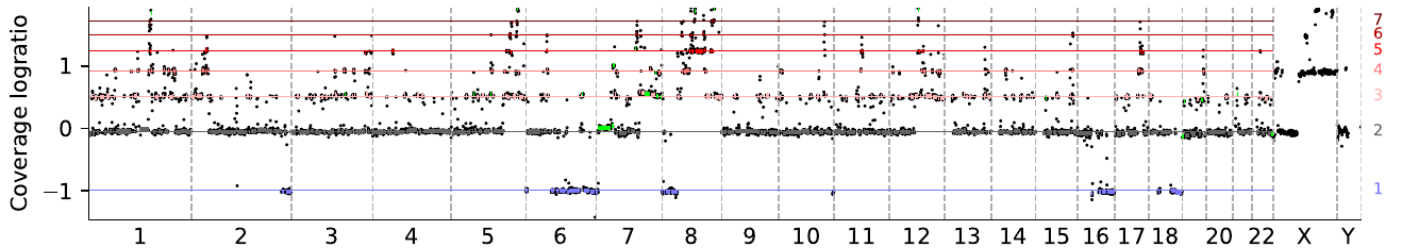
DTB-175 copy number deconvolution



Patient DTB-183

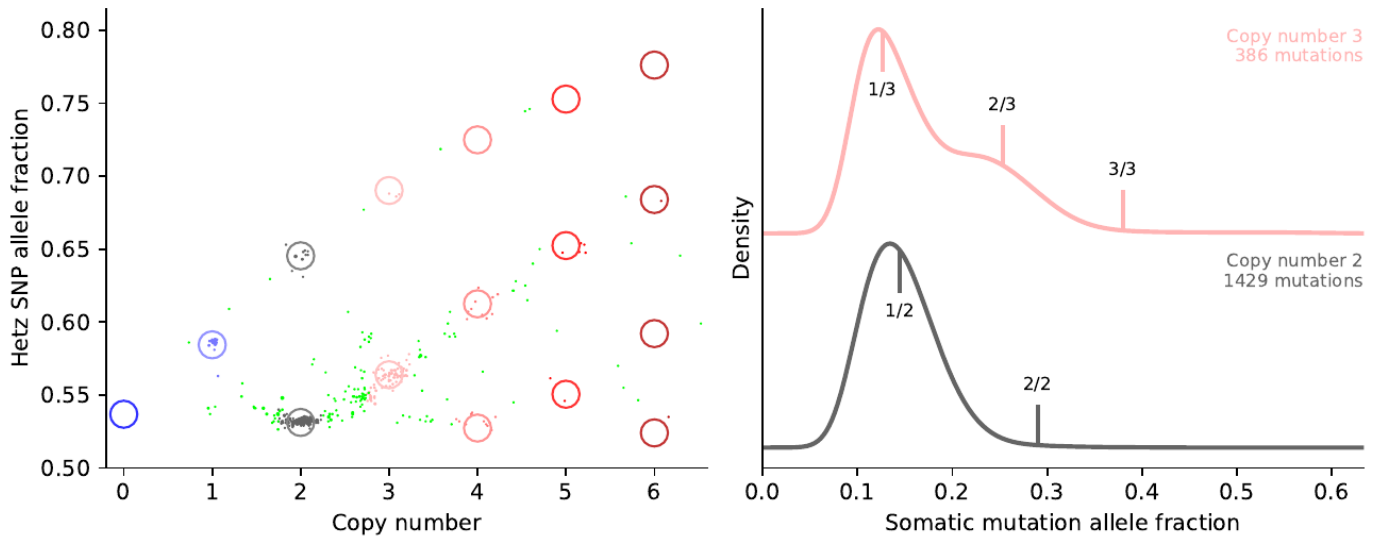
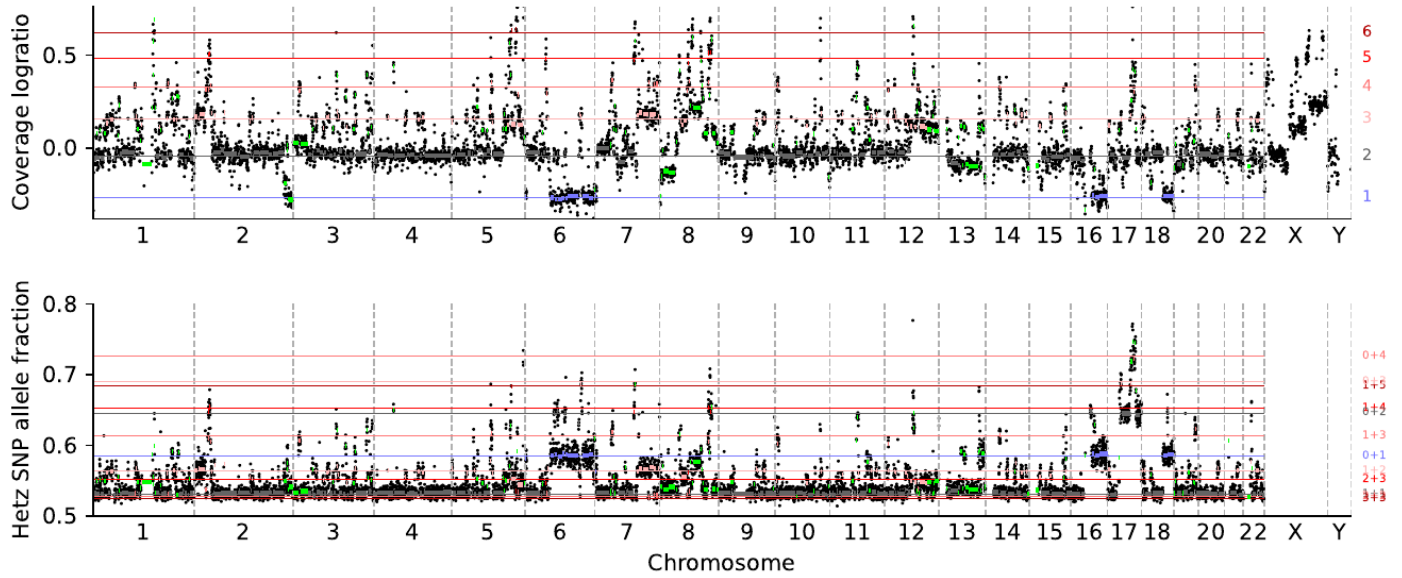
DTB-183-Baseline-Tumor

Cancer fraction: 96%, Diploid level: -0.05

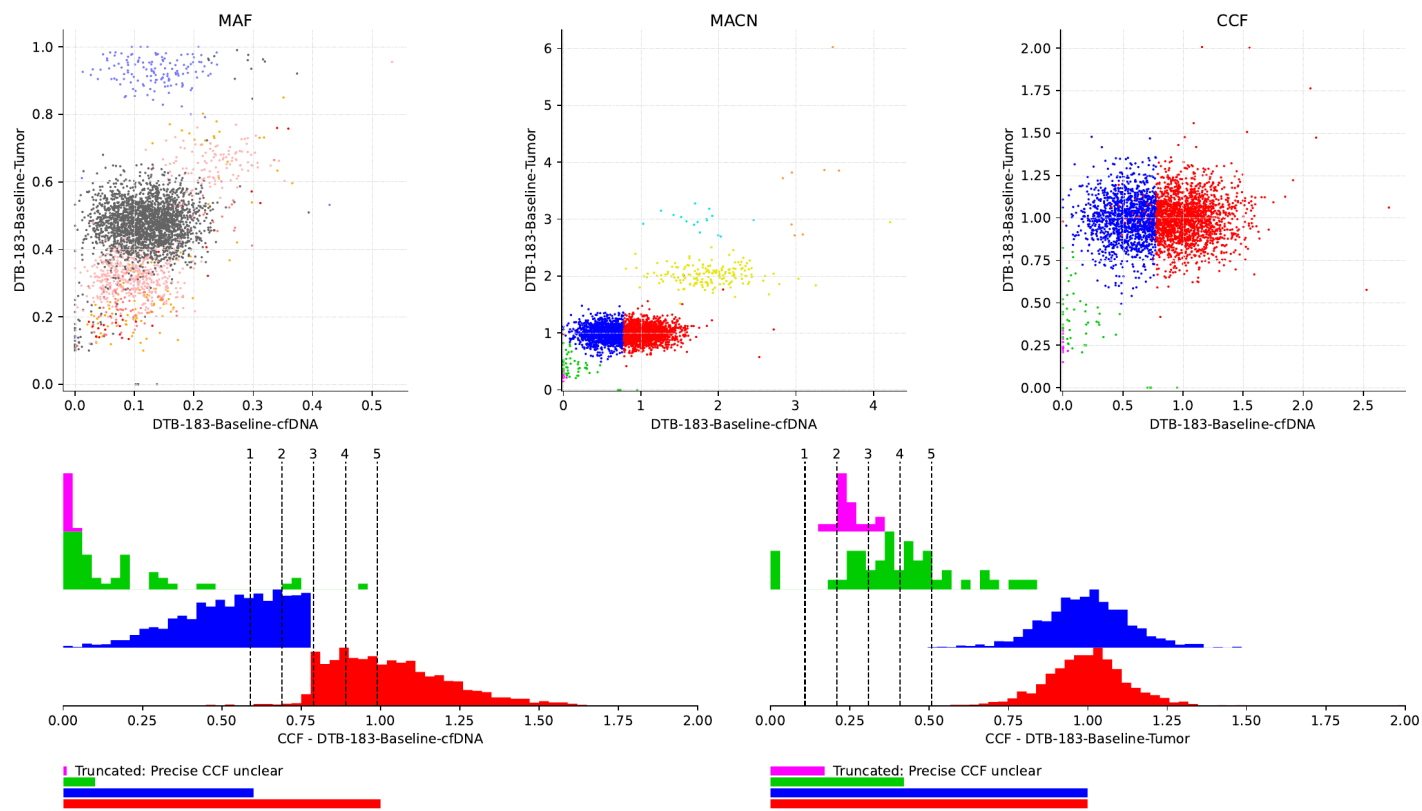


DTB-183-Baseline-cfDNA

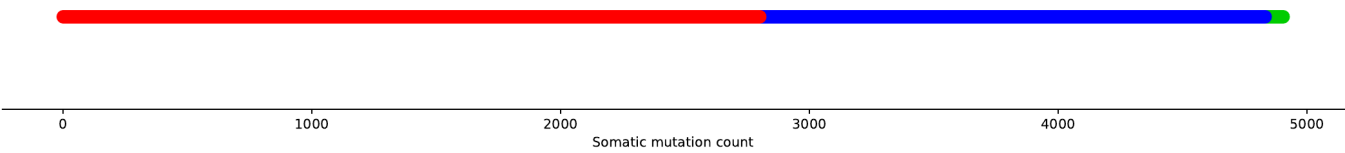
Cancer fraction: 29%, Diploid level: -0.04



DTB-183 mutation clustering

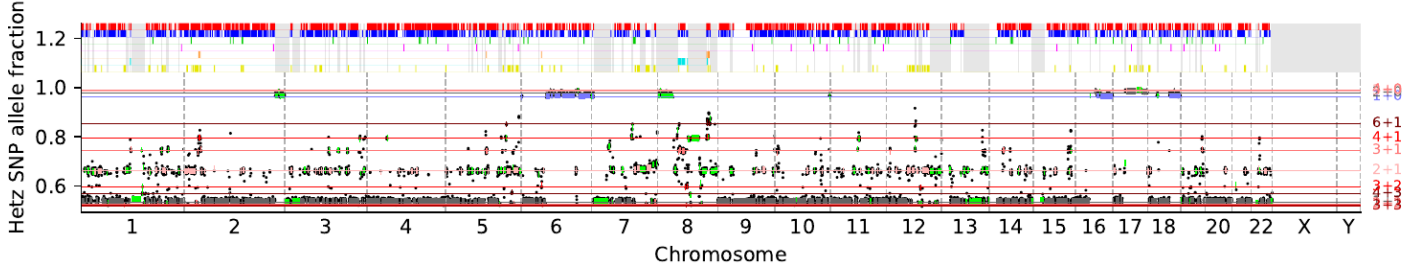
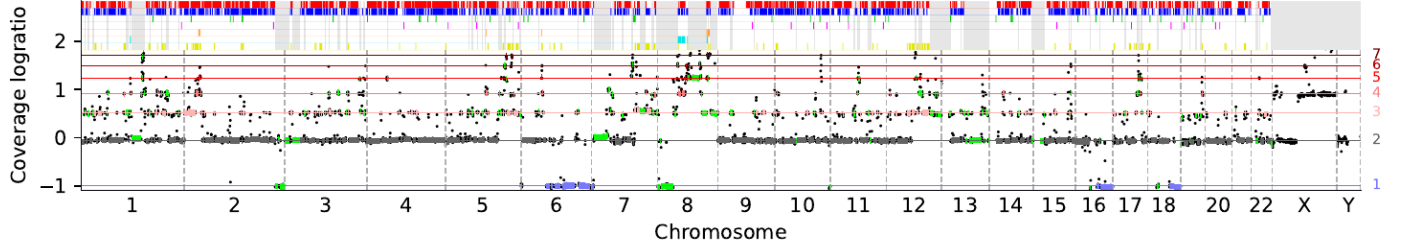


Phylogenetic tree

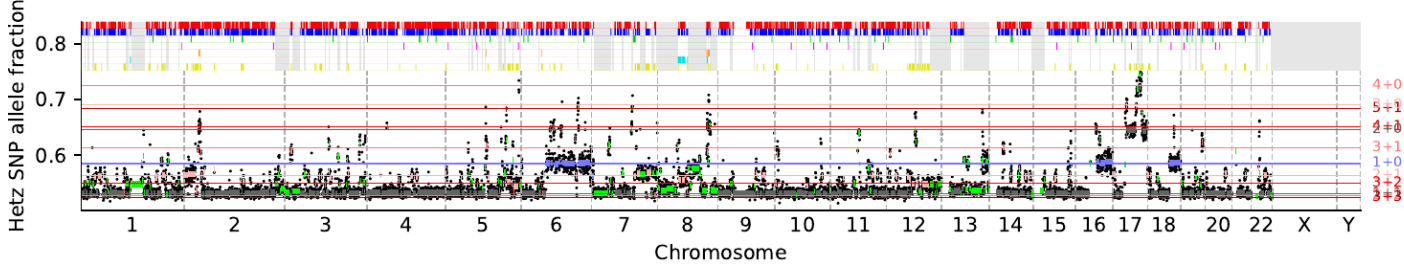
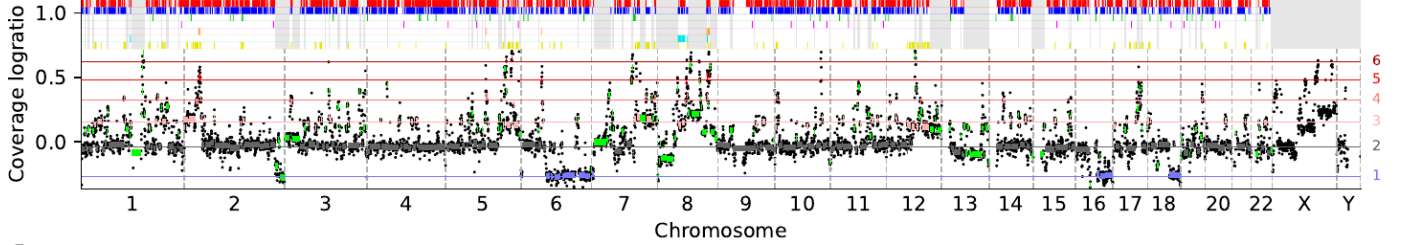


DTB-183 mutation cluster positions

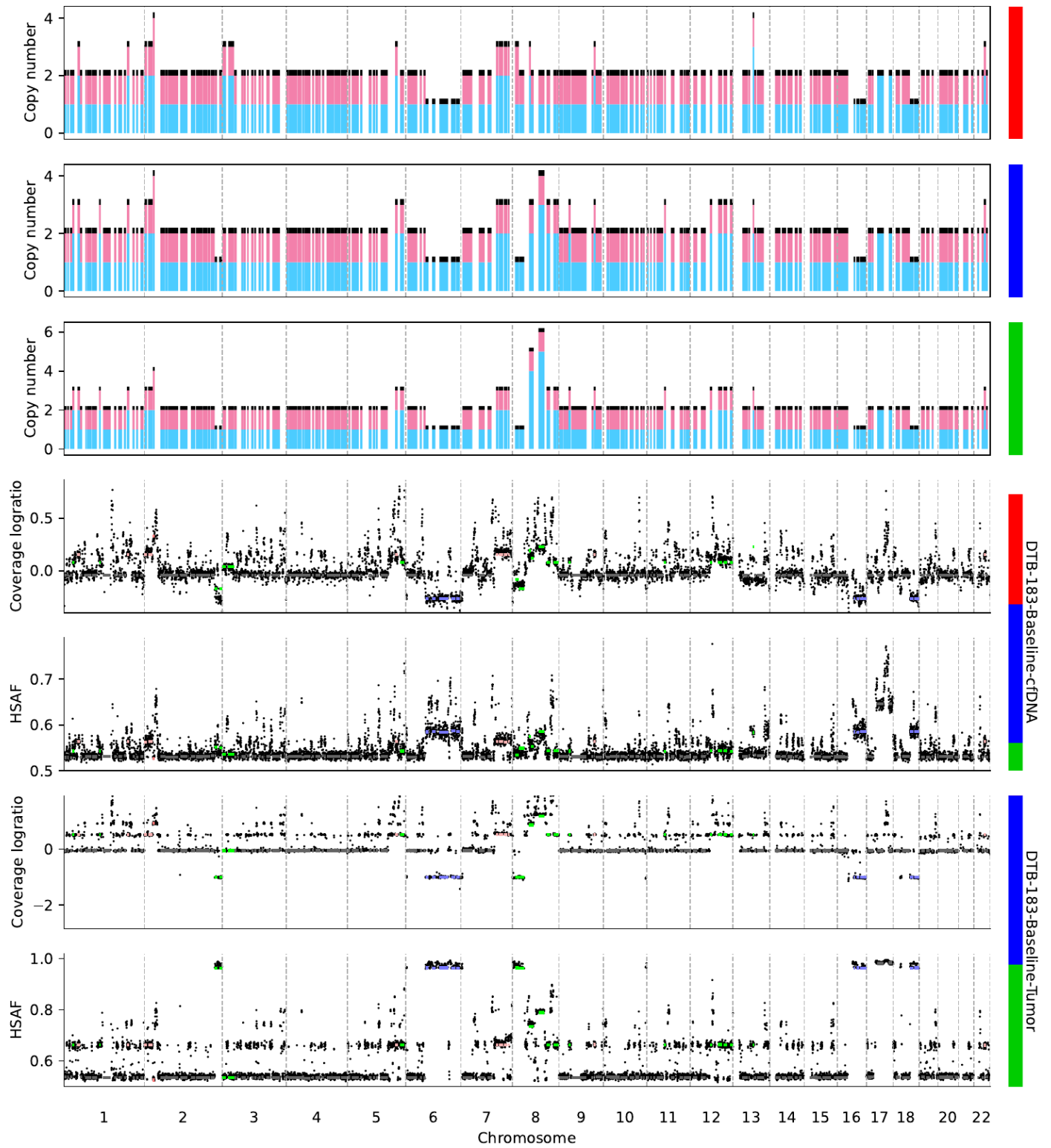
DTB-183-Baseline-Tumor



DTB-183-Baseline-cfDNA



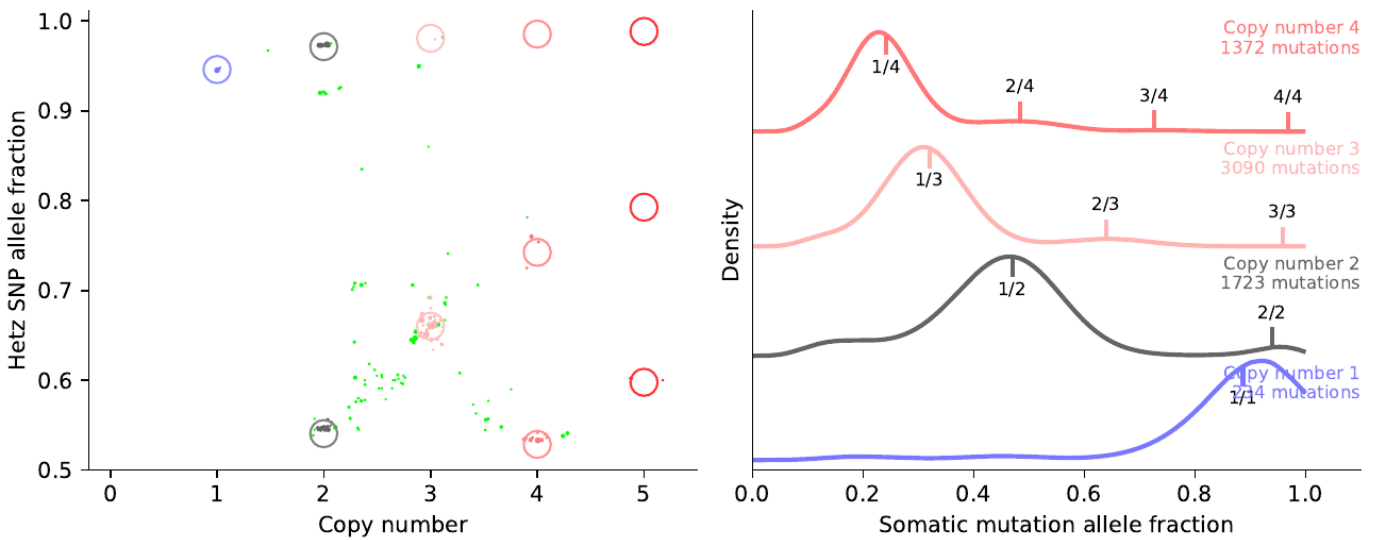
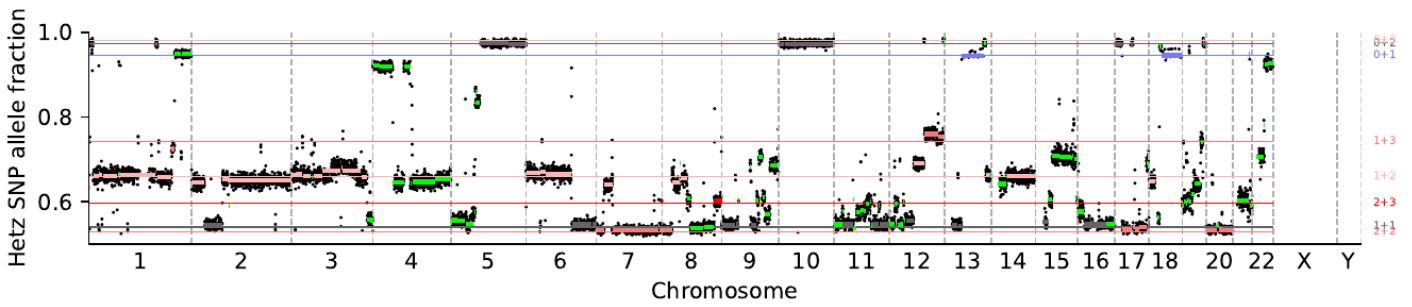
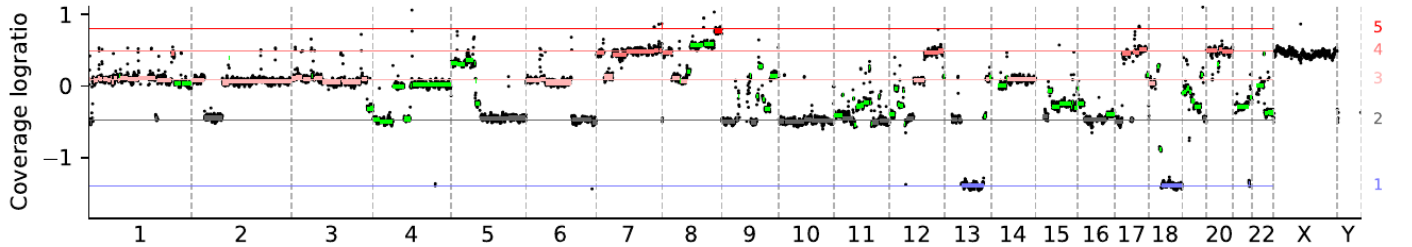
DTB-183 copy number deconvolution



Patient DTB-205

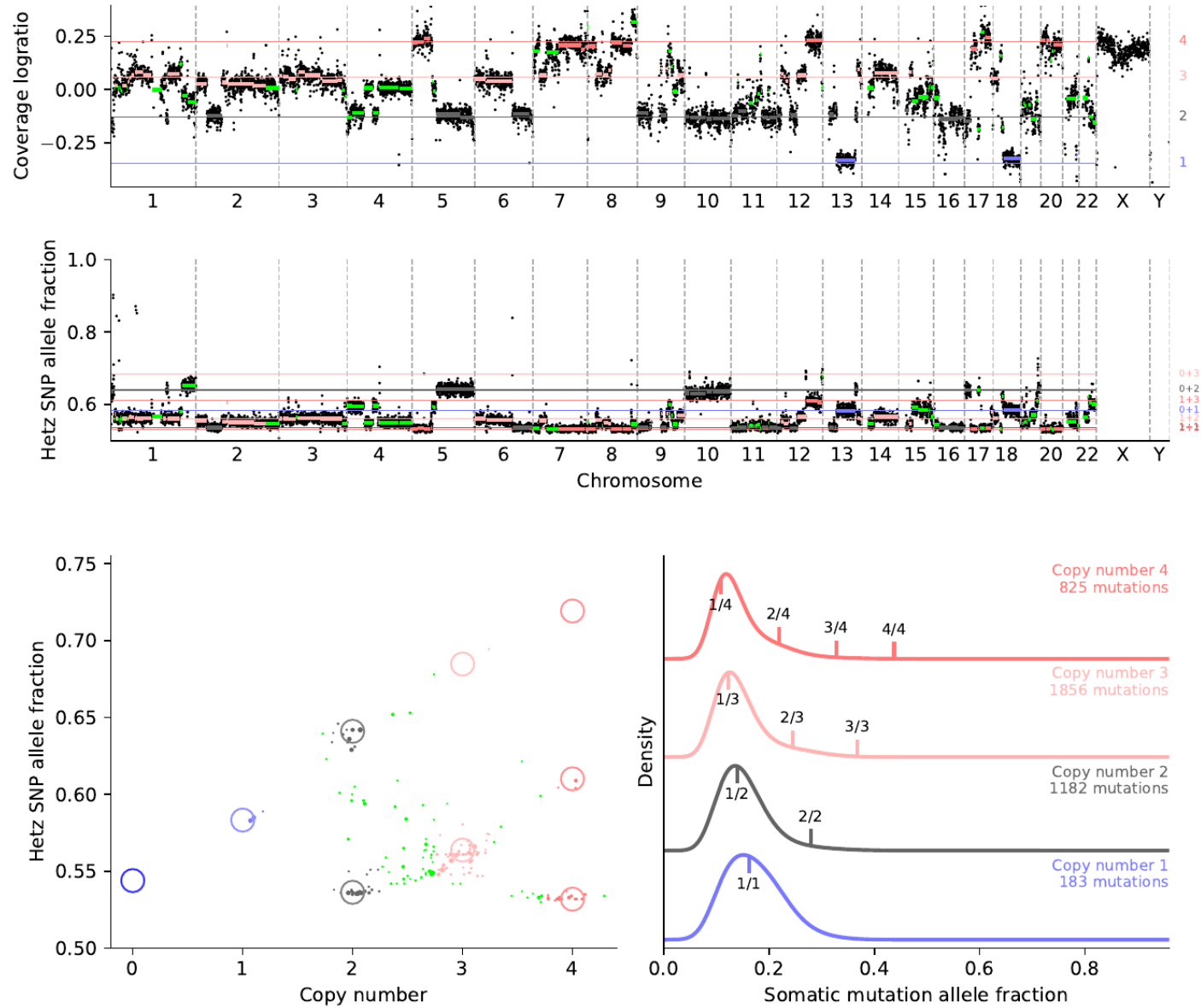
DTB-205-Baseline-Tumor

Cancer fraction: 94%, Diploid level: -0.47

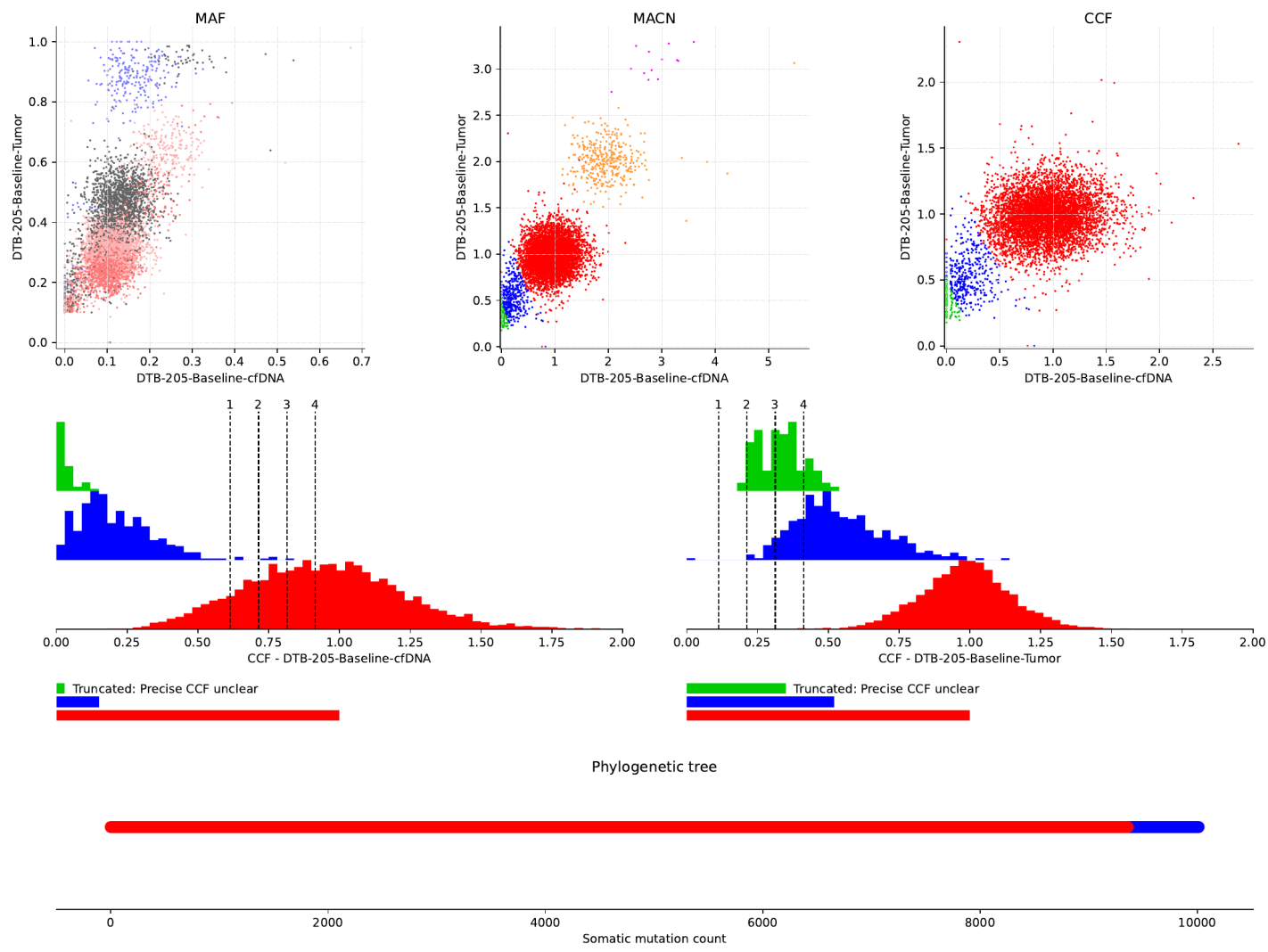


DTB-205-Baseline-cfDNA

Cancer fraction: 28%, Diploid level: -0.13

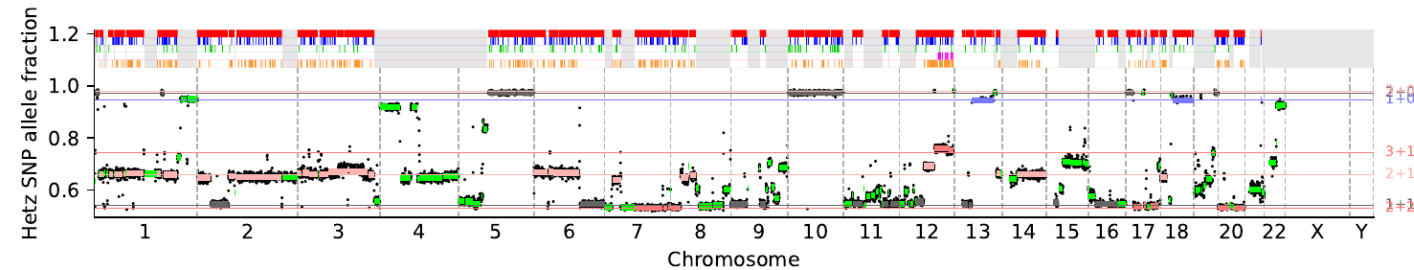
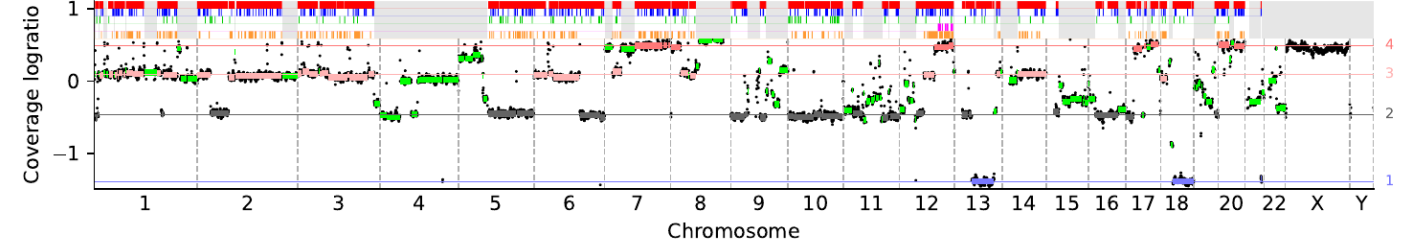


DTB-205 mutation clustering

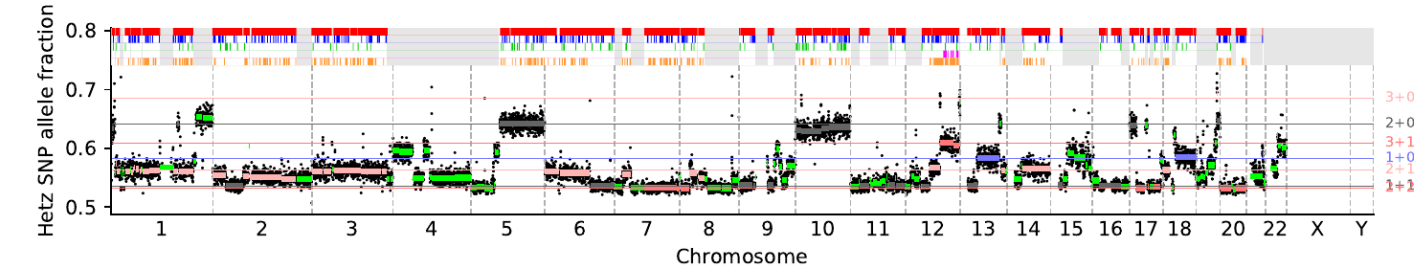
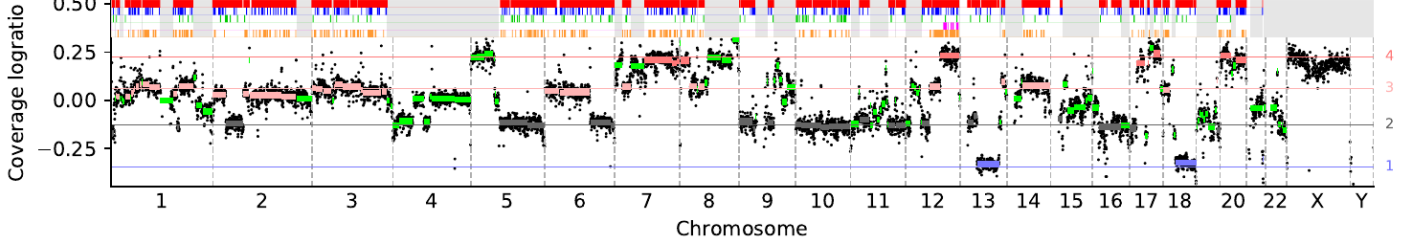


DTB-205 mutation cluster positions

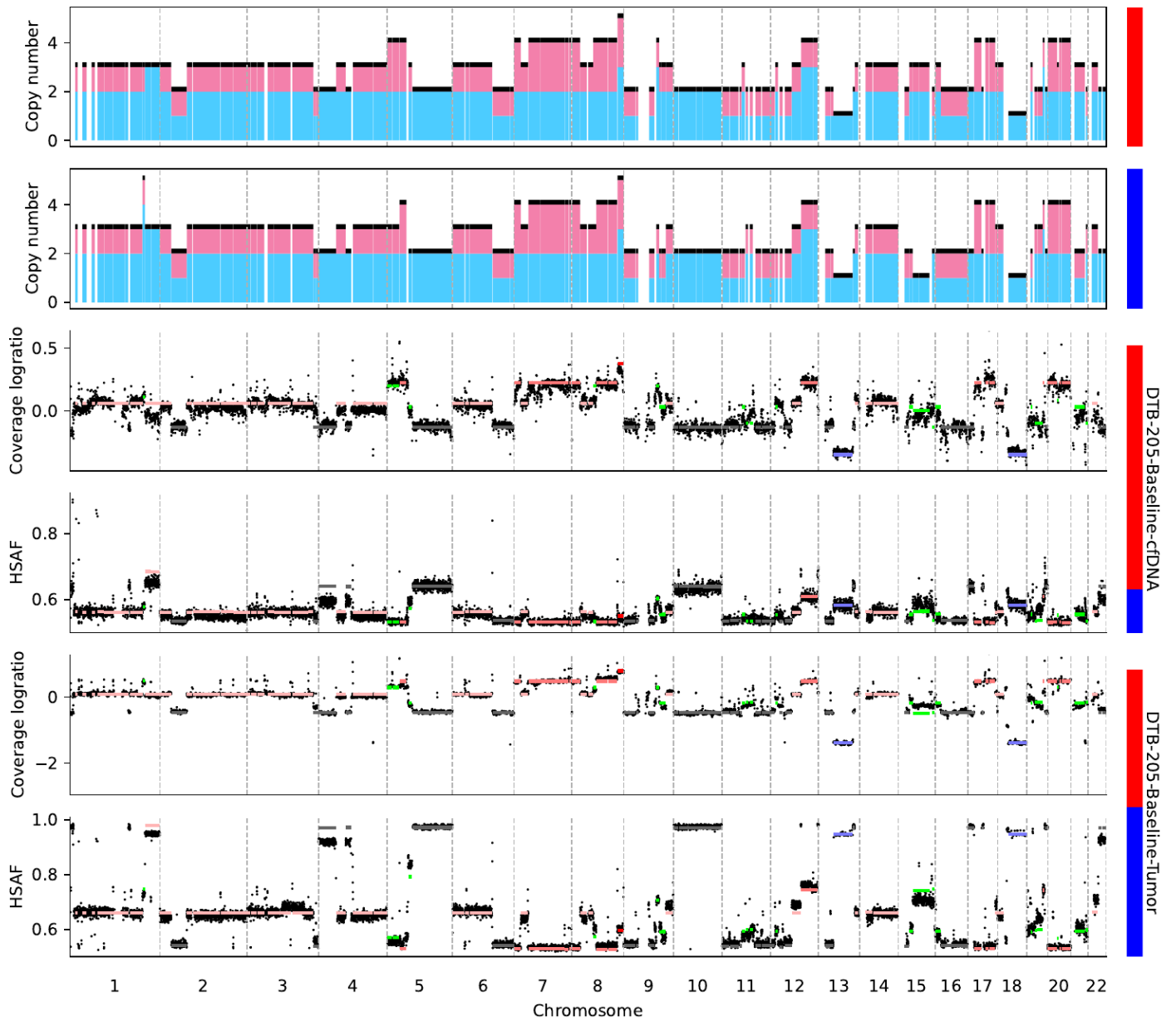
DTB-205-Baseline-Tumor



DTB-205-Baseline-cfDNA



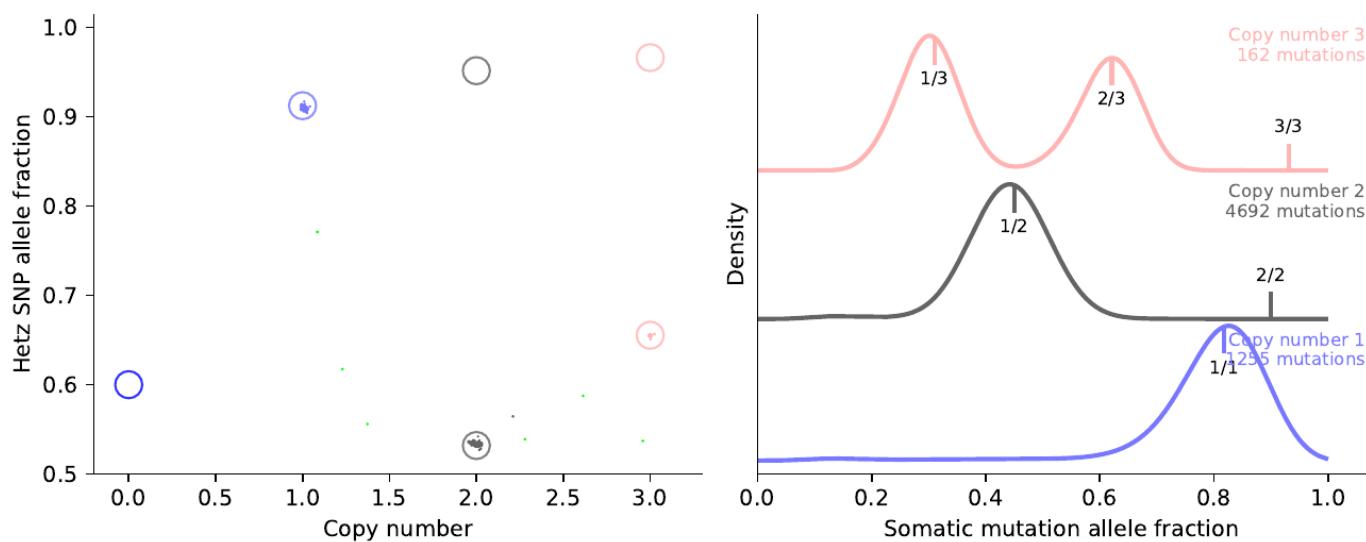
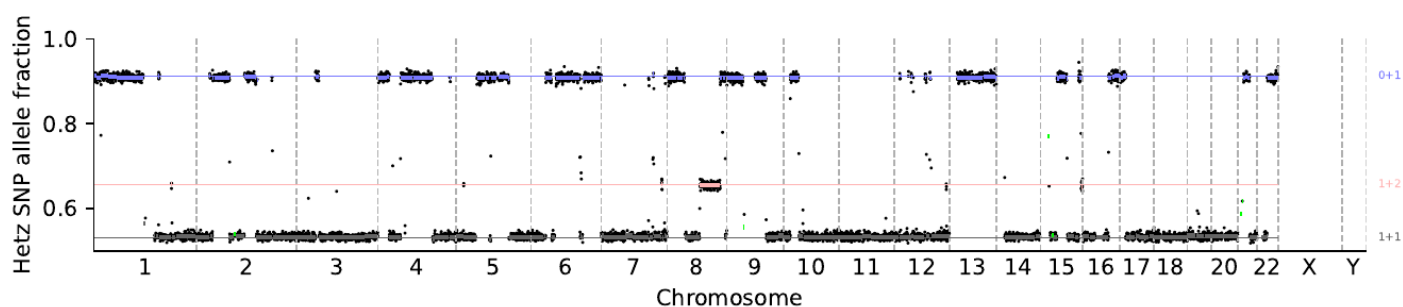
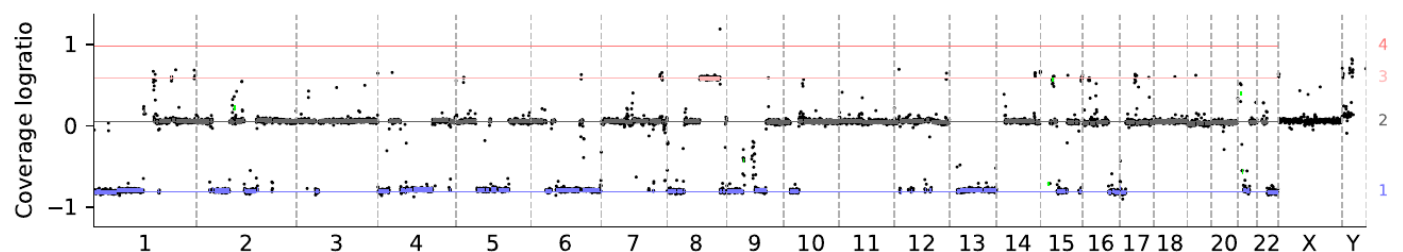
DTB-205 copy number deconvolution



Patient DTB-210

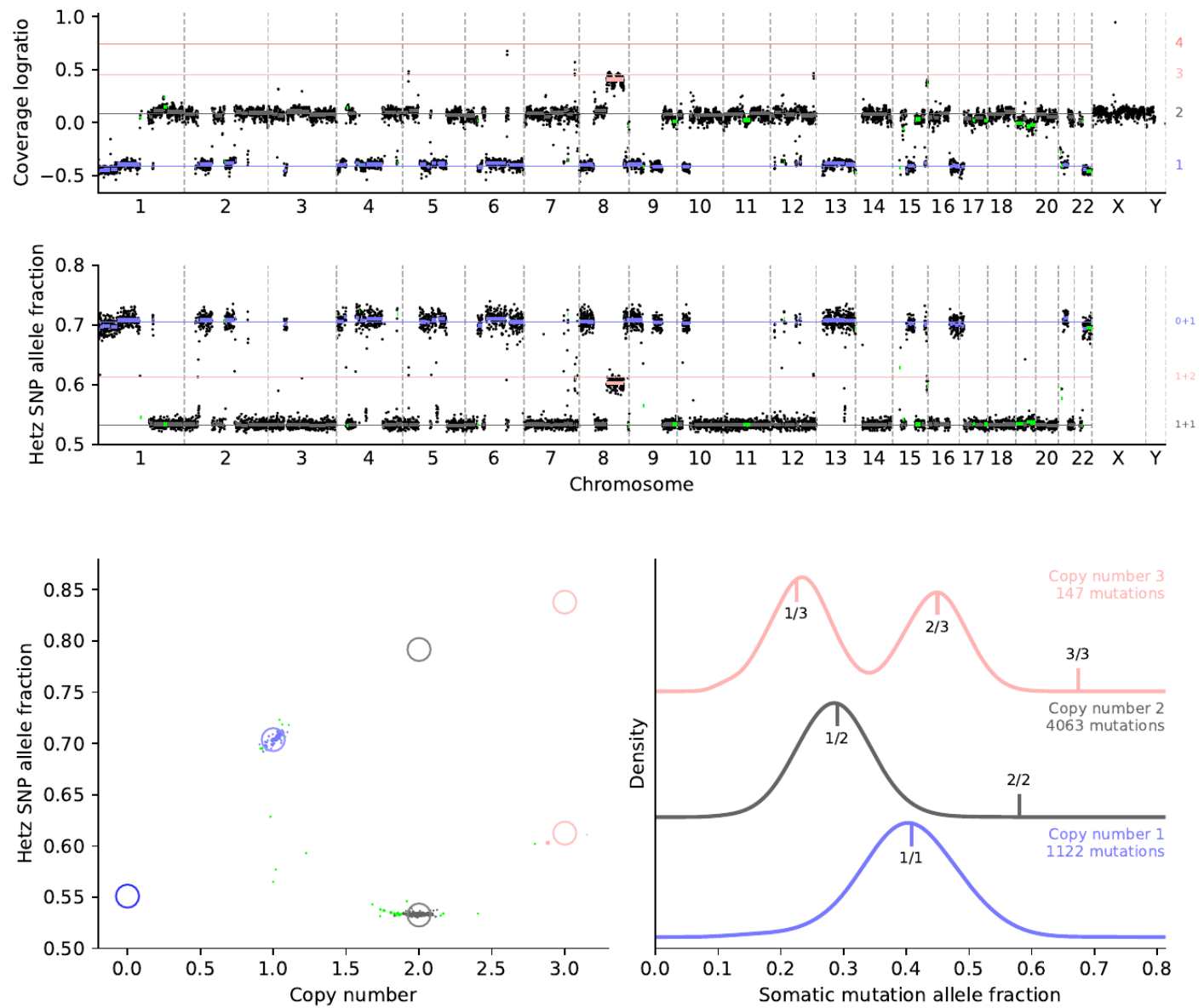
DTB-210-Baseline-Tumor

Cancer fraction: 90%, Diploid level: 0.05



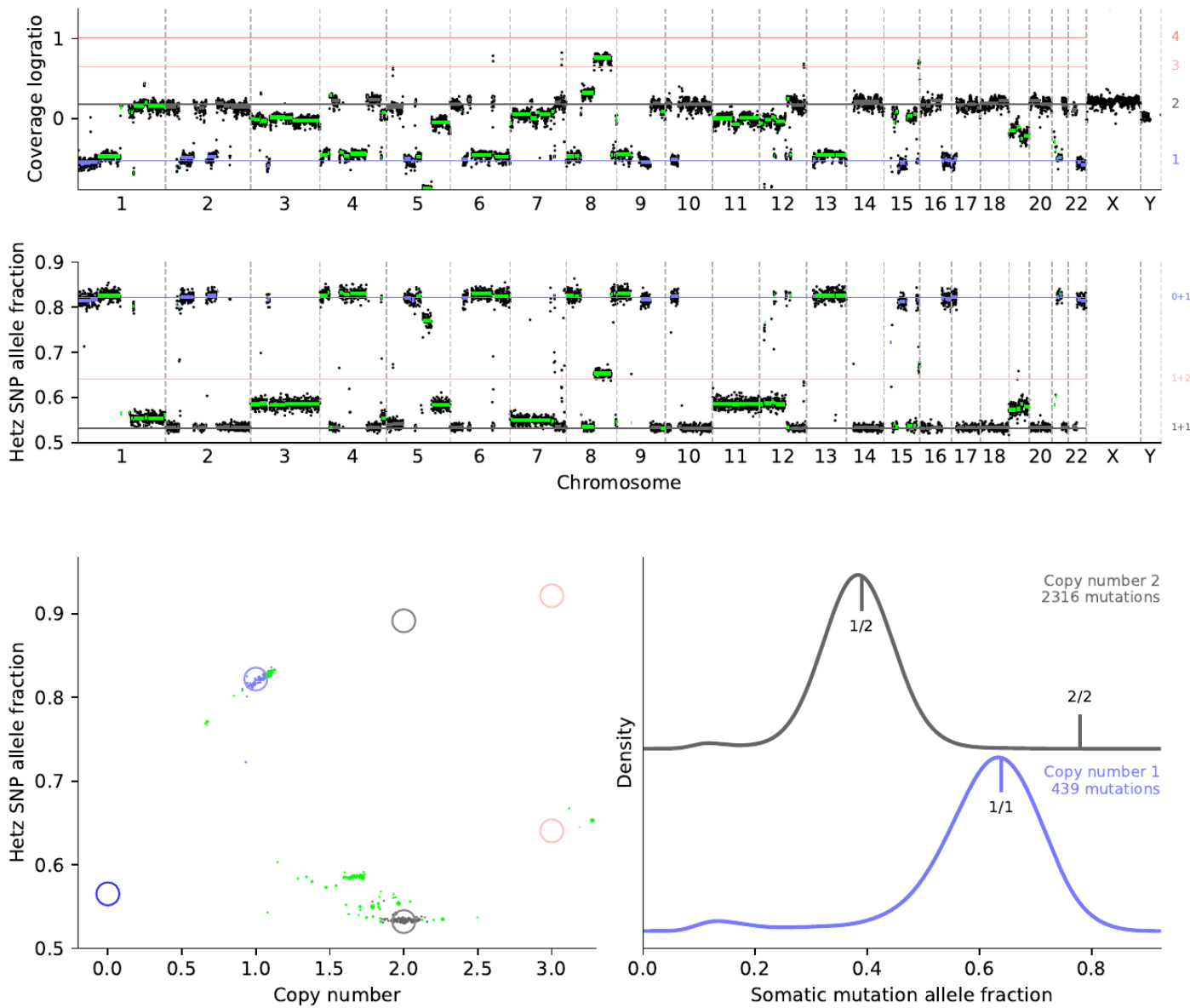
DTB-210-Baseline-cfDNA

Cancer fraction: 58%, Diploid level: 0.08

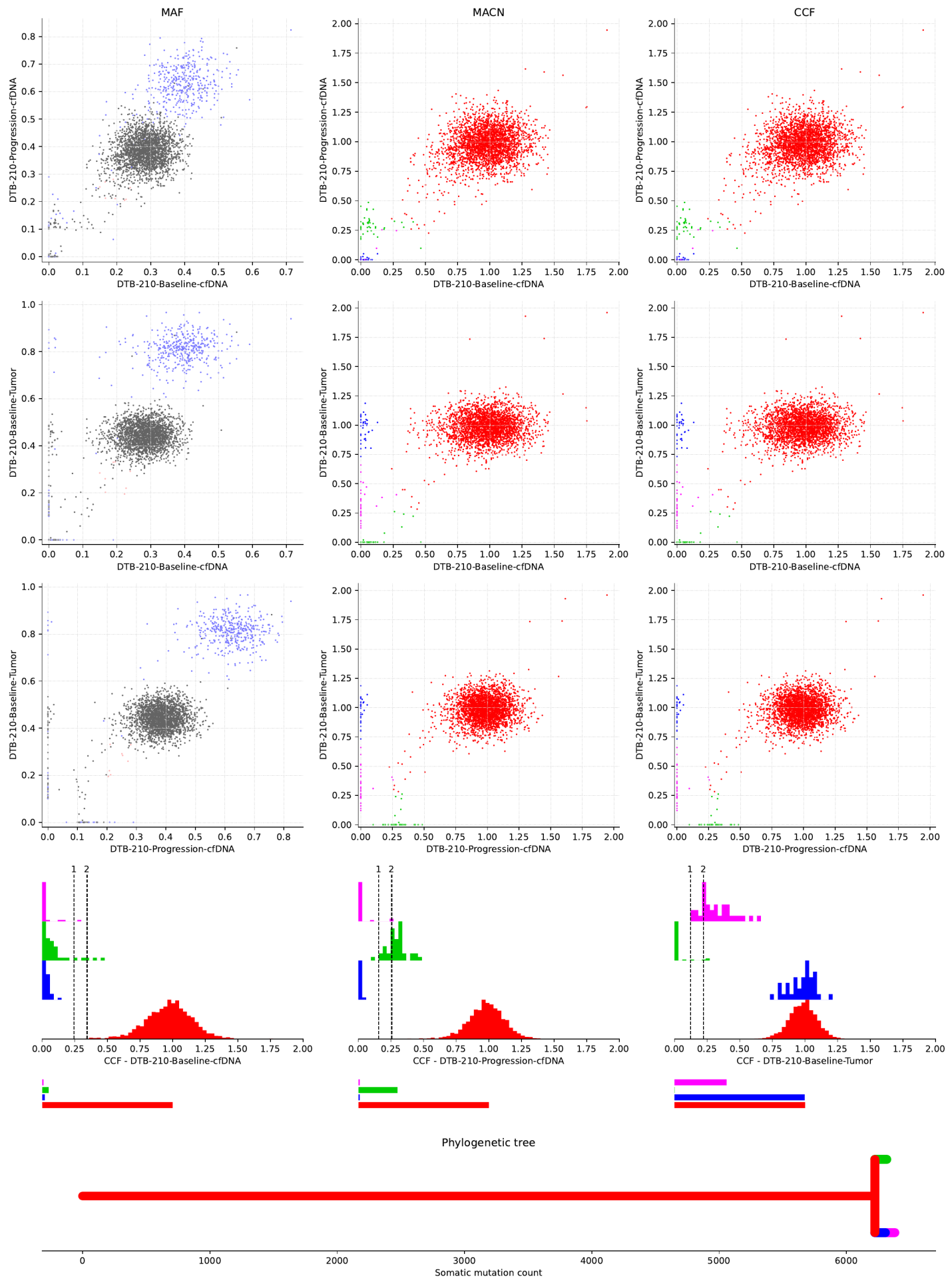


DTB-210-Progression-cfDNA

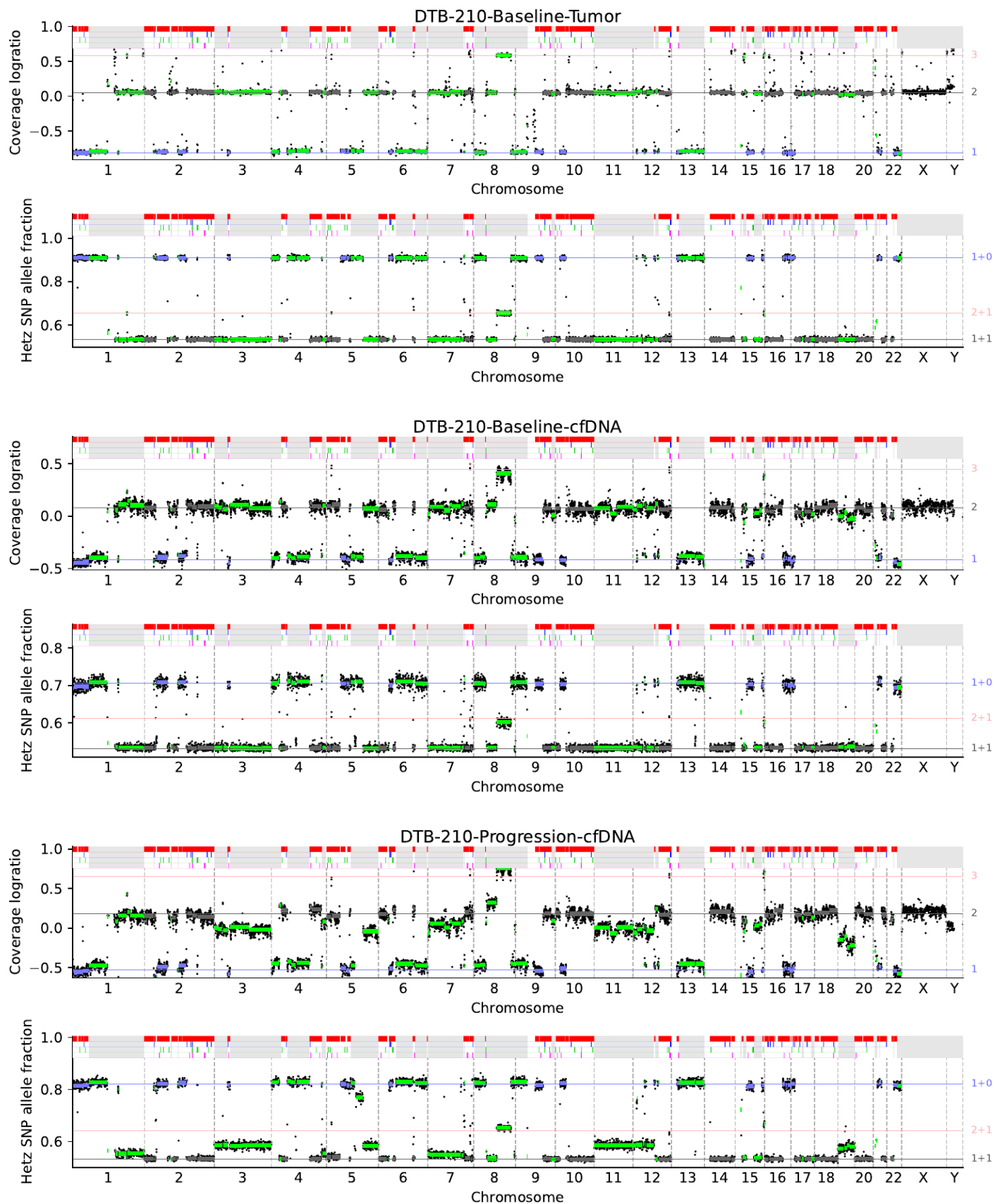
Cancer fraction: 78%, Diploid level: 0.18



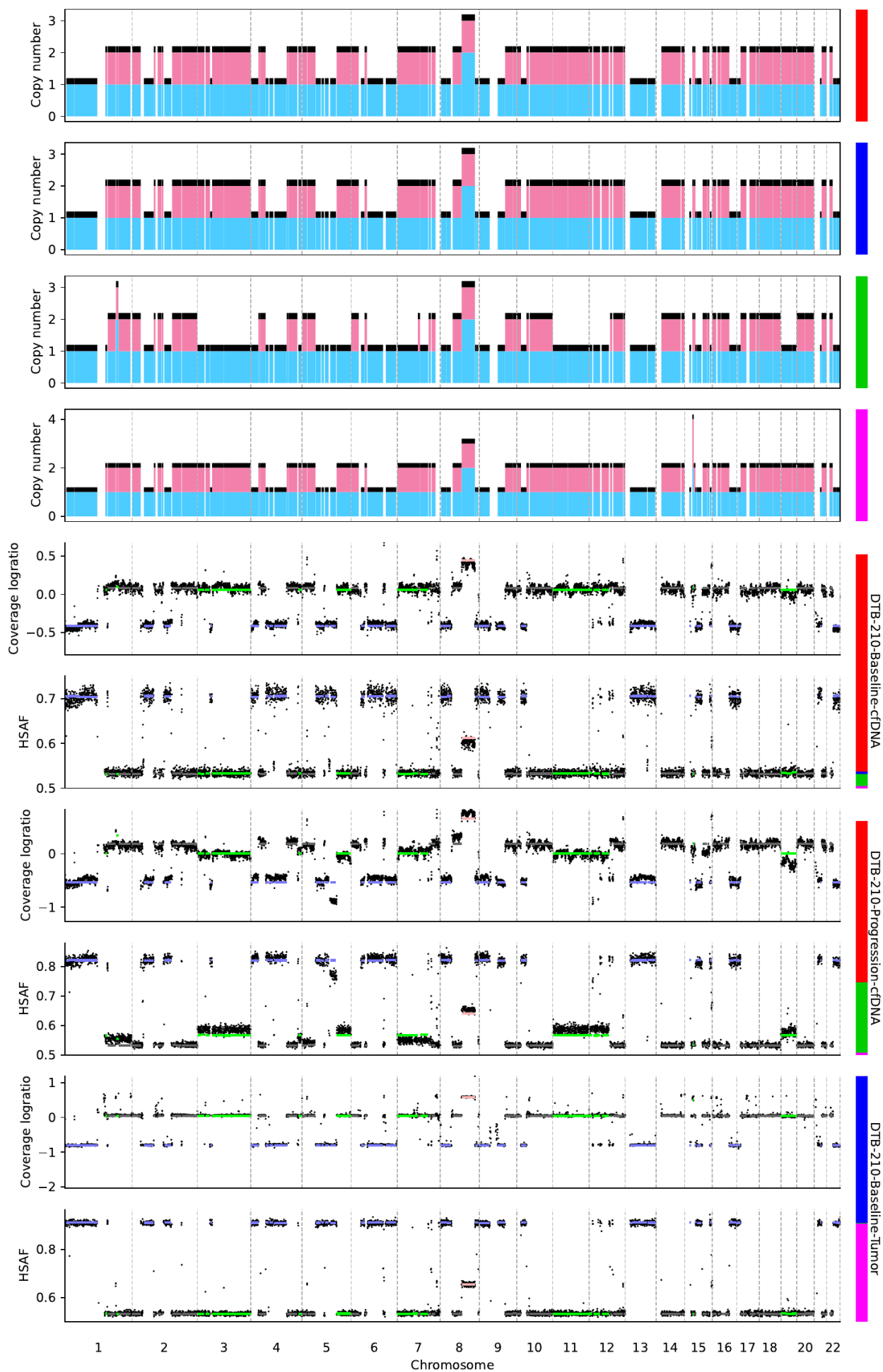
DTB-210 mutation clustering



DTB-210 mutation cluster positions



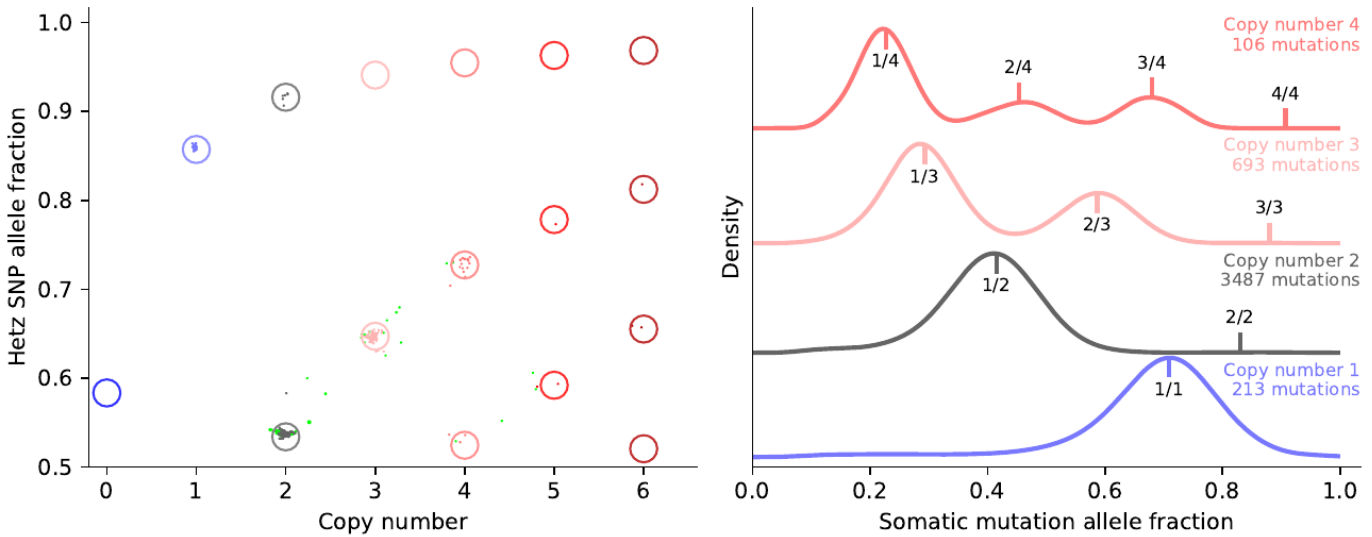
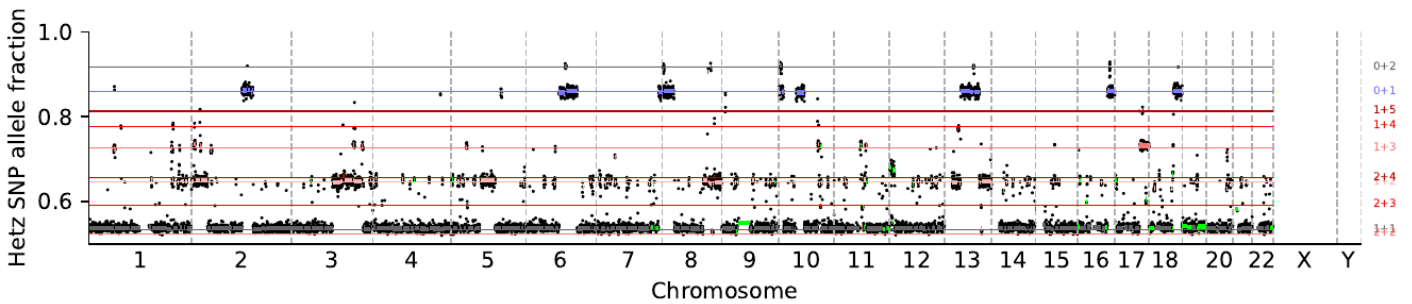
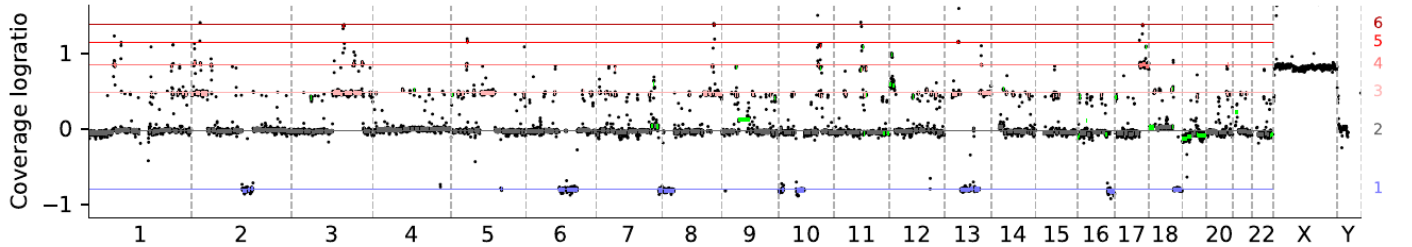
DTB-210 copy number deconvolution



Patient DTB-214

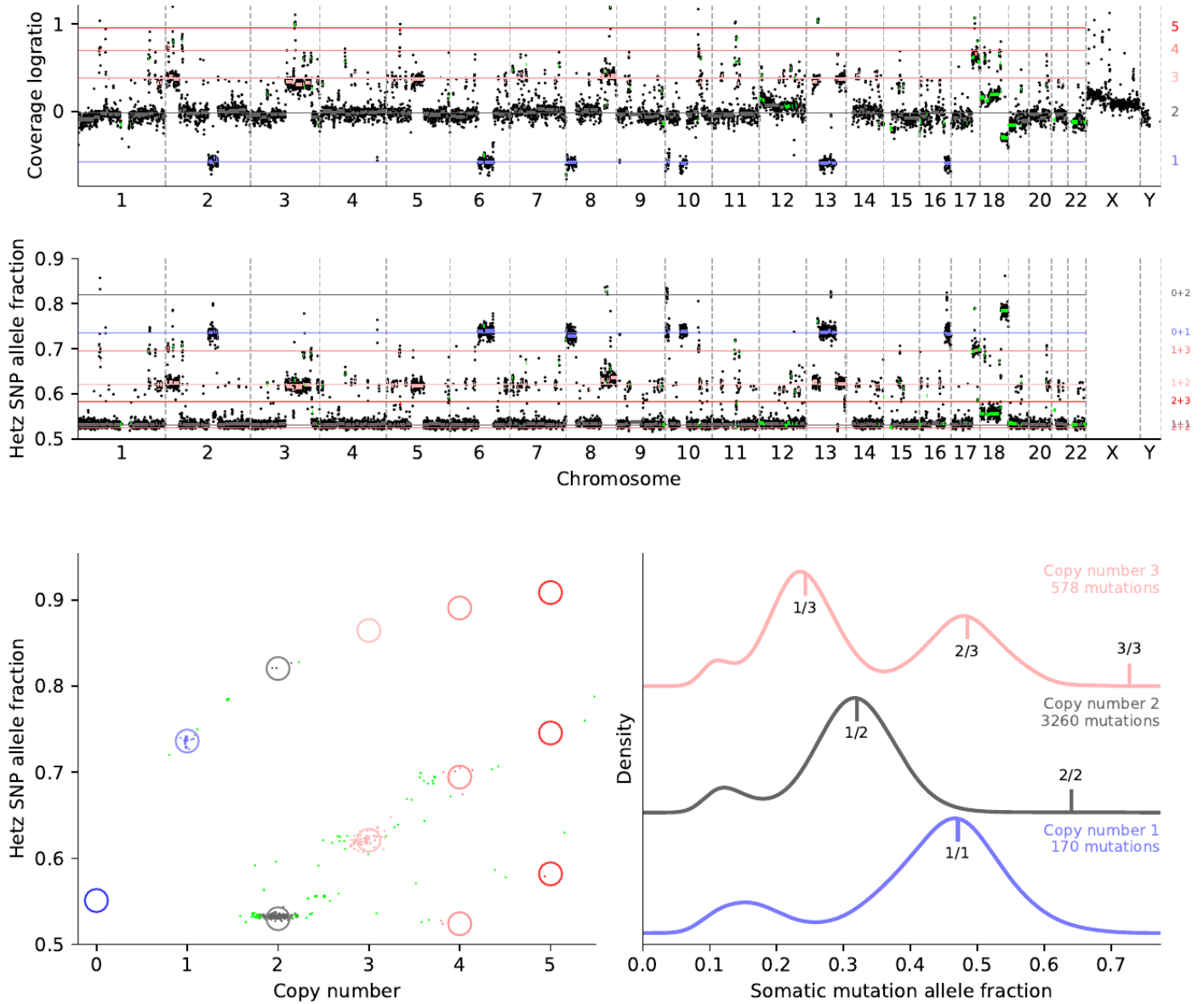
DTB-214-Baseline-Tumor

Cancer fraction: 83%, Diploid level: -0.02

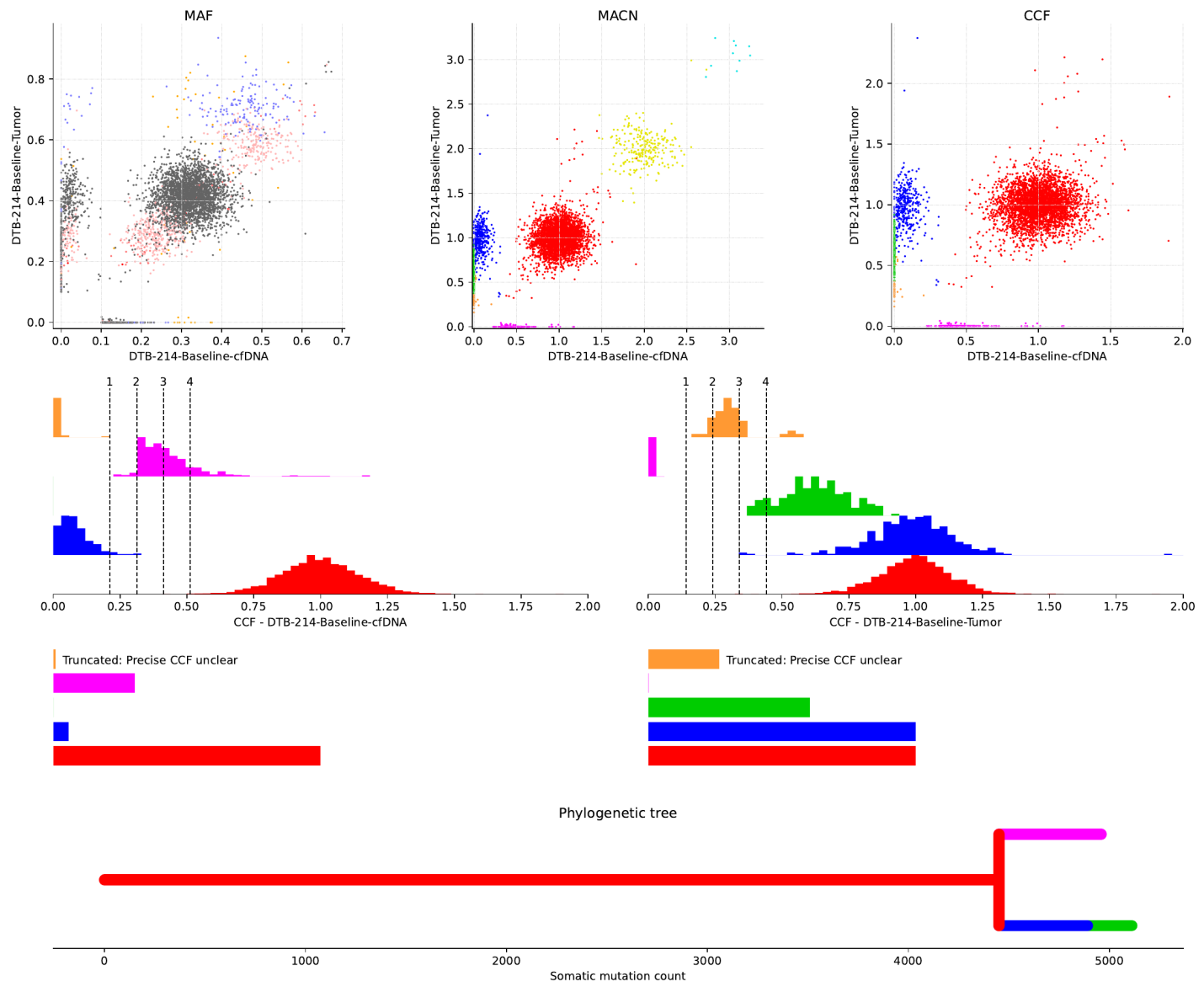


DTB-214-Baseline-cfDNA

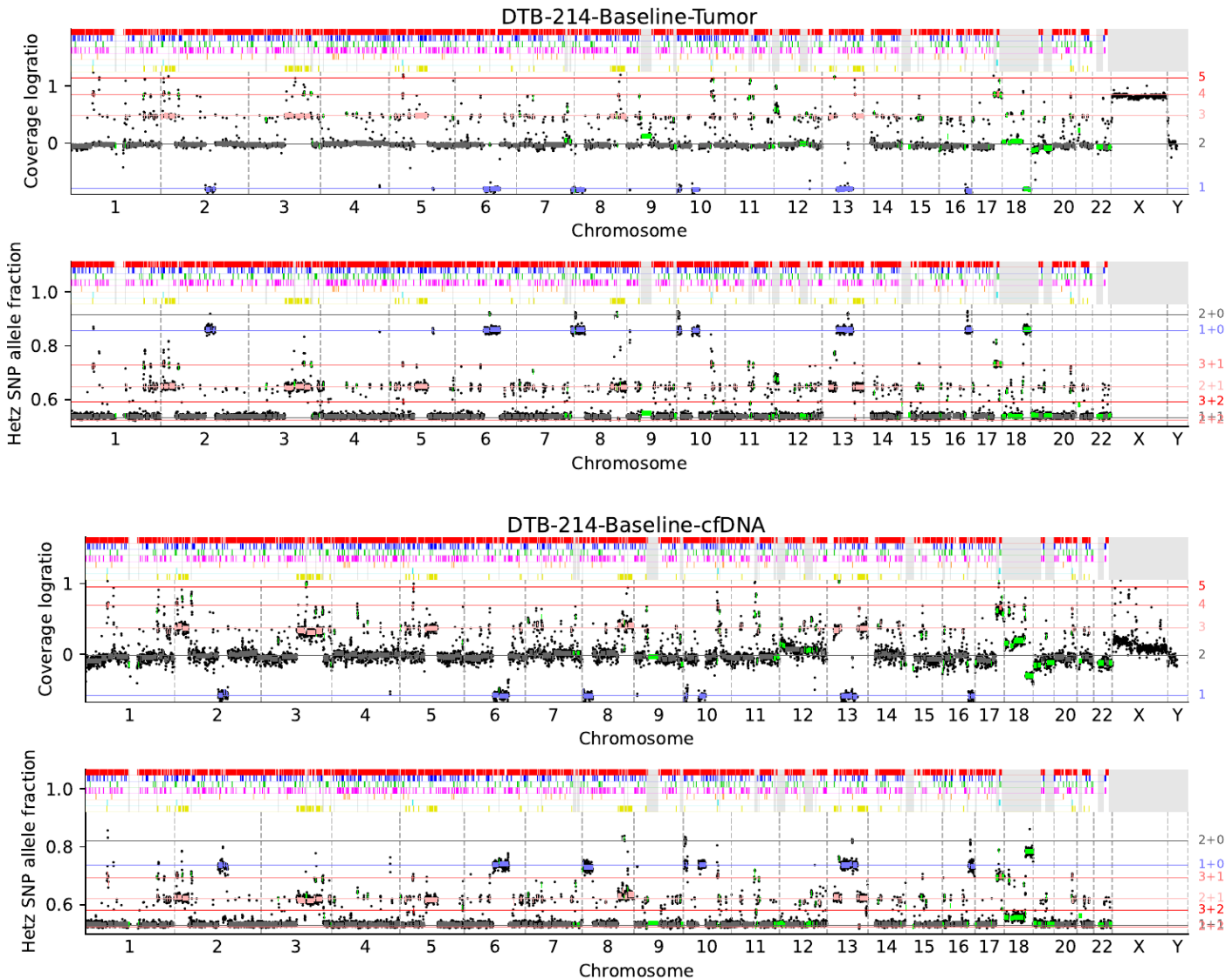
Cancer fraction: 64%, Diploid level: -0.02



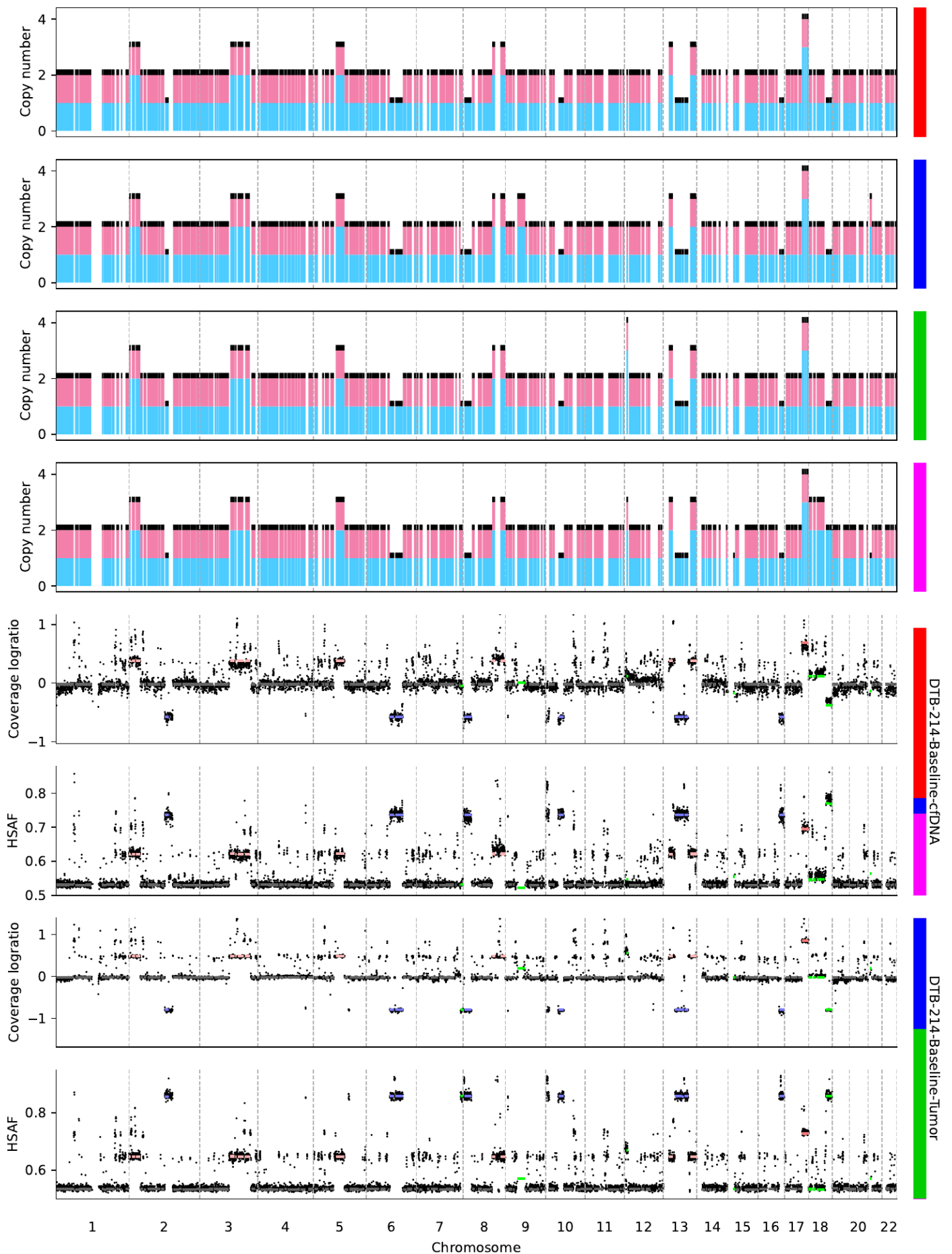
DTB-214 mutation clustering



DTB-214 mutation cluster positions



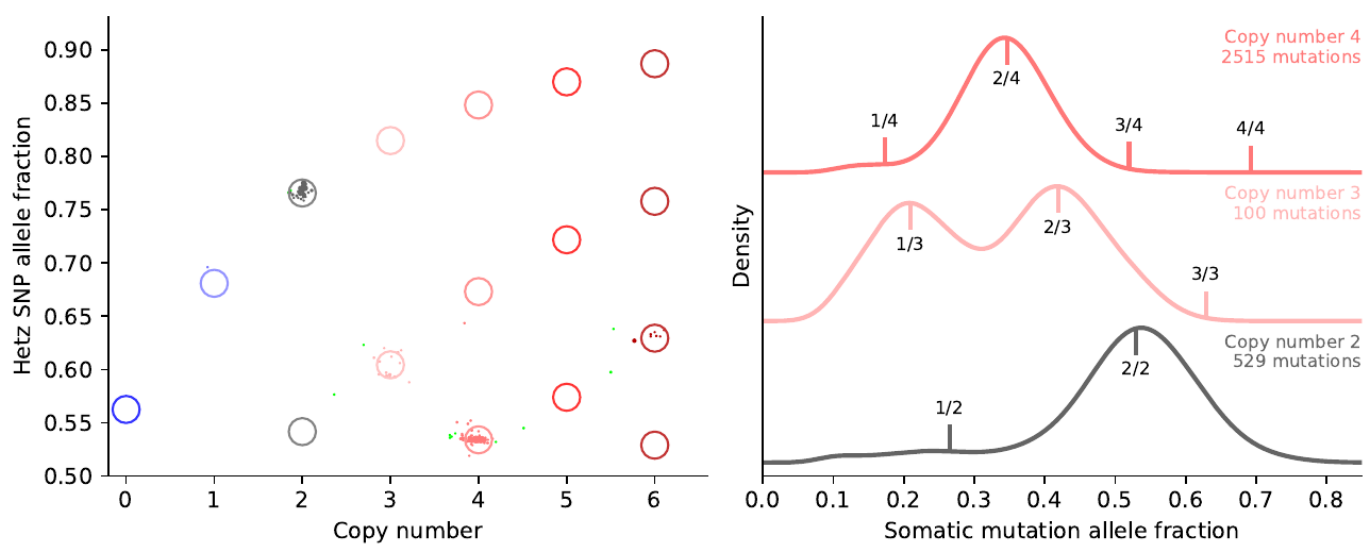
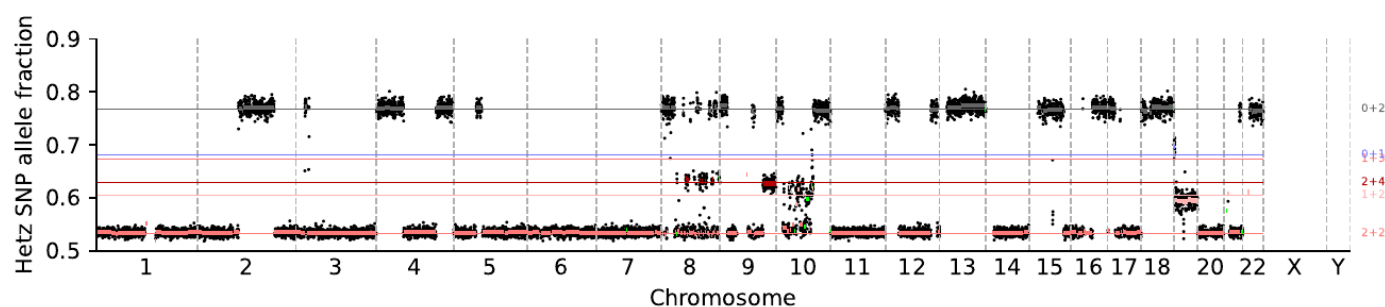
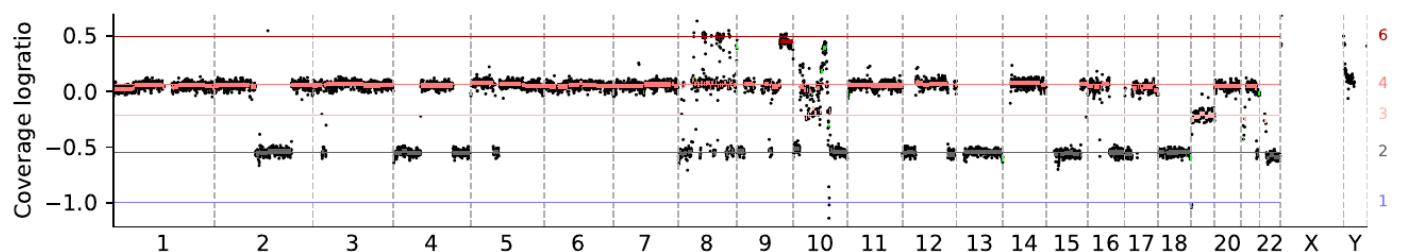
DTB-214 copy number deconvolution



Patient DTB-216

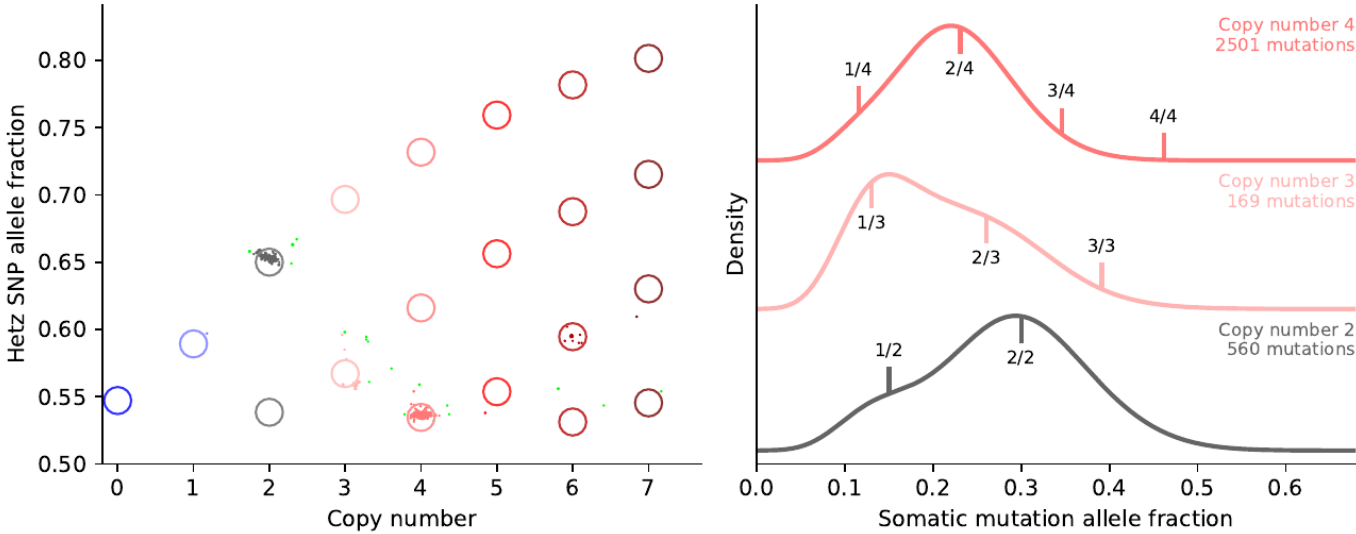
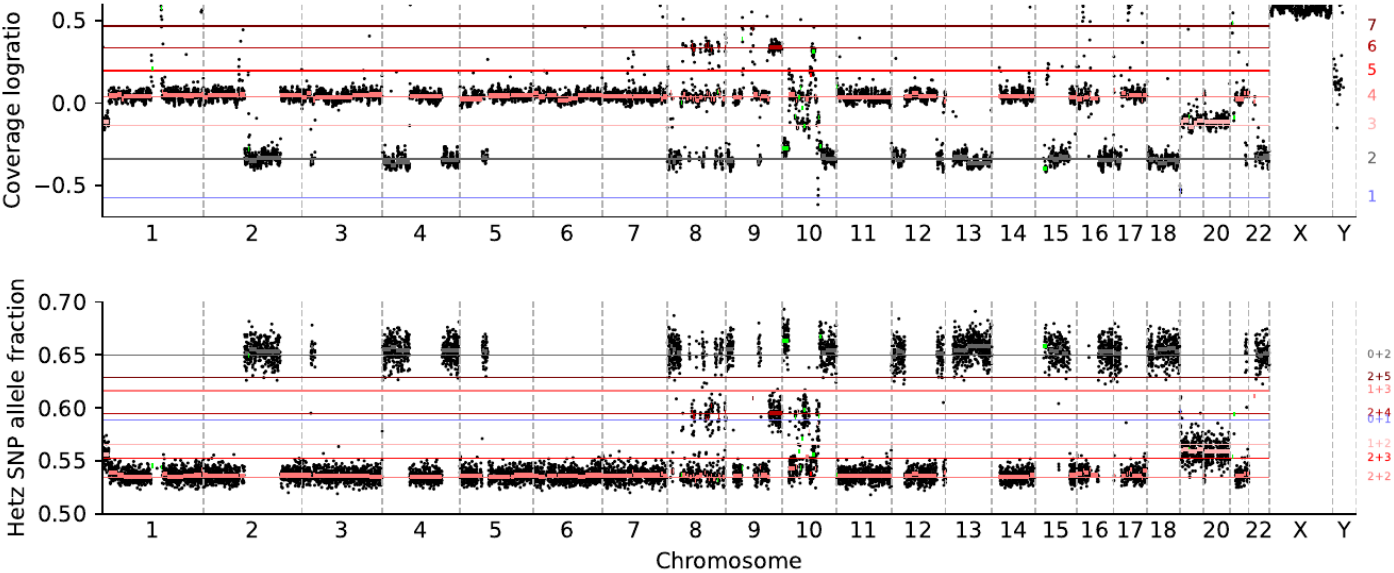
DTB-216-Baseline-cfDNA

Cancer fraction: 53%, Diploid level: -0.55



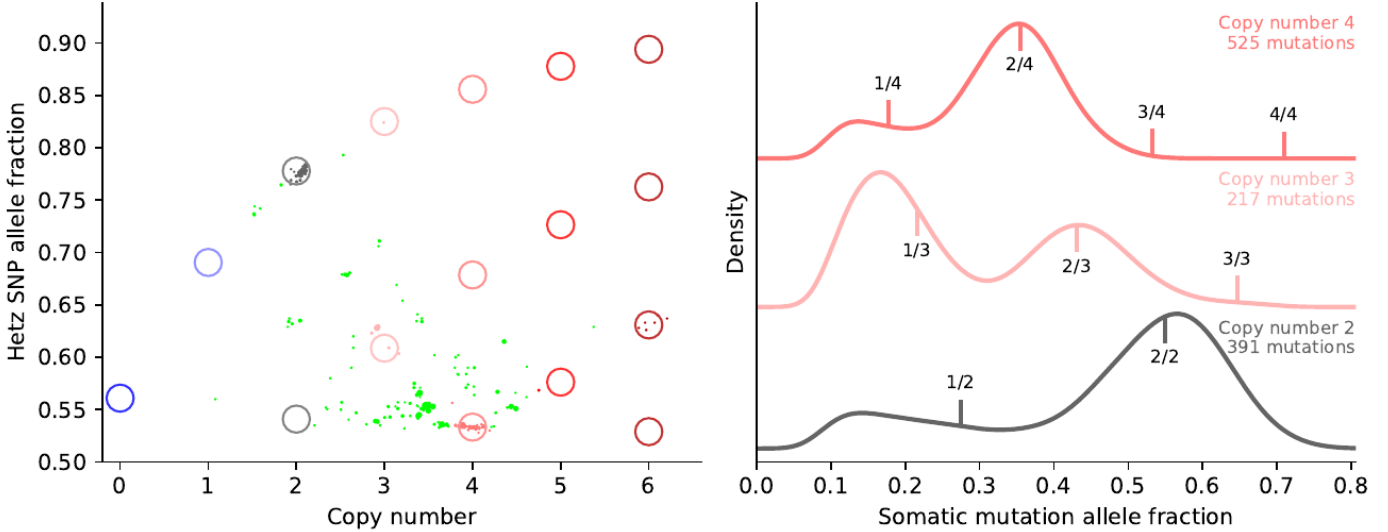
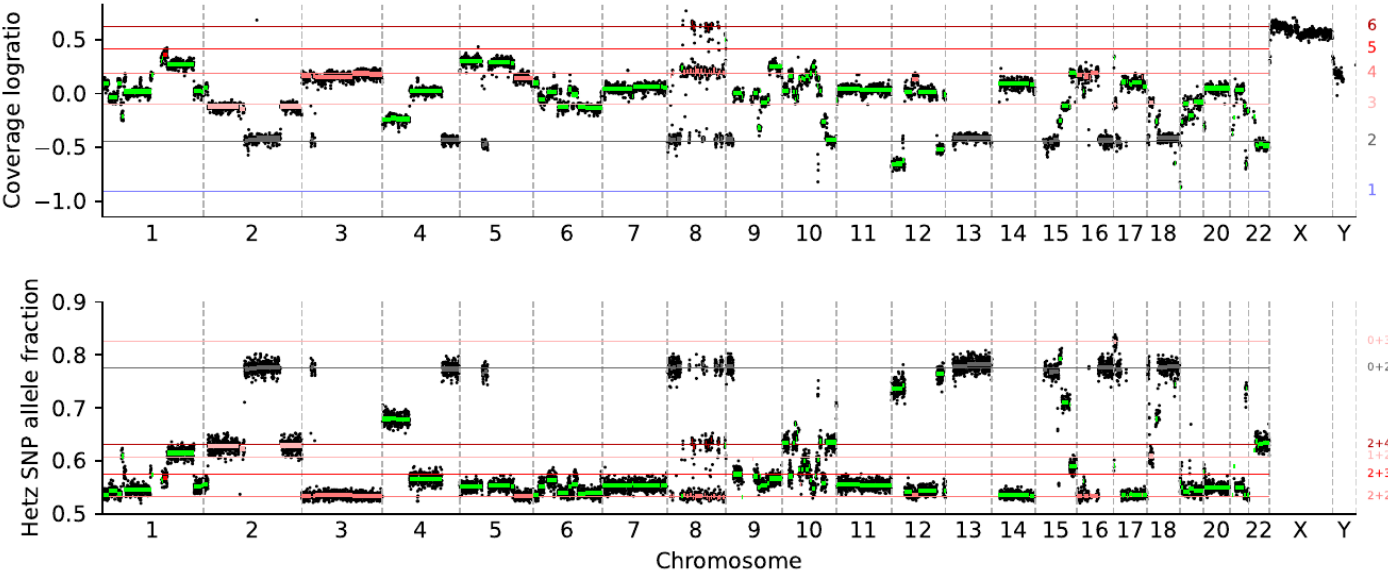
DTB-216-Progression-Tumor

Cancer fraction: 30%, Diploid level: -0.34

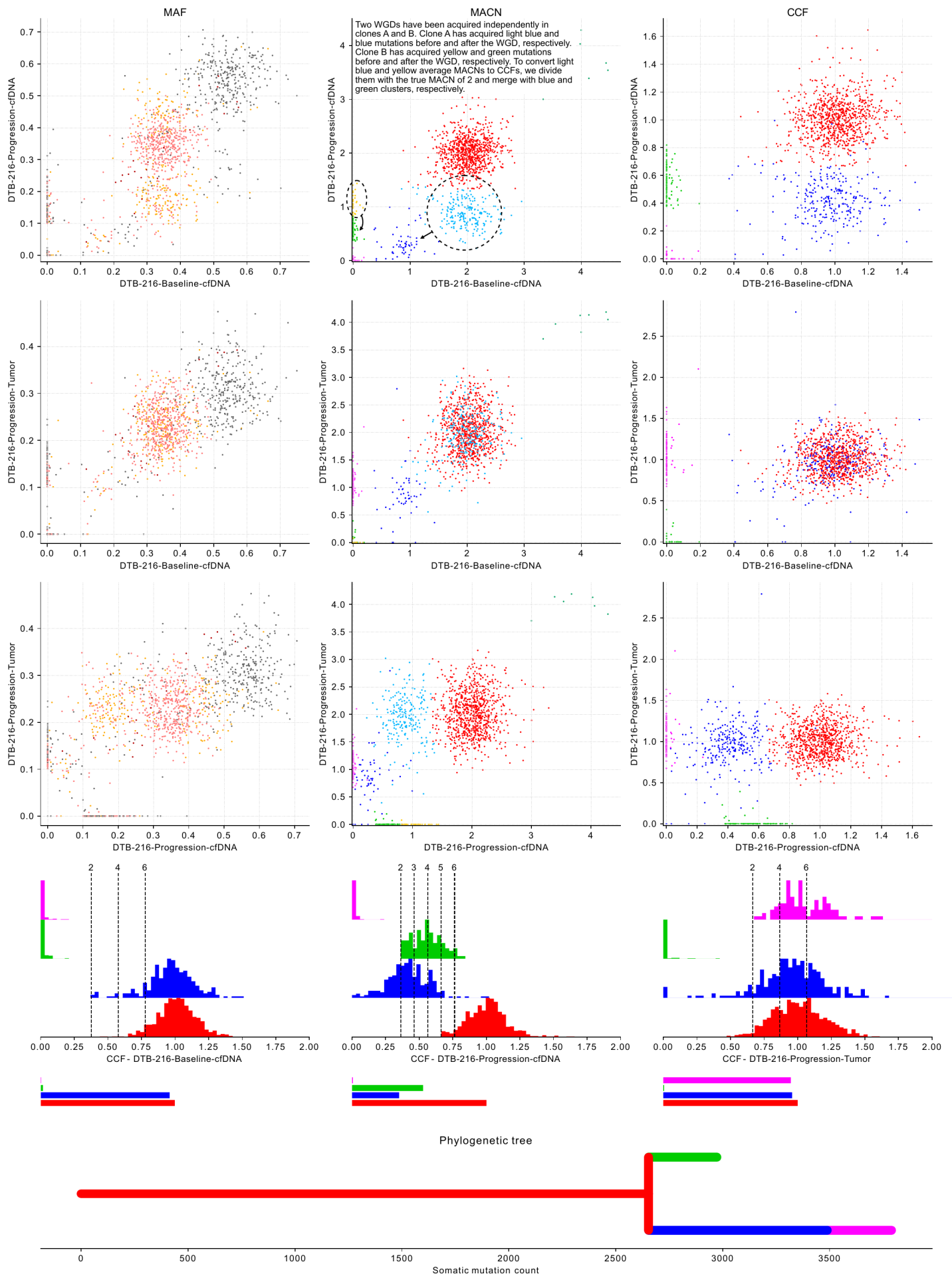


DTB-216-Progression-cfDNA

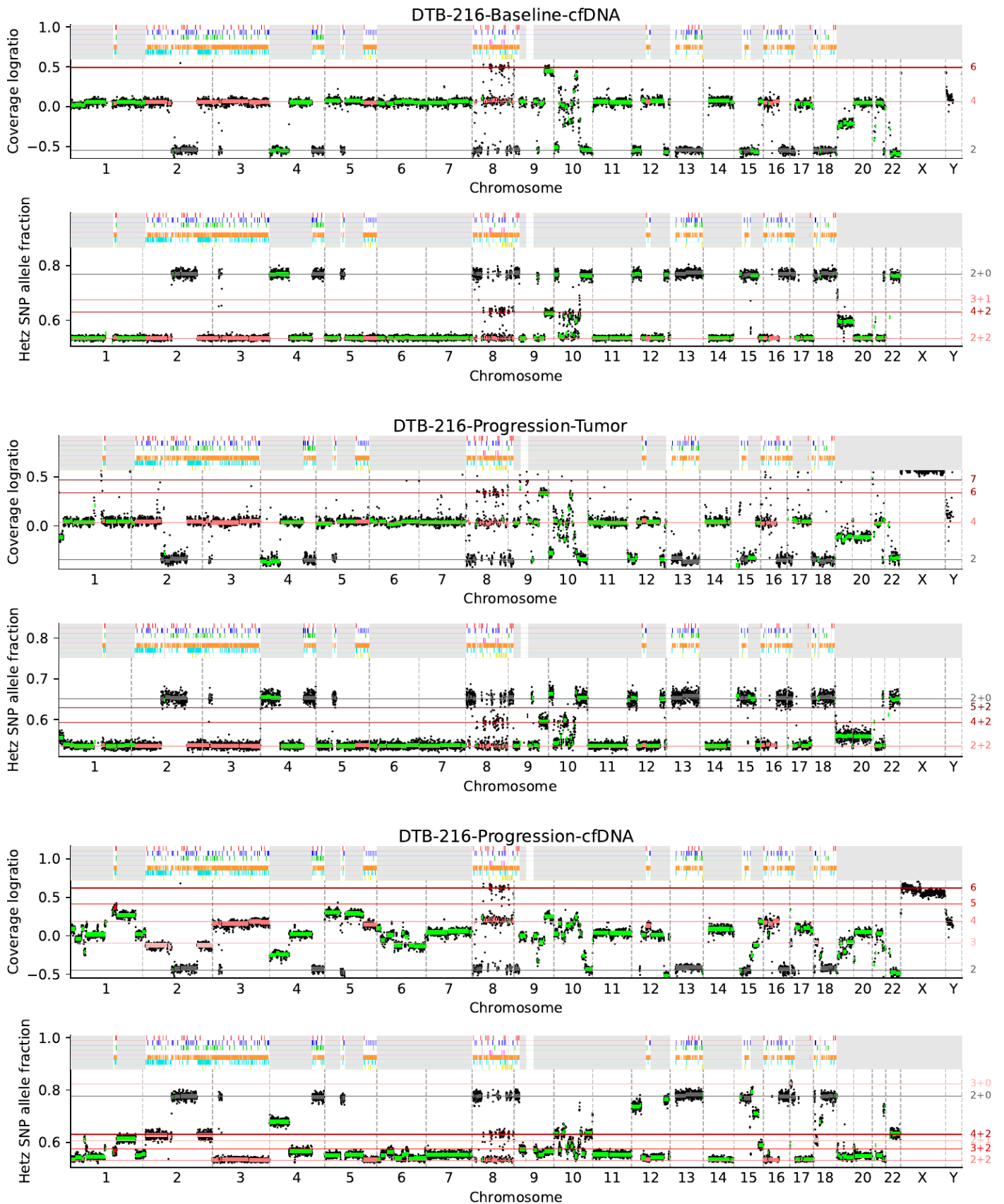
Cancer fraction: 55%, Diploid level: -0.45



DTB-216 mutation clustering



DTB-216 mutation cluster positions



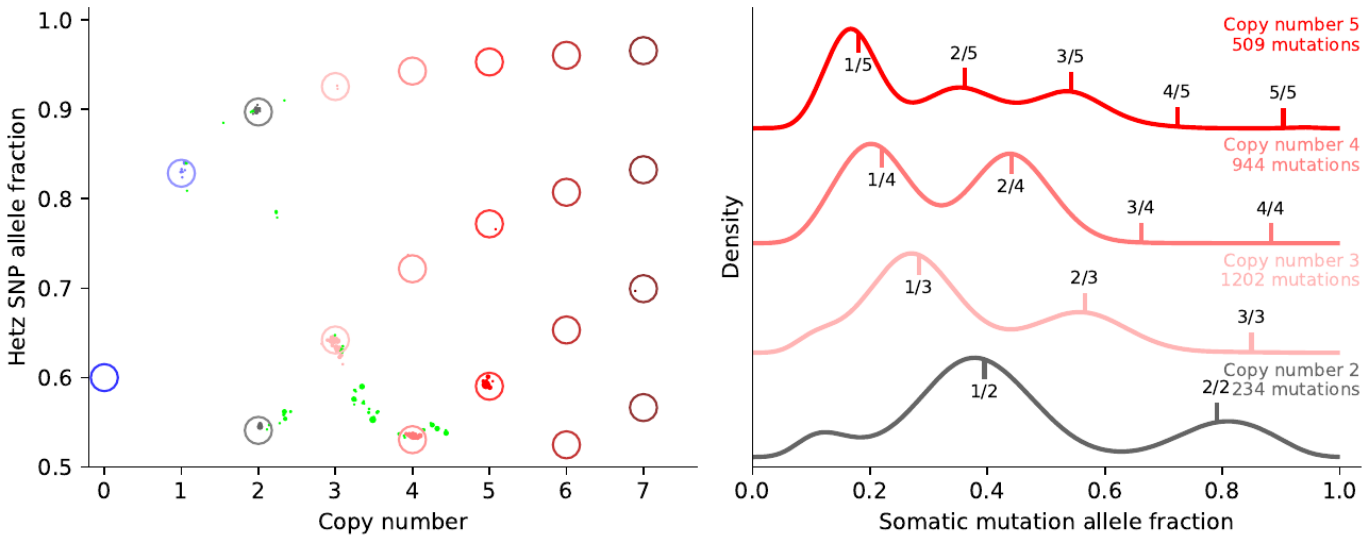
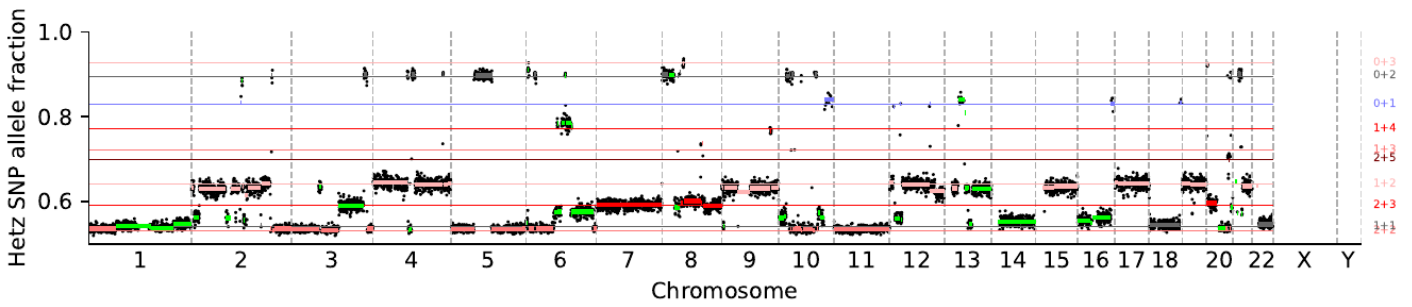
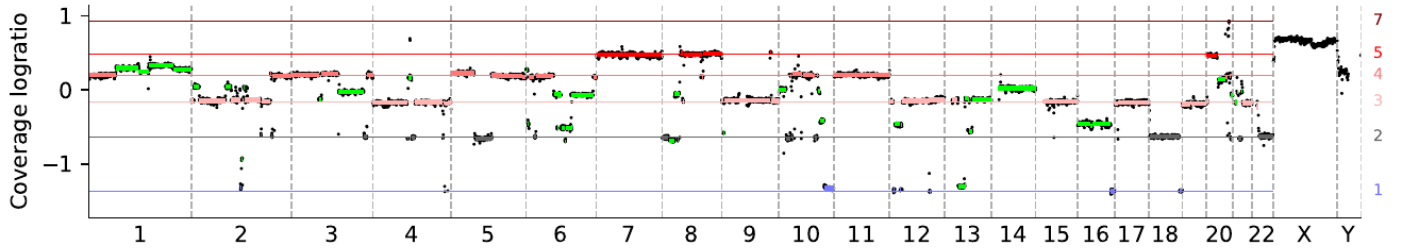
DTB-216 copy number deconvolution

Subclonal copy number deconvolution was not carried out for DTB-216, because our current deconvolution algorithm is not optimally suited for analyzing samples that carry one or more populations with independent WGD events.

Patient DTB-258

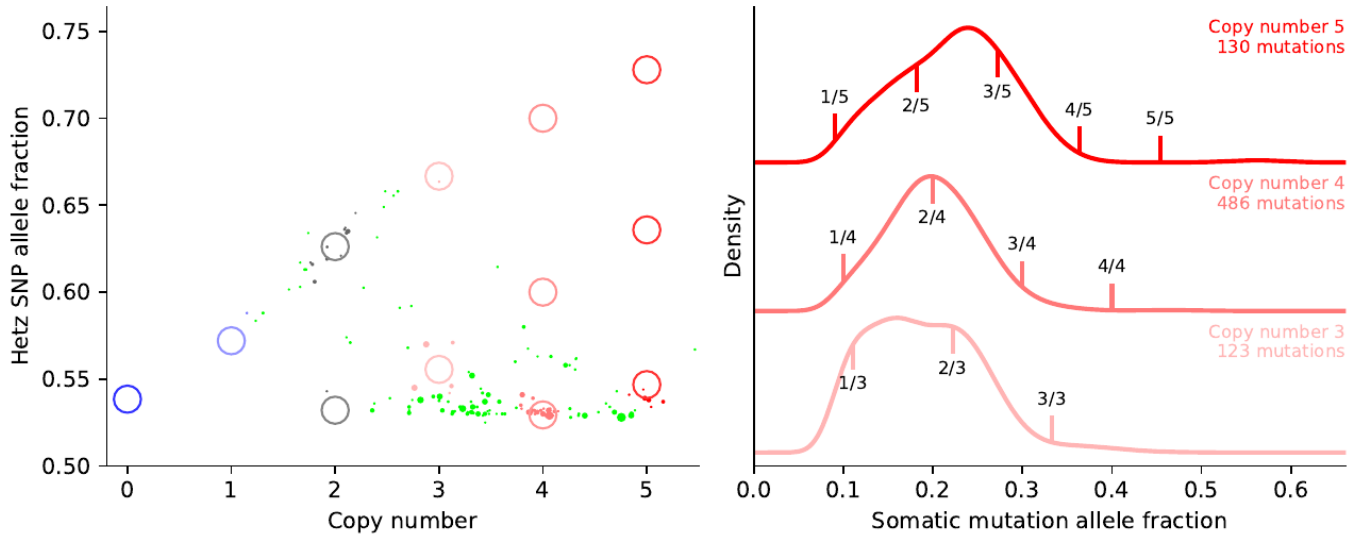
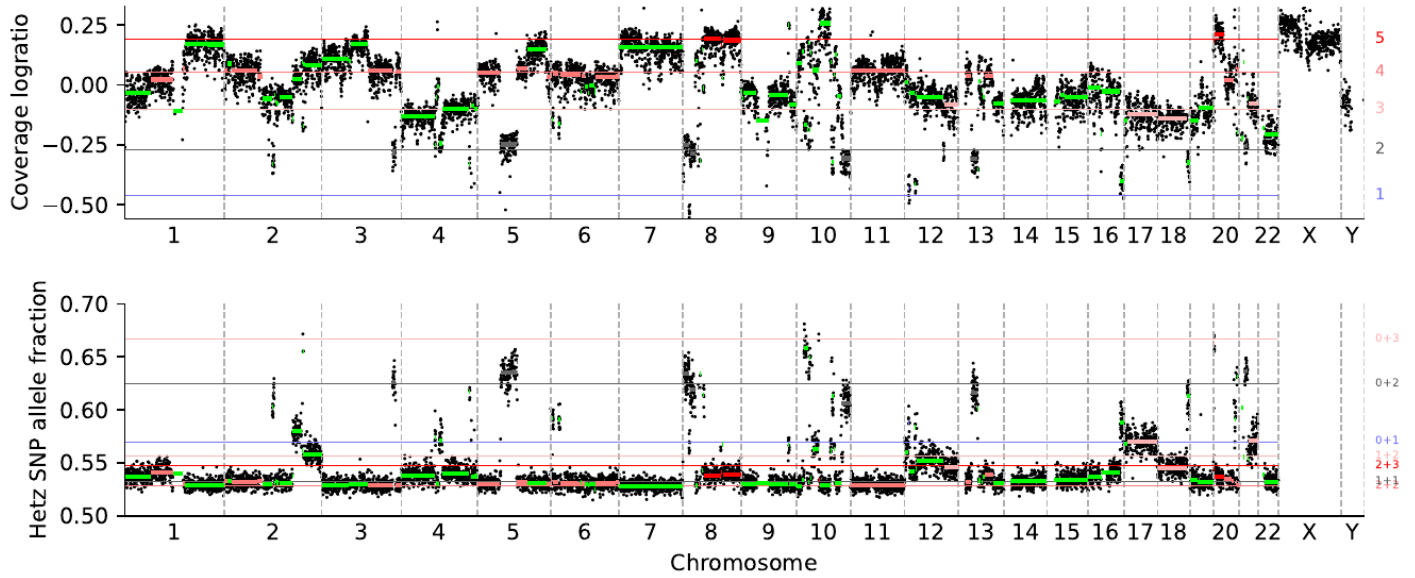
DTB-258-Baseline-Tumor

Cancer fraction: 79%, Diploid level: -0.64

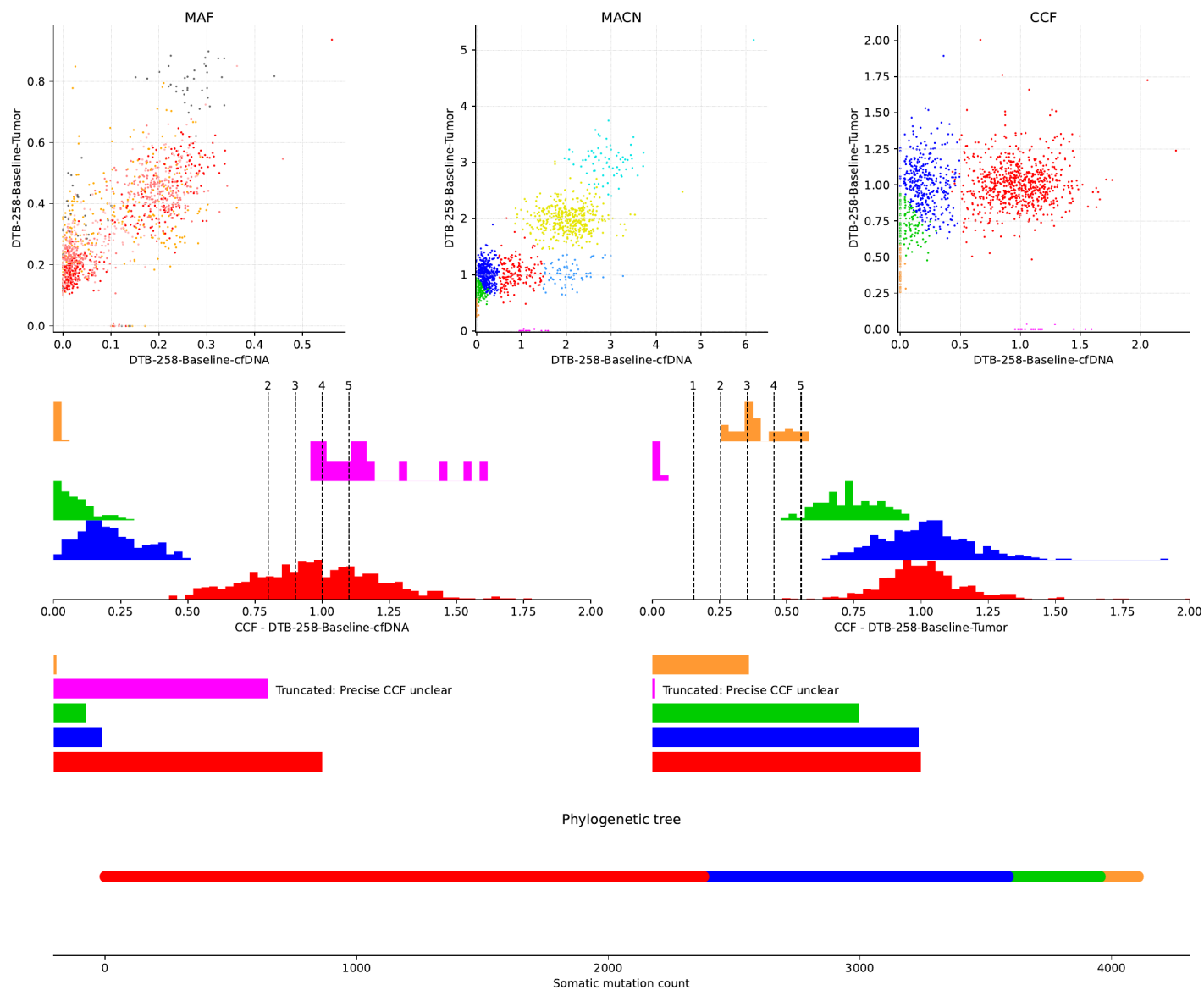


DTB-258-Baseline-cfDNA

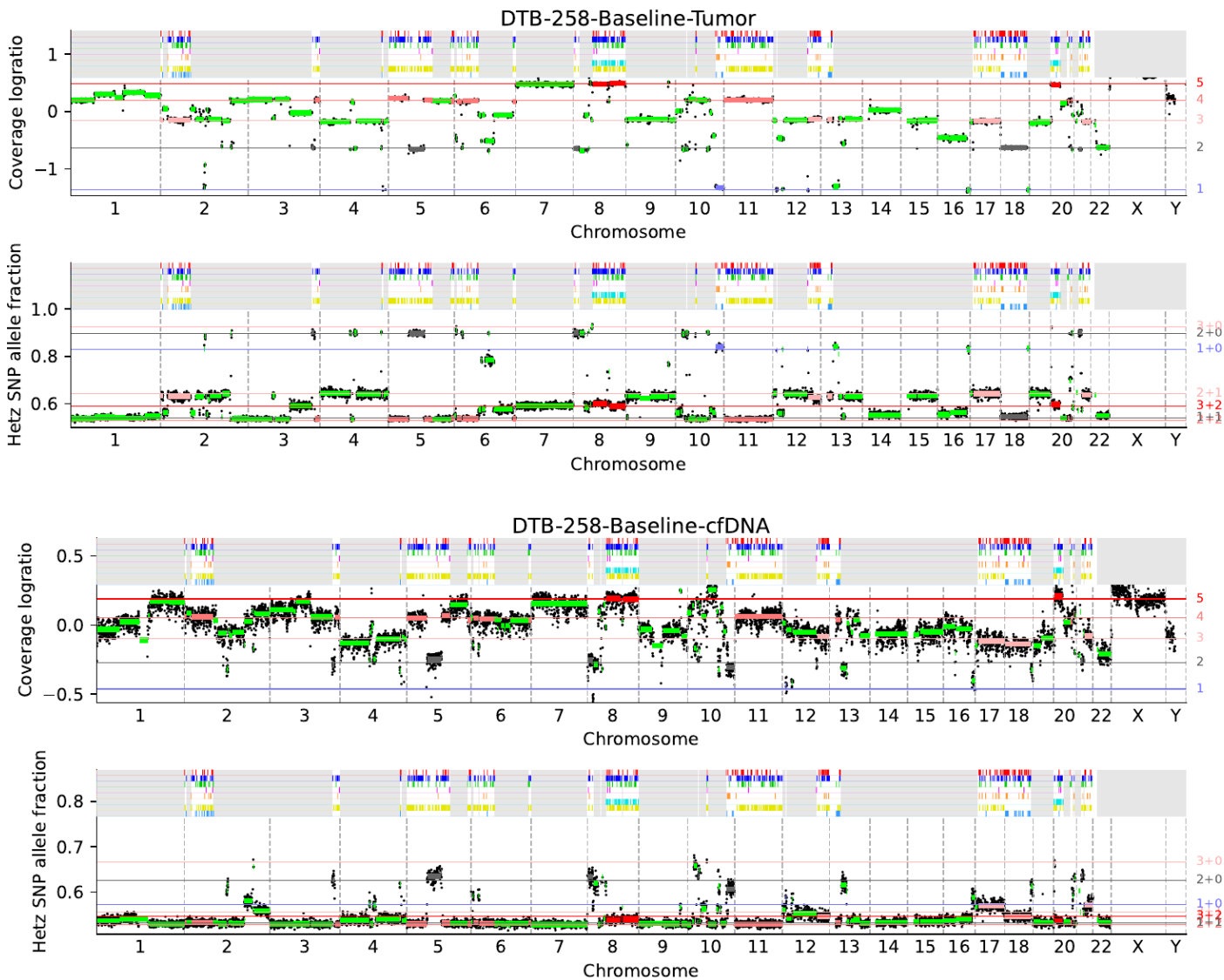
Cancer fraction: 25%, Diploid level: -0.27



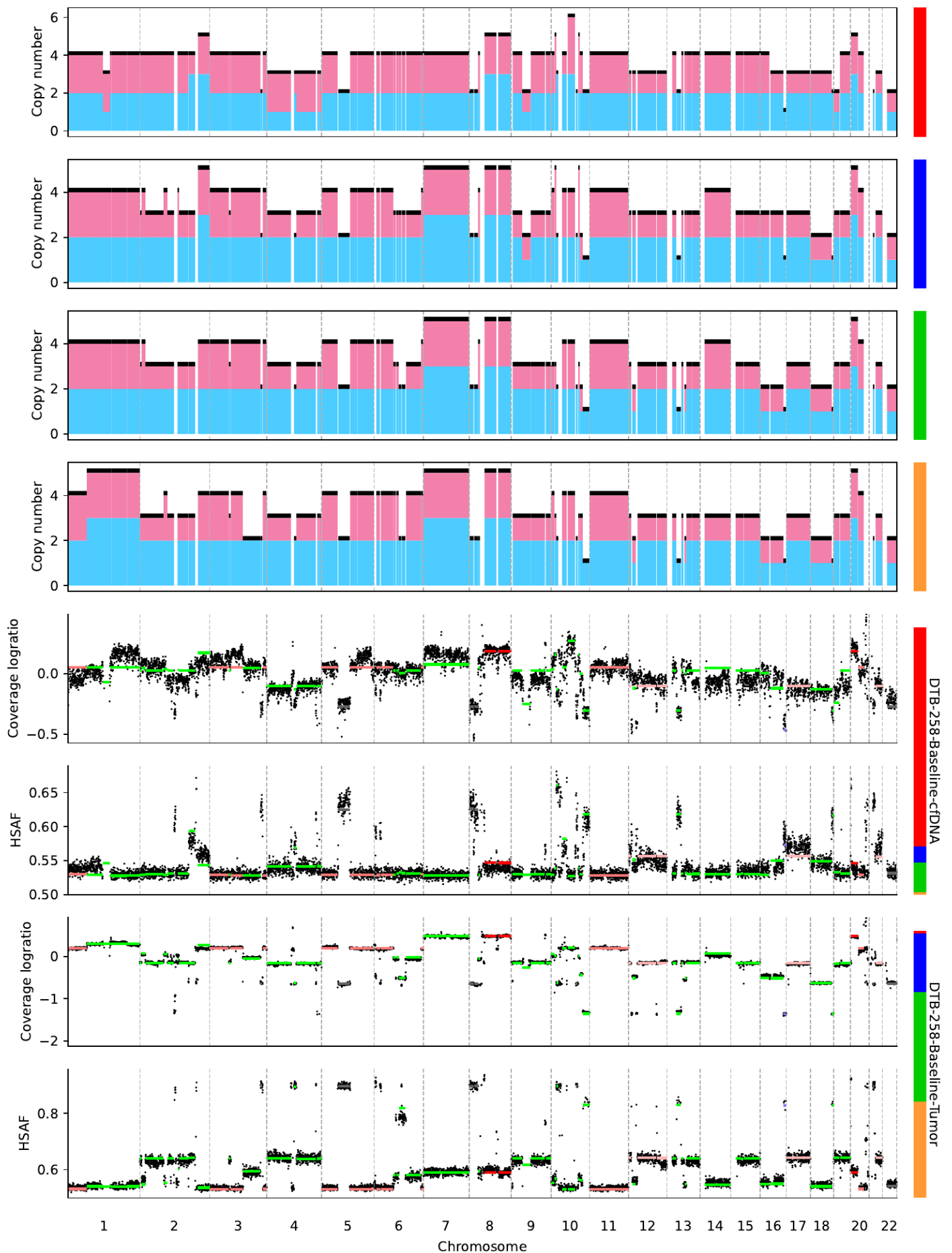
DTB-258 mutation clustering



DTB-258 mutation cluster positions



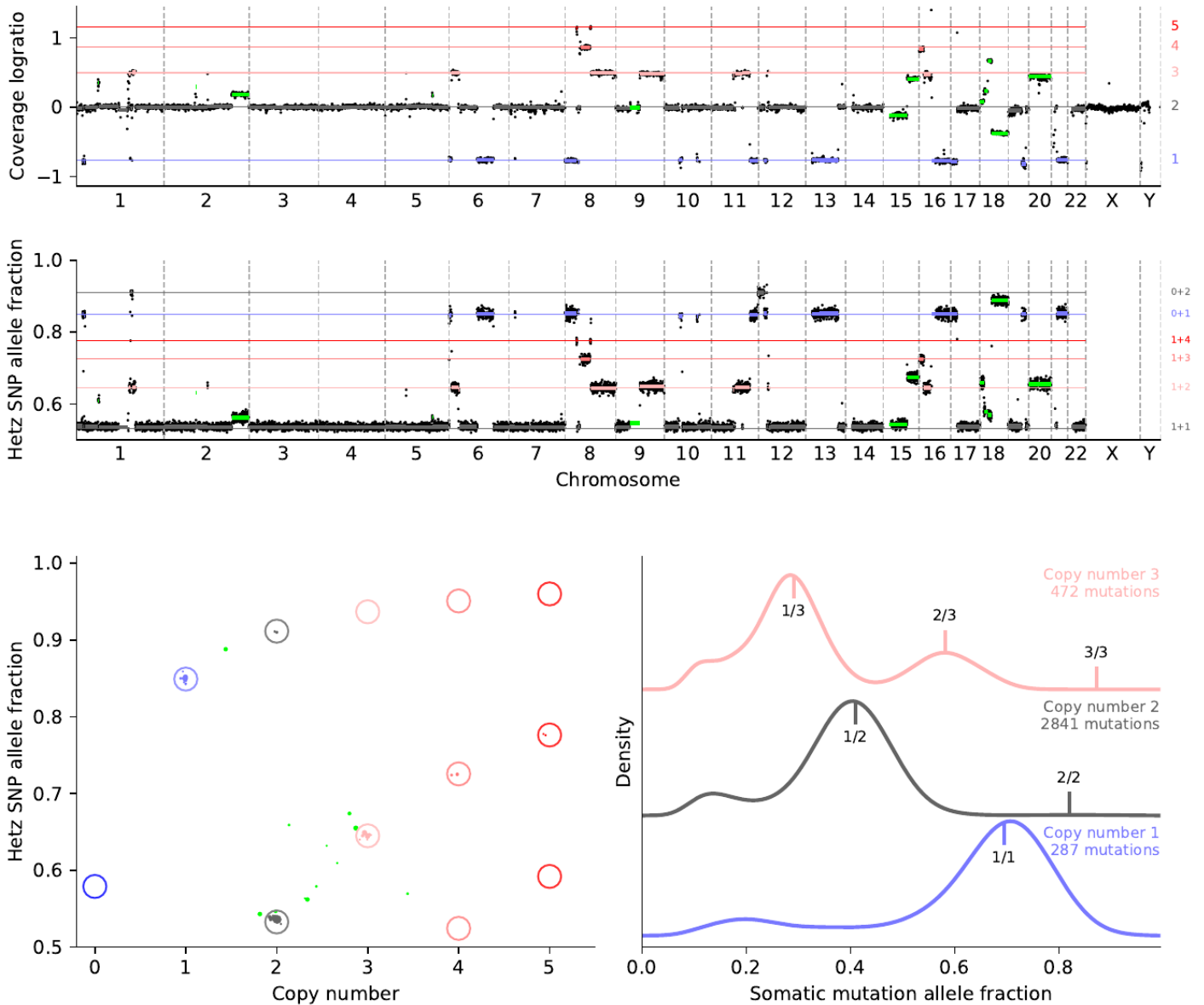
DTB-258 copy number deconvolution



Patient DTB-261

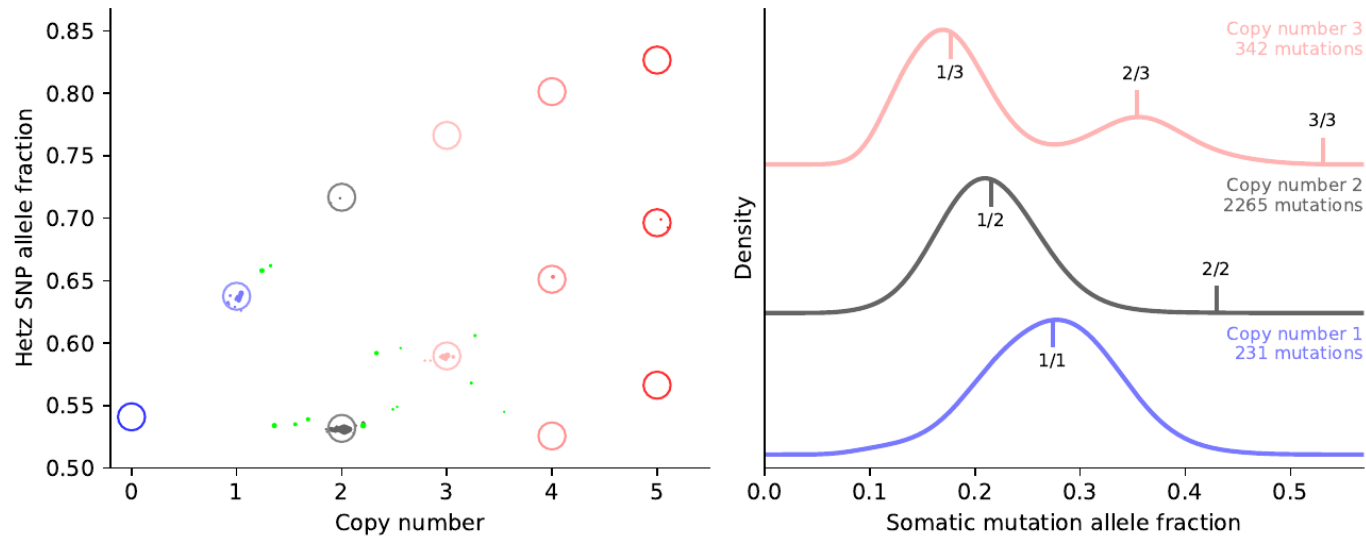
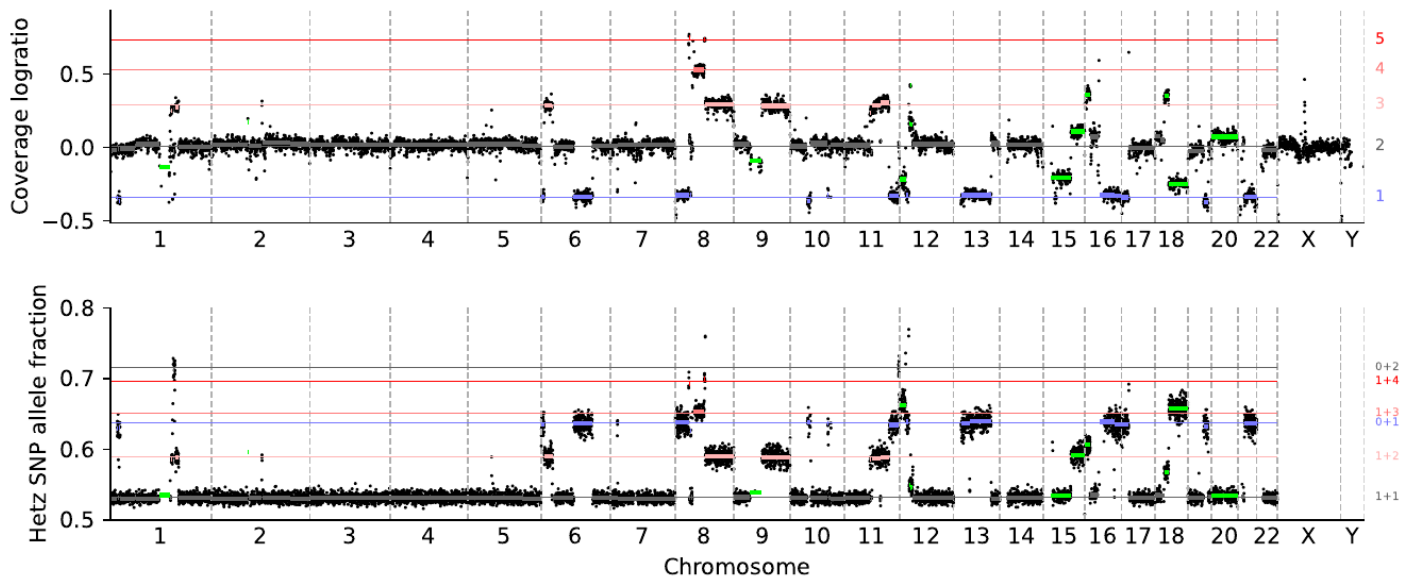
DTB-261-Baseline-Tumor

Cancer fraction: 82%, Diploid level: 0.00



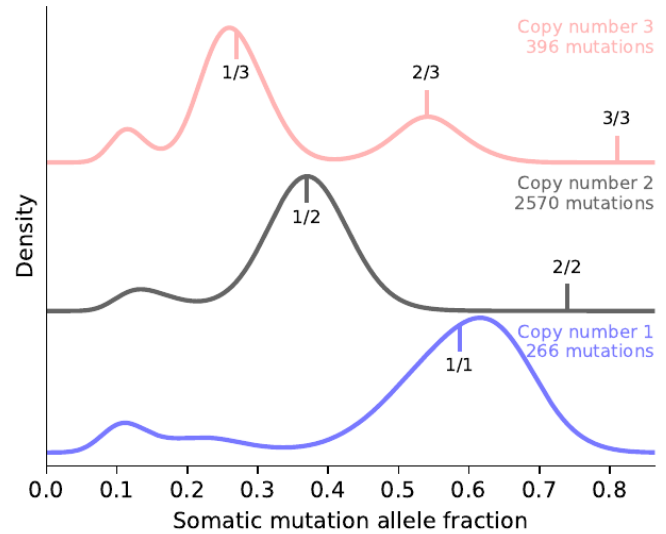
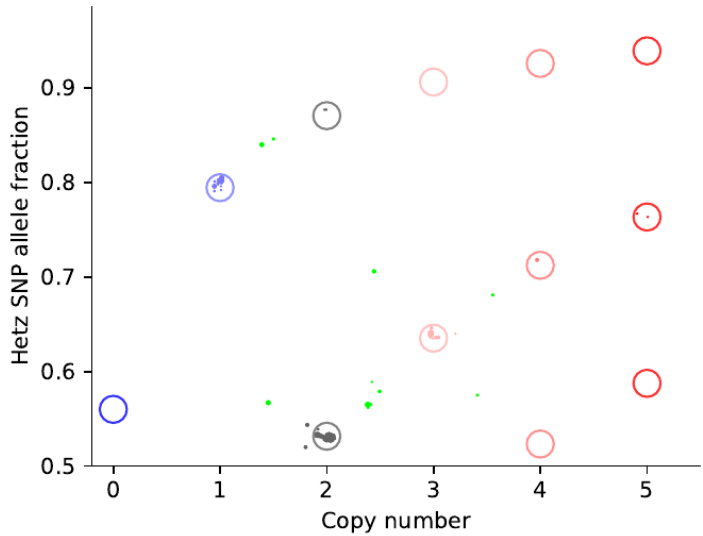
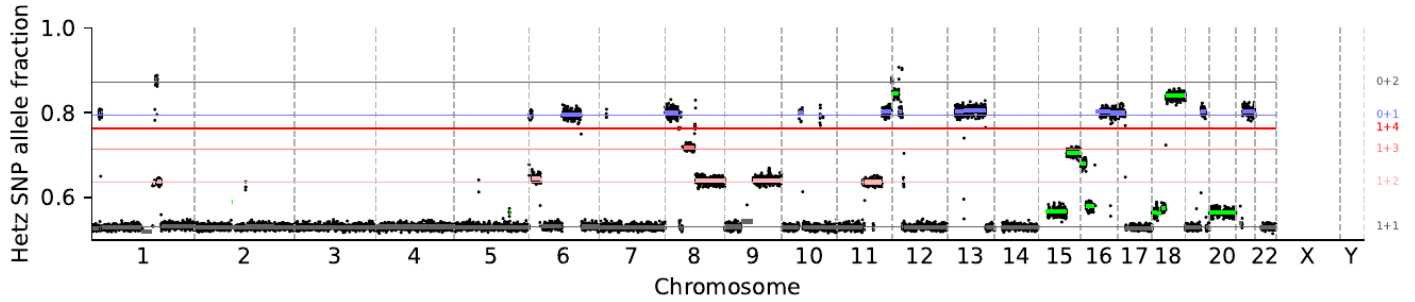
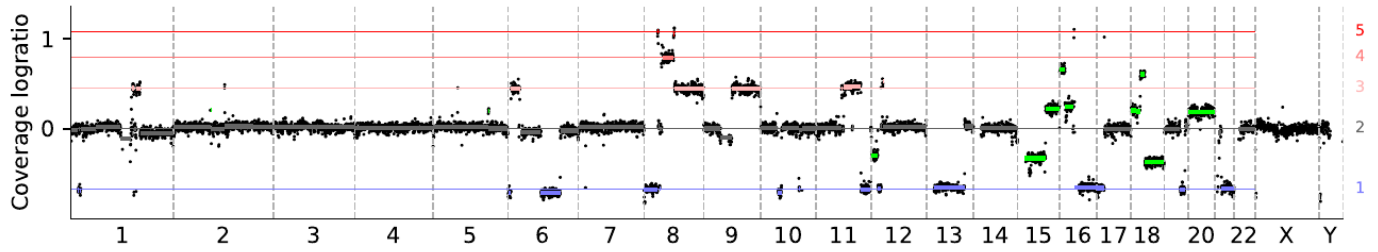
DTB-261-Baseline-cfDNA

Cancer fraction: 43%, Diploid level: 0.01

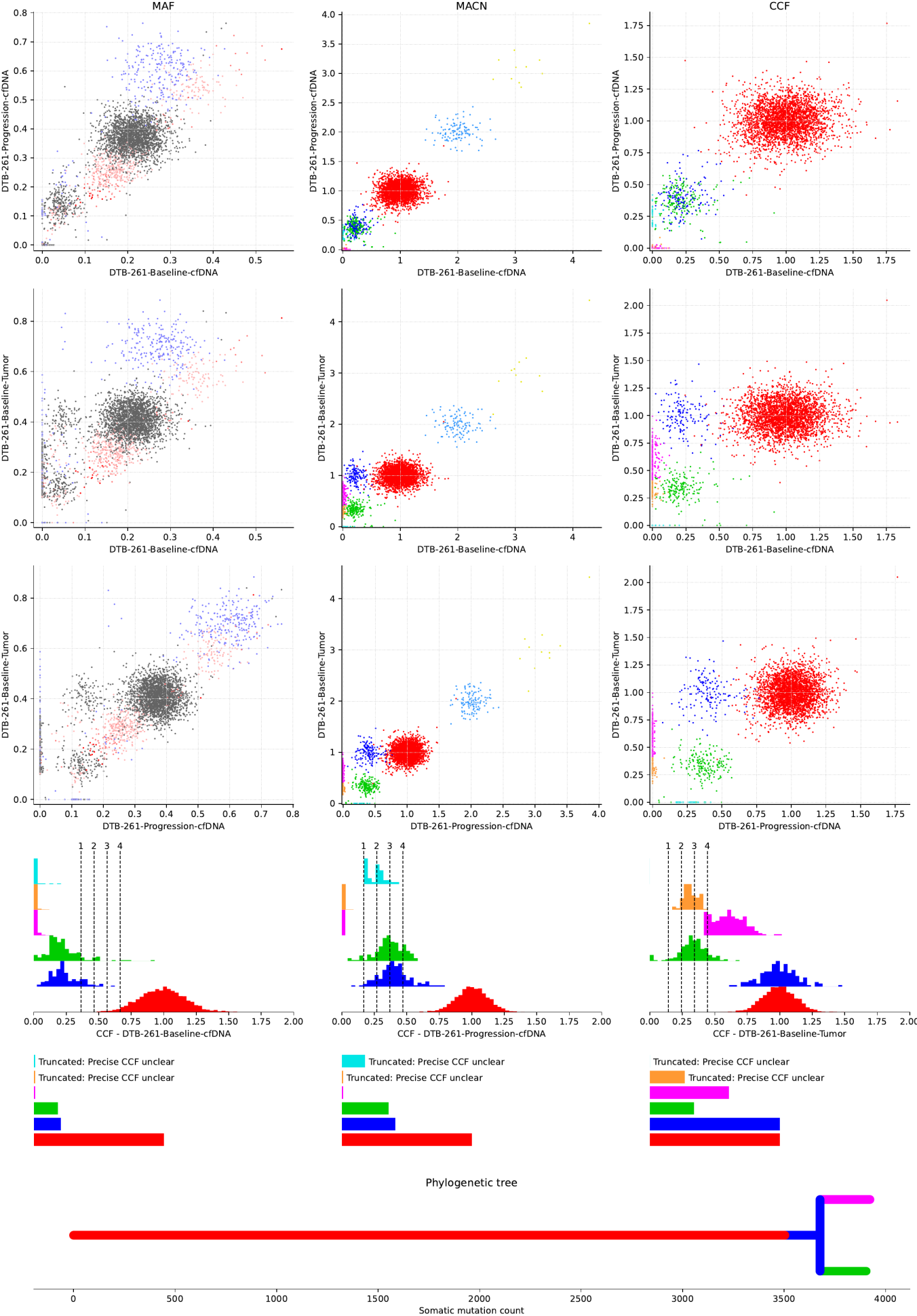


DTB-261-Progression-cfDNA

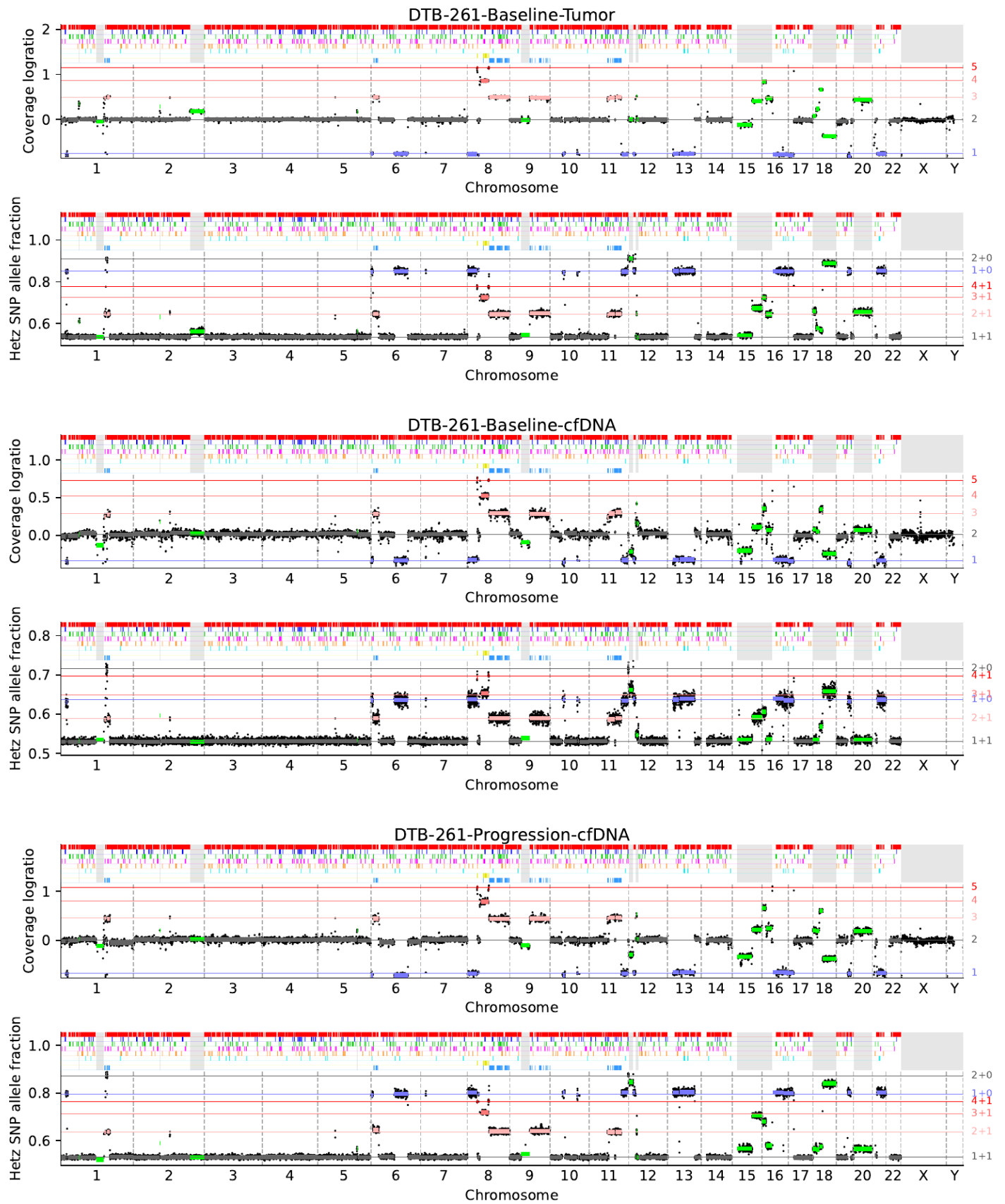
Cancer fraction: 74%, Diploid level: 0.00



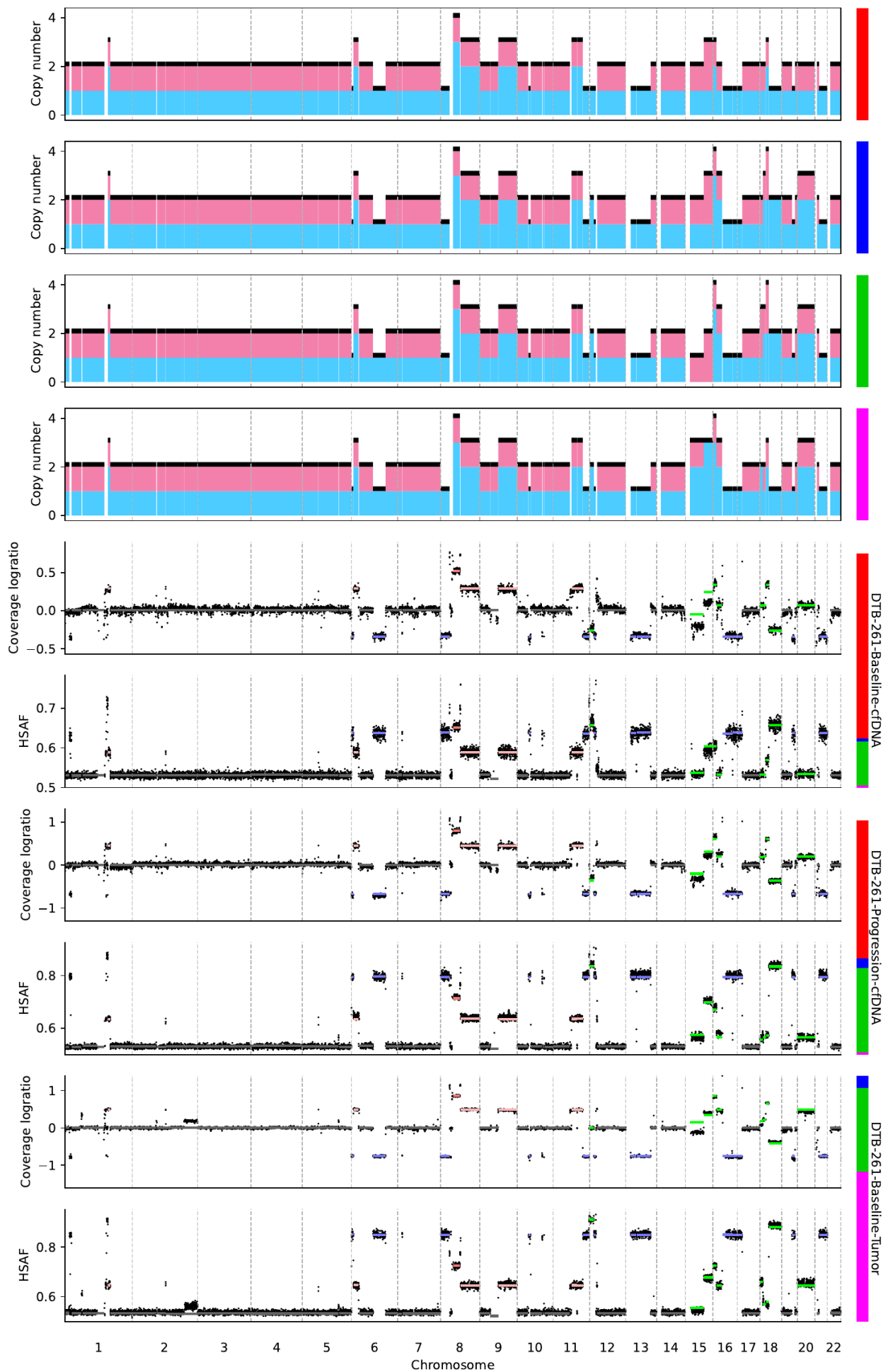
DTB-261 mutation clustering



DTB-261 mutation cluster positions



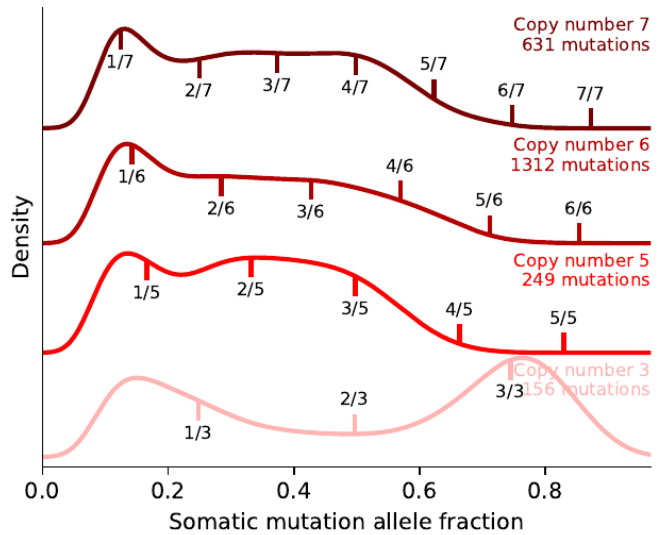
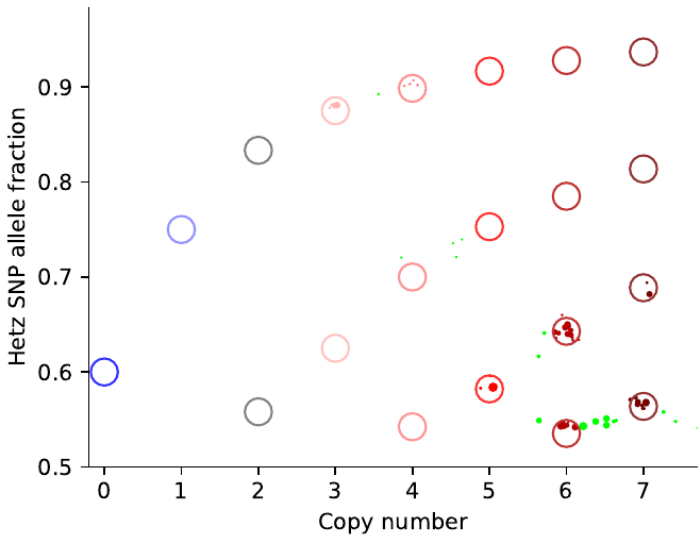
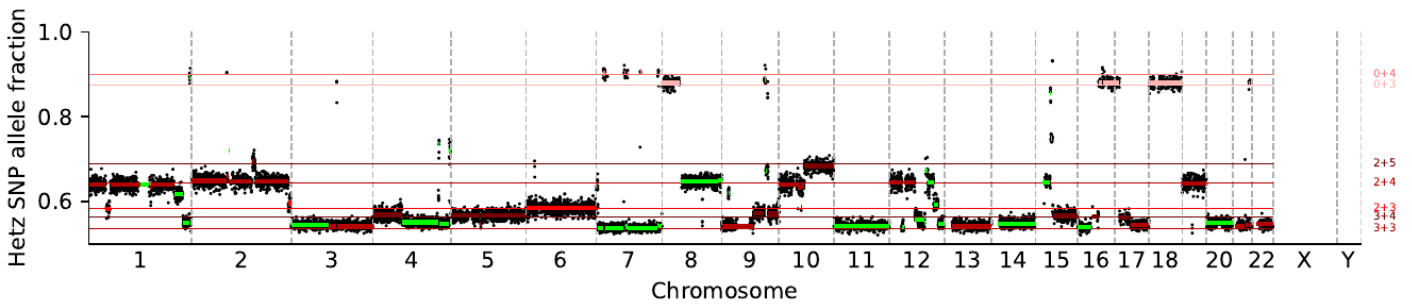
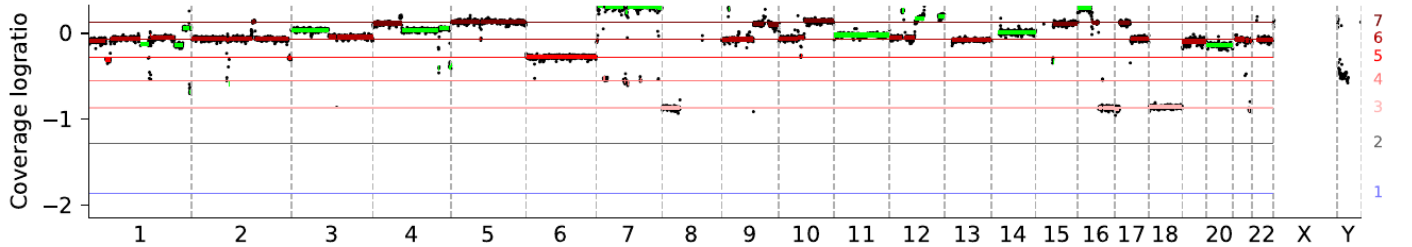
DTB-261 copy number deconvolution



Patient DTB-266

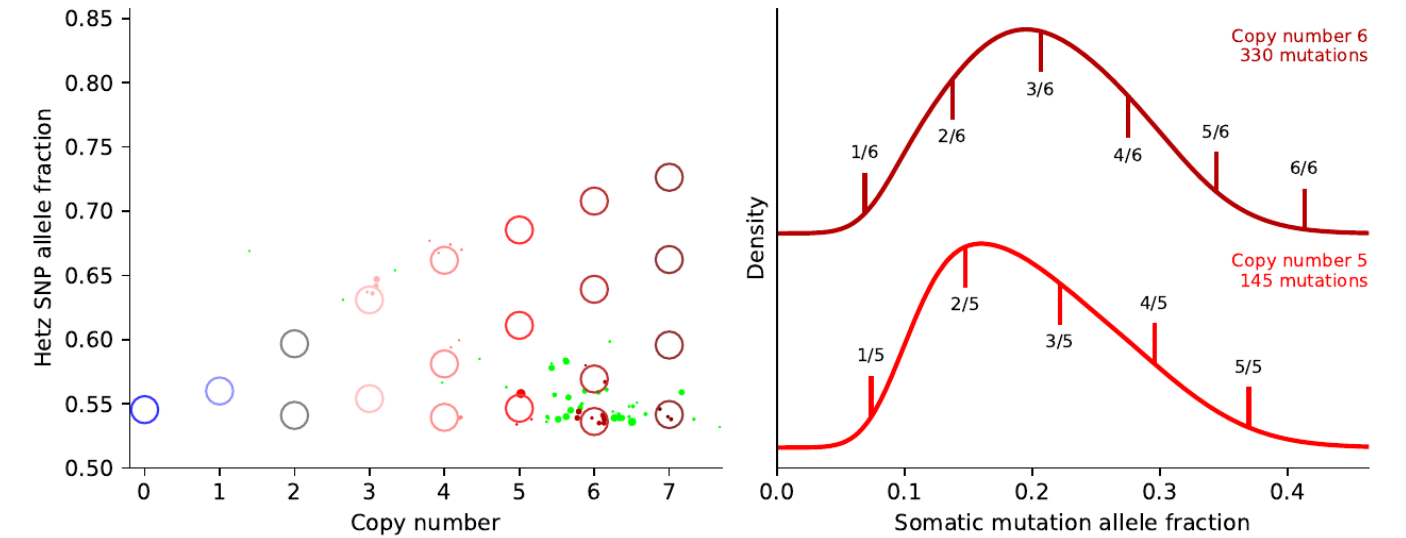
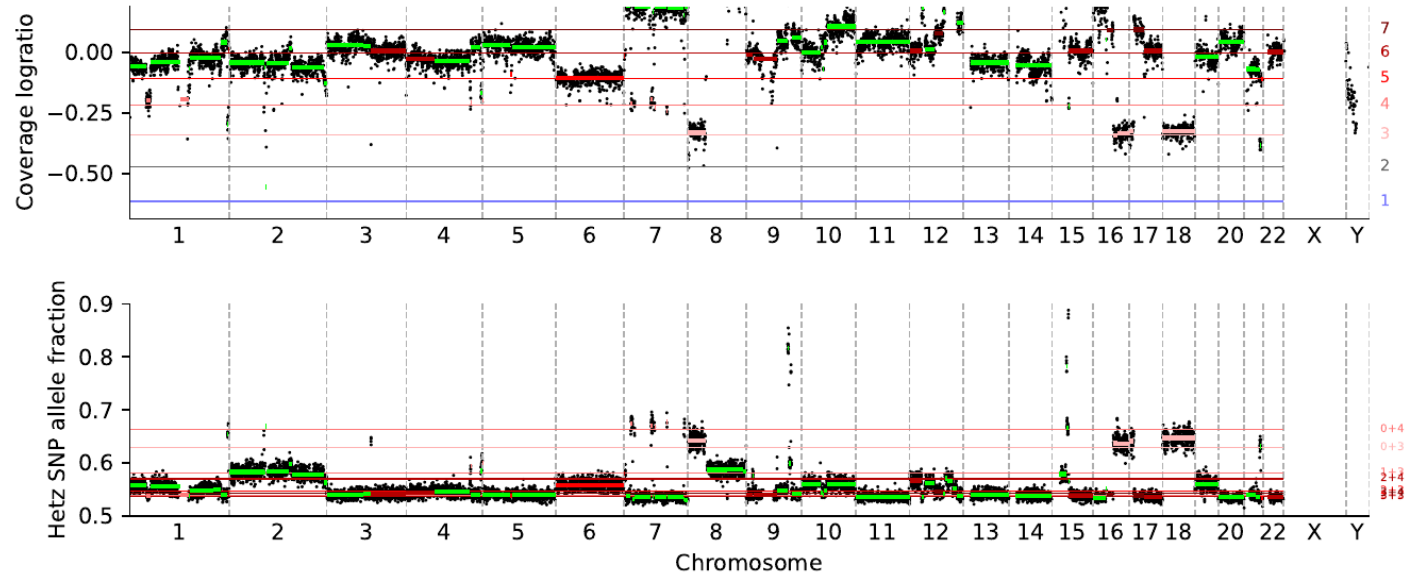
DTB-266-Baseline-Tumor

Cancer fraction: 66%, Diploid level: -1.28

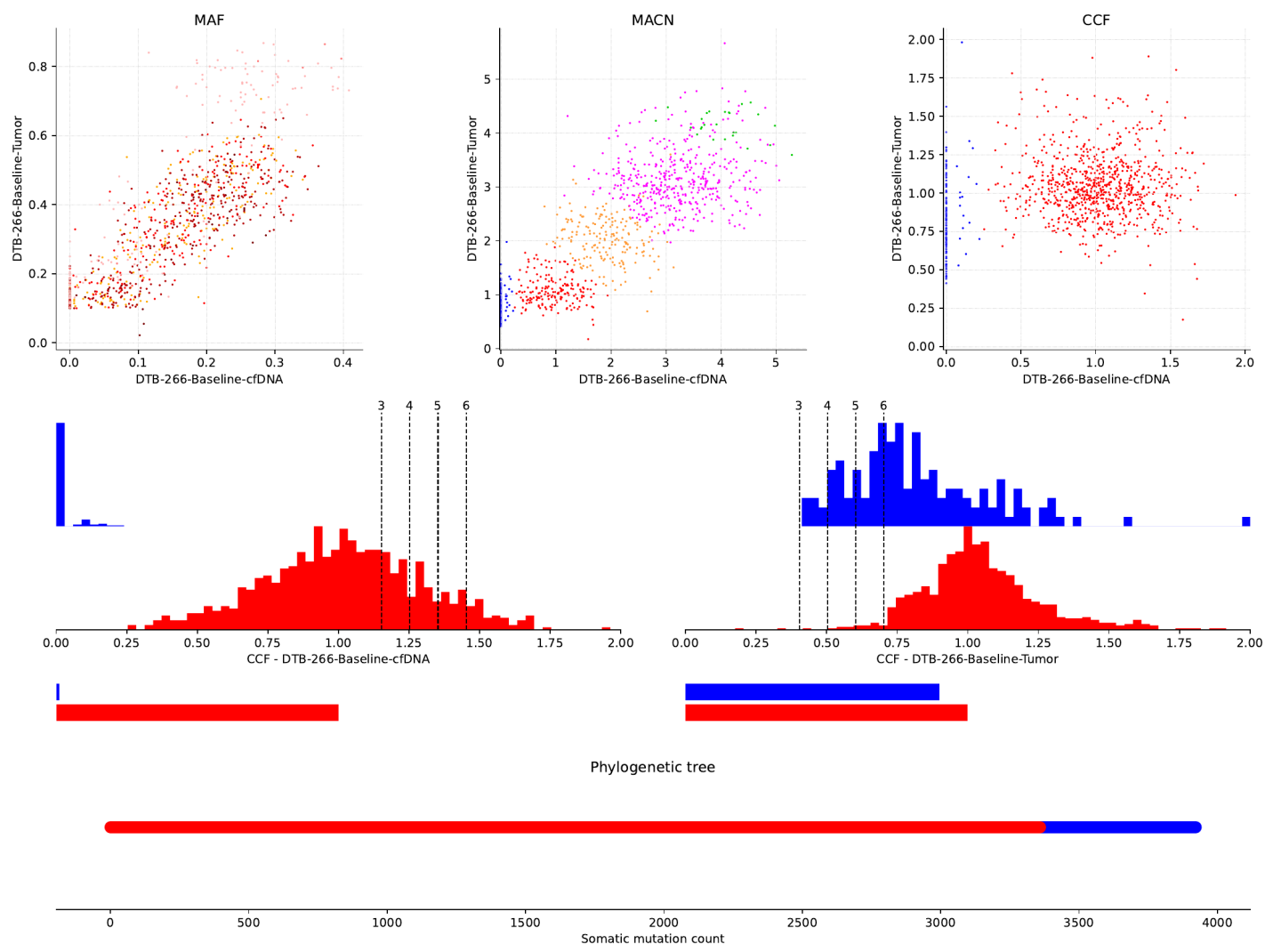


DTB-266-Baseline-cfDNA

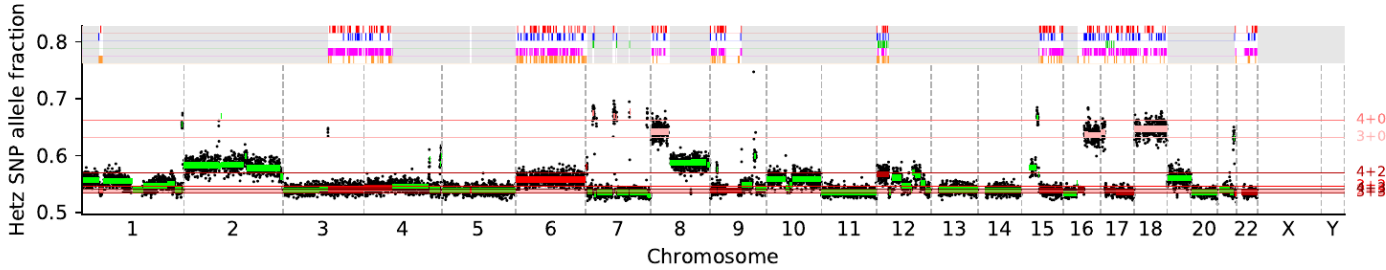
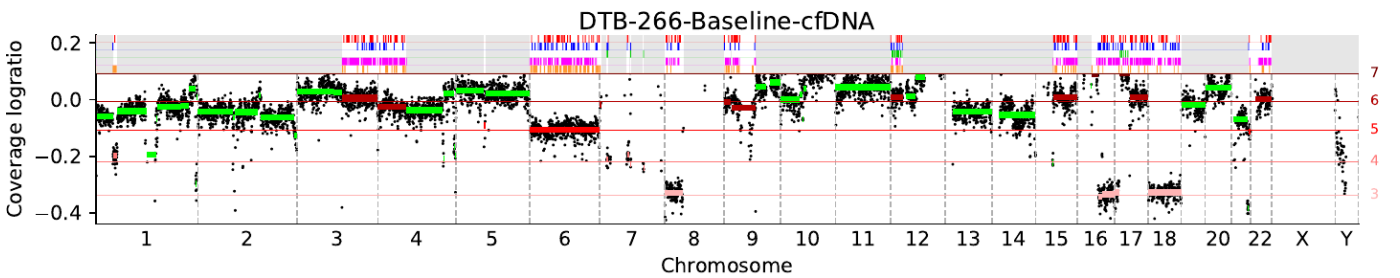
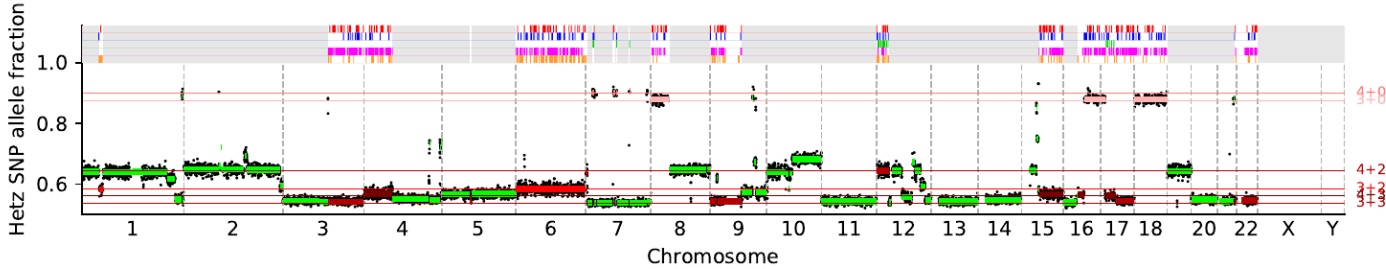
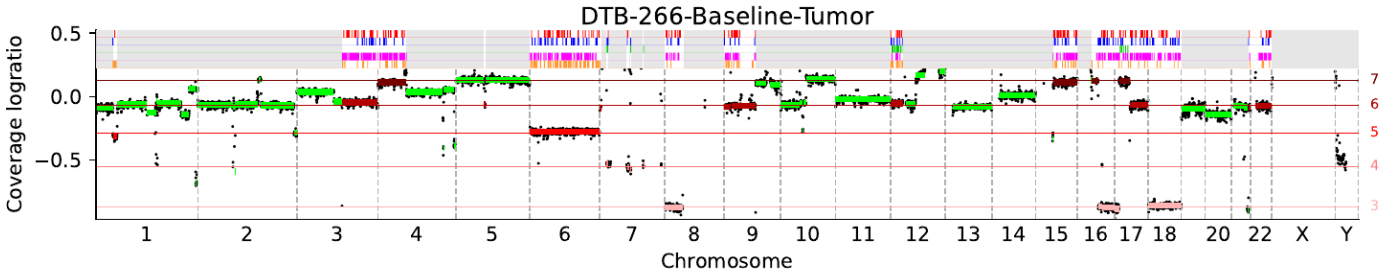
Cancer fraction: 19%, Diploid level: -0.47



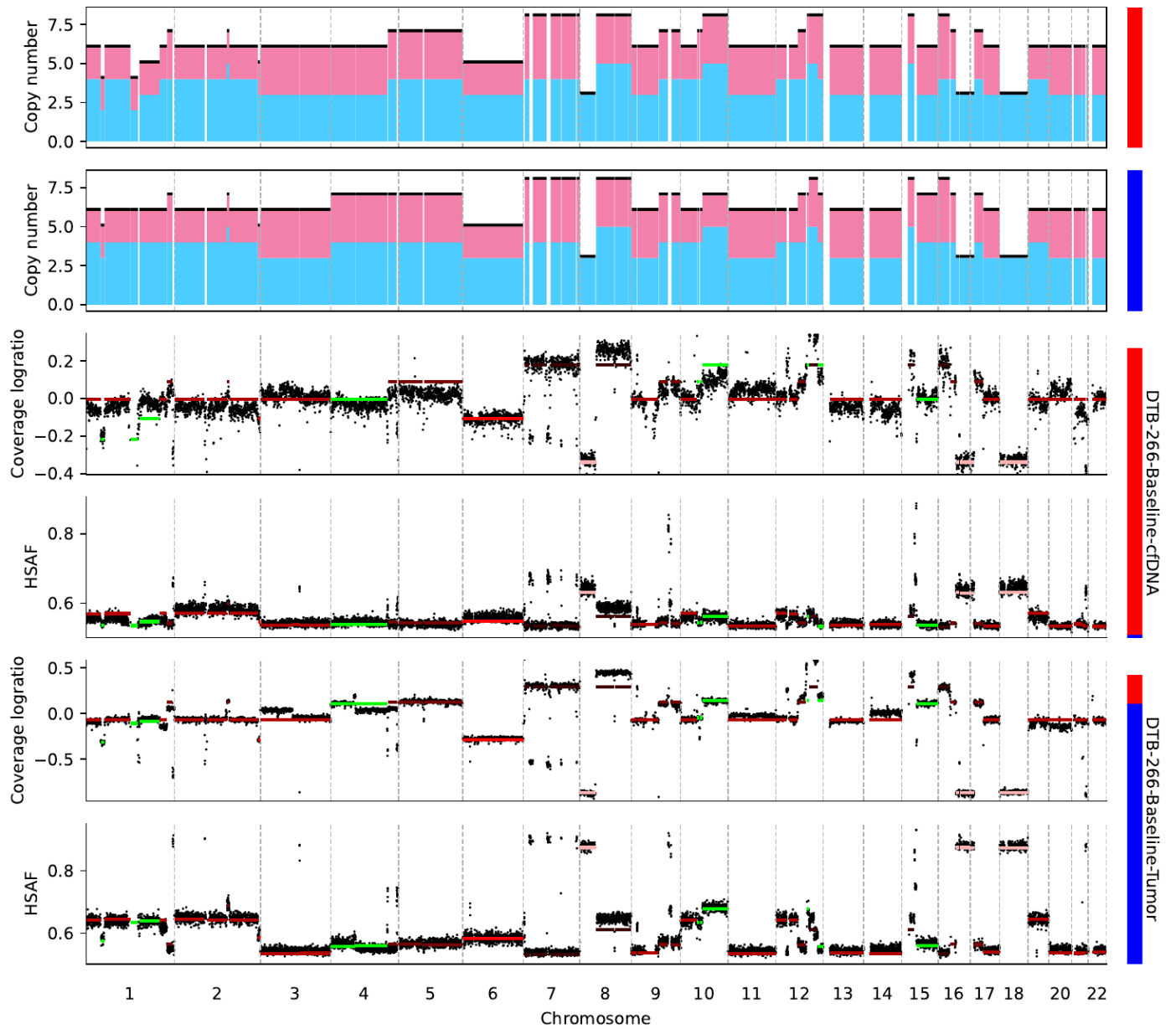
DTB-266 mutation clustering



DTB-266 mutation cluster positions



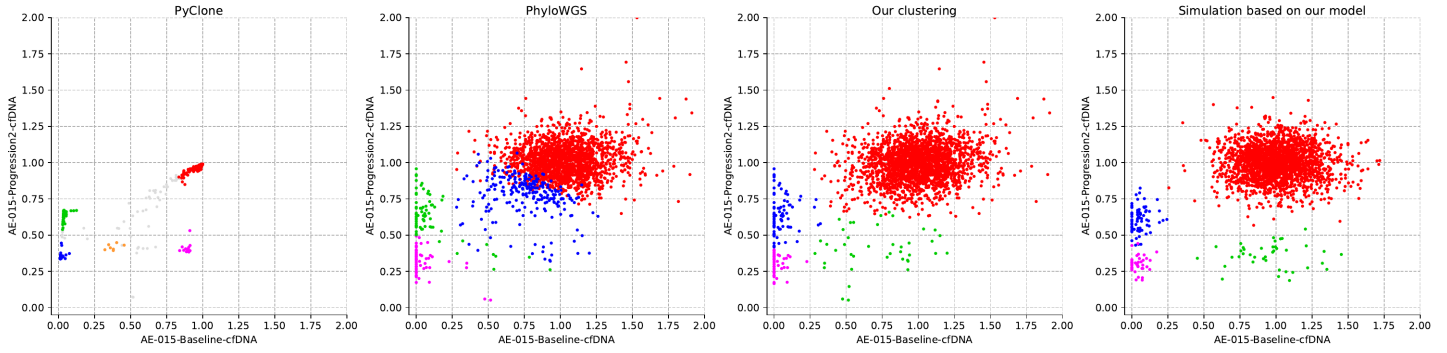
DTB-266 copy number deconvolution



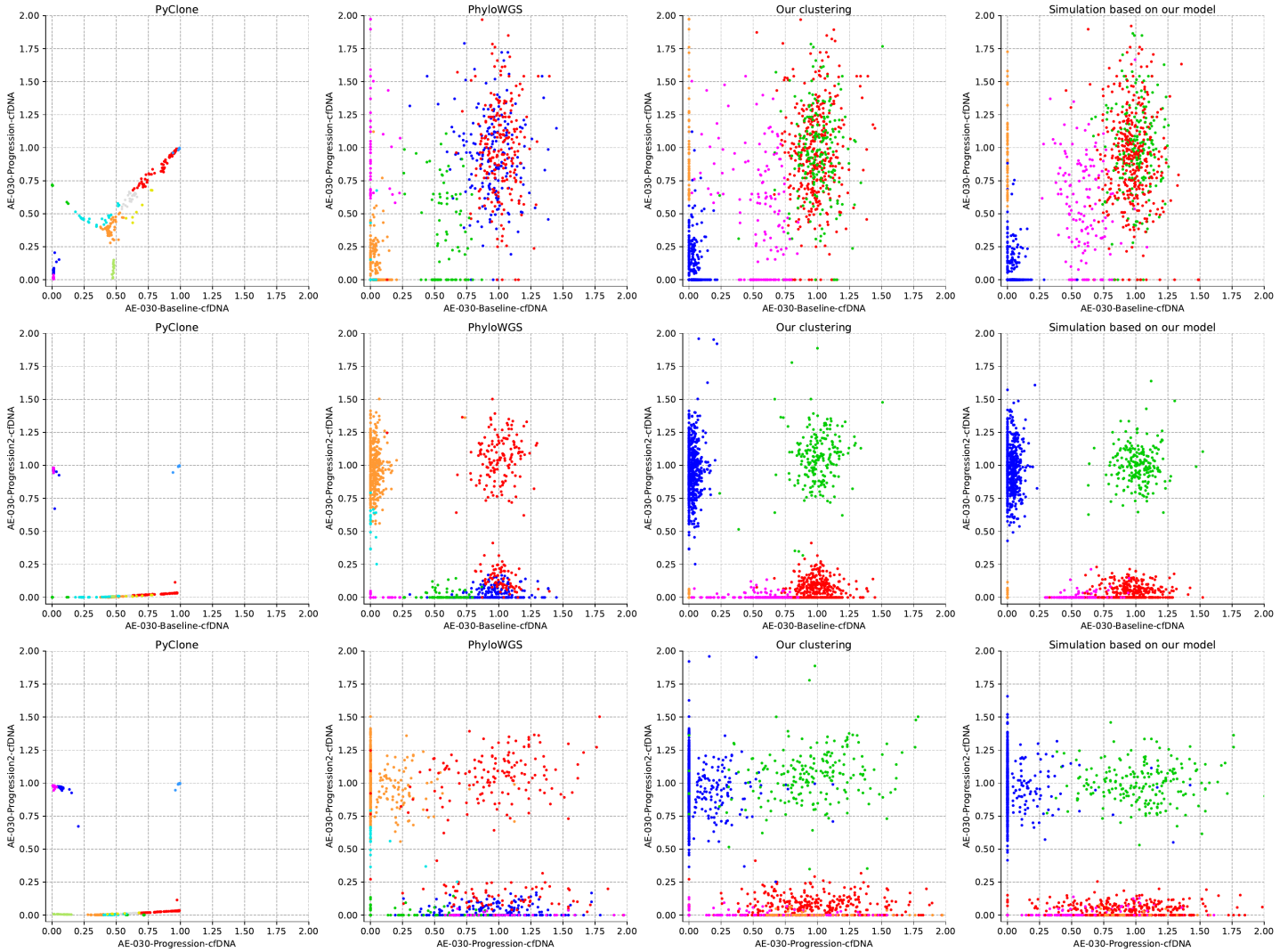
PyClone and PhyloWGS subclonal reconstruction validation data

In this section we provide a side-by-side comparison of the somatic mutation clustering step carried out by our algorithm, PyClone, and PhyloWGS in order to identify subclones and their CCFs in patient samples. For each patient, we show mutation CCF scatterplots generated by the three approaches, for all two-sample combinations.

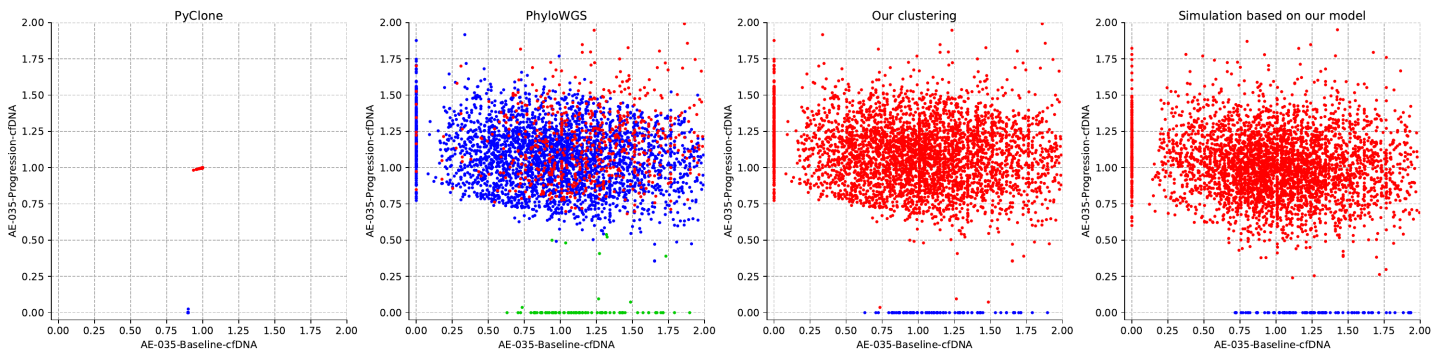
Patient AE-015



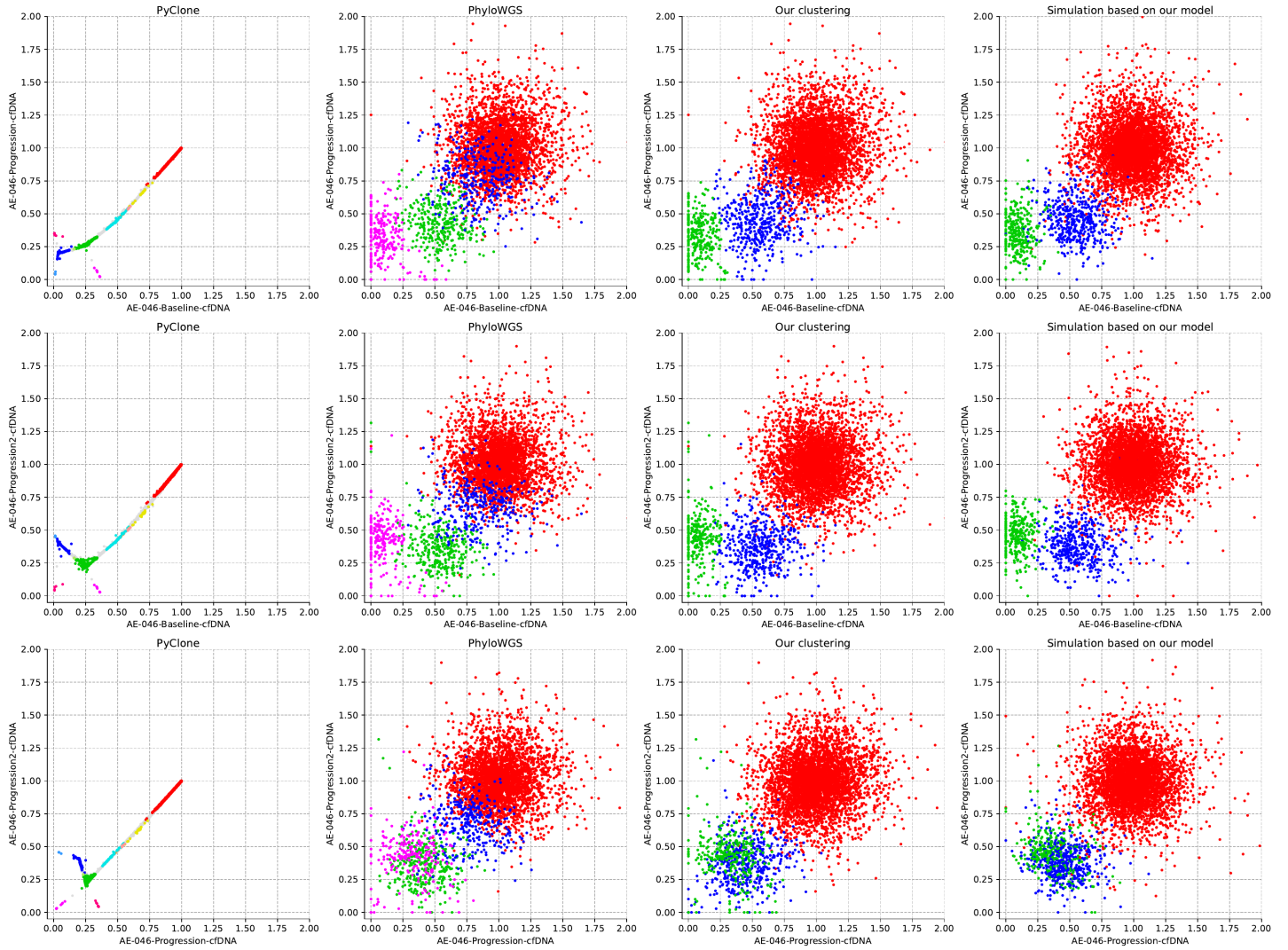
Patient AE-030



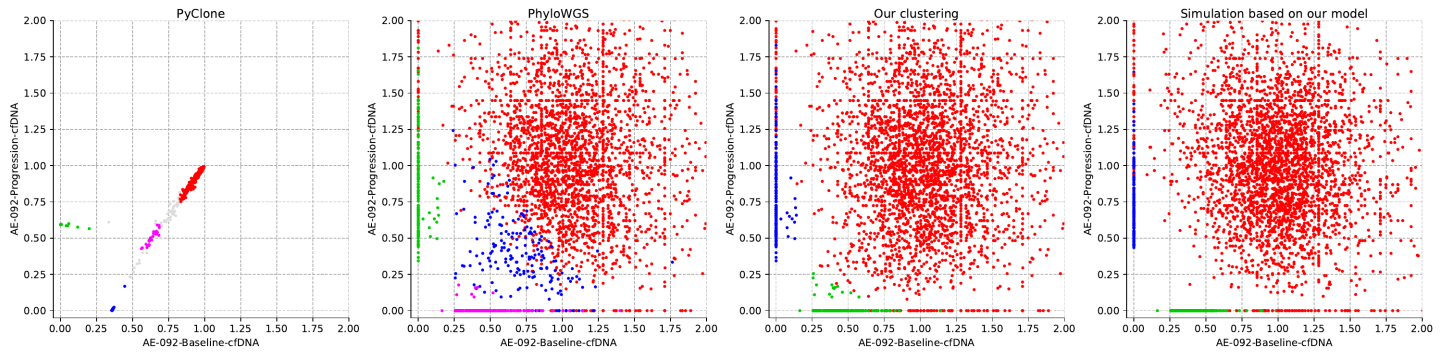
Patient AE-035



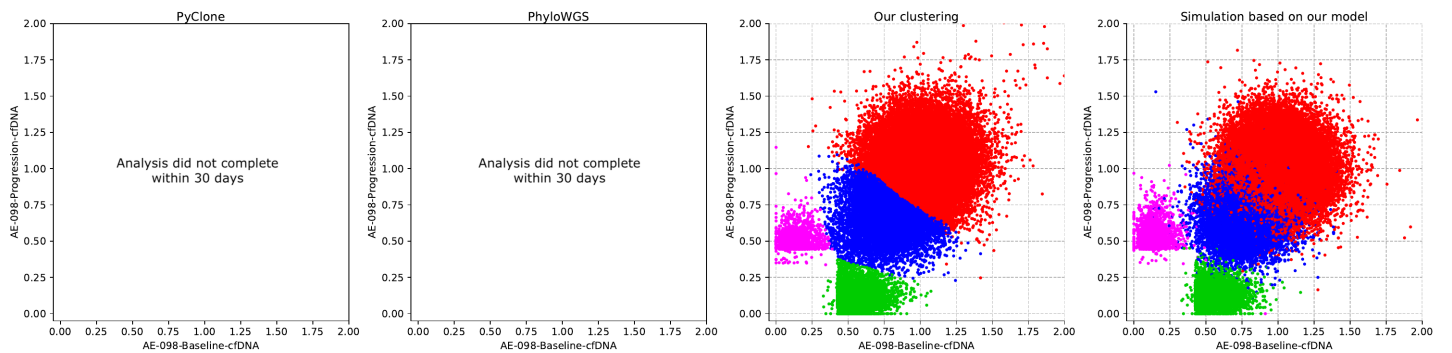
Patient AE-046



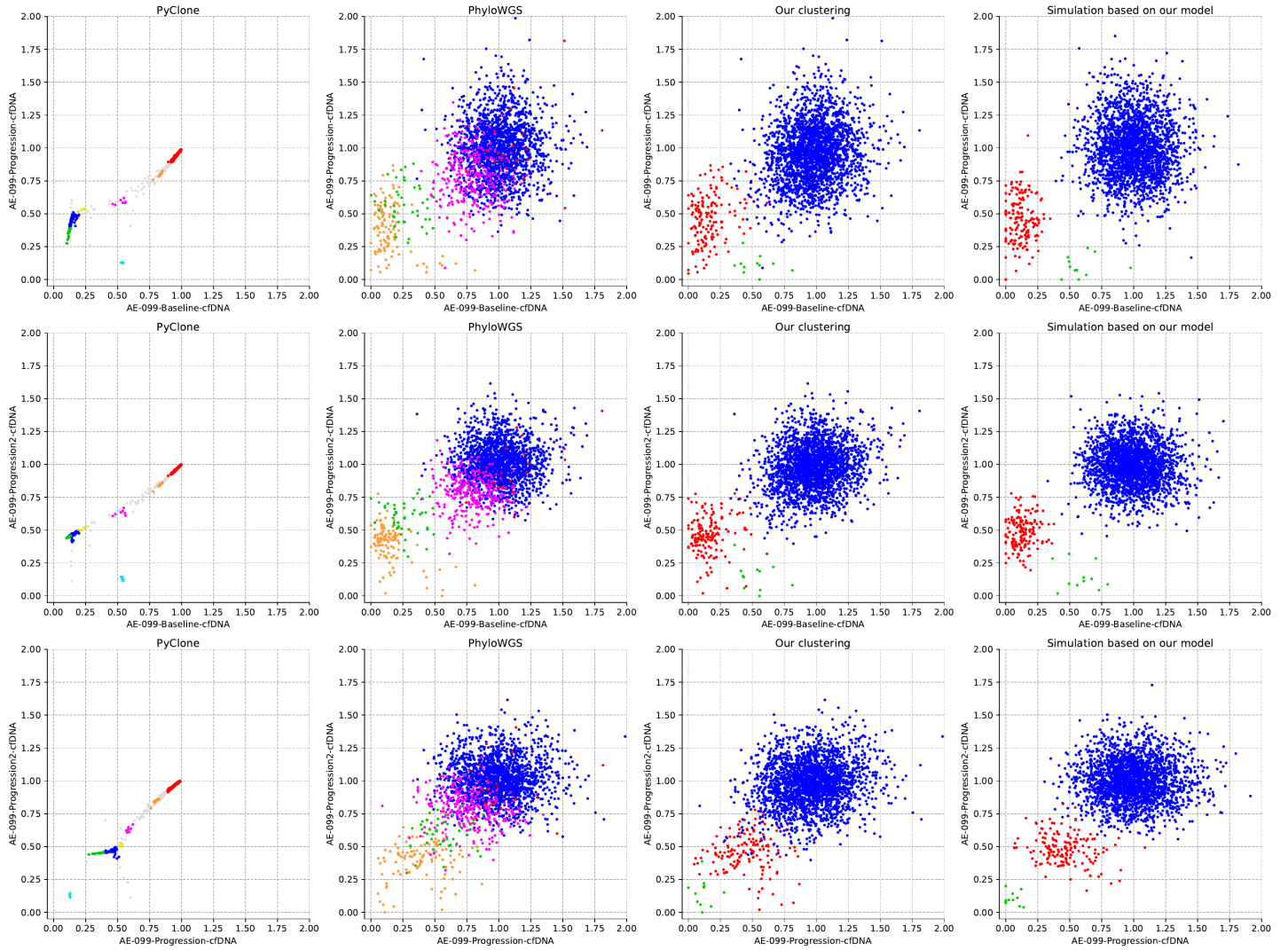
Patient AE-092



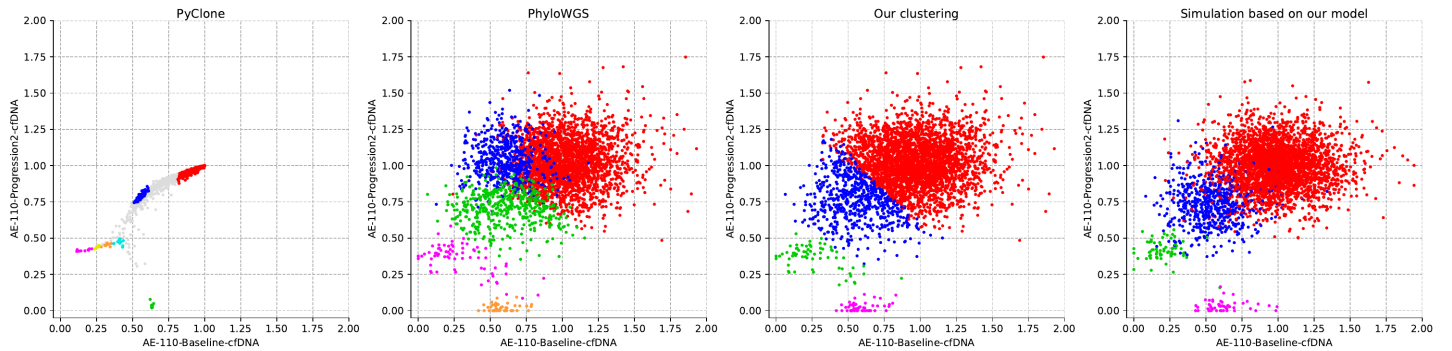
Patient AE-098



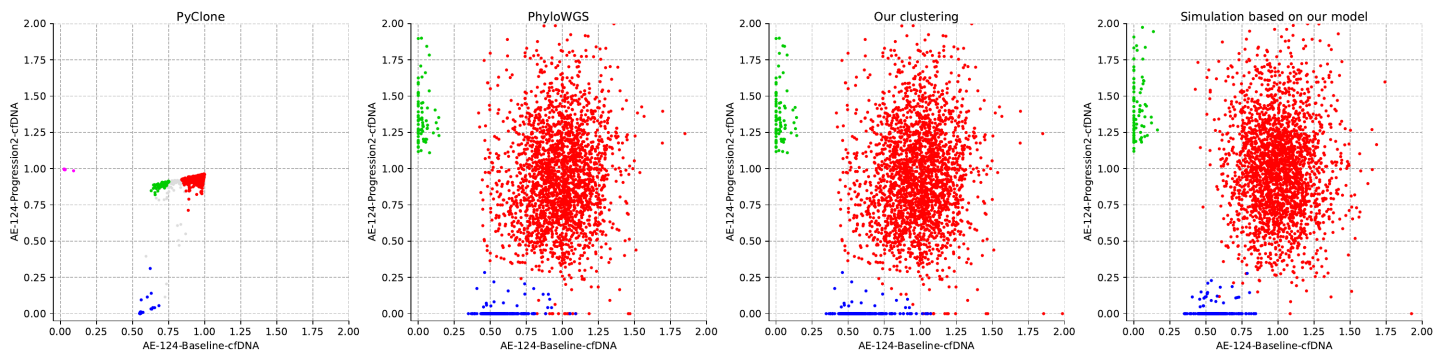
Patient AE-099



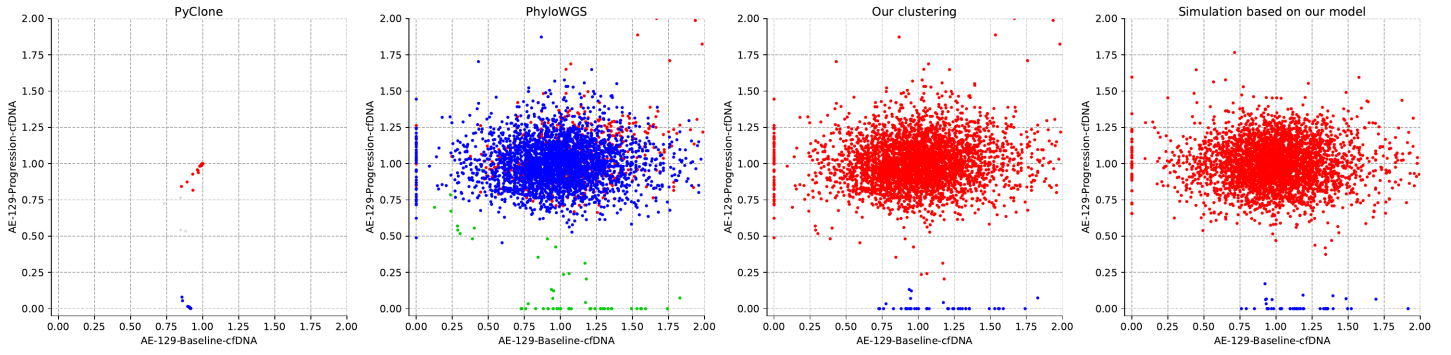
Patient AE-110



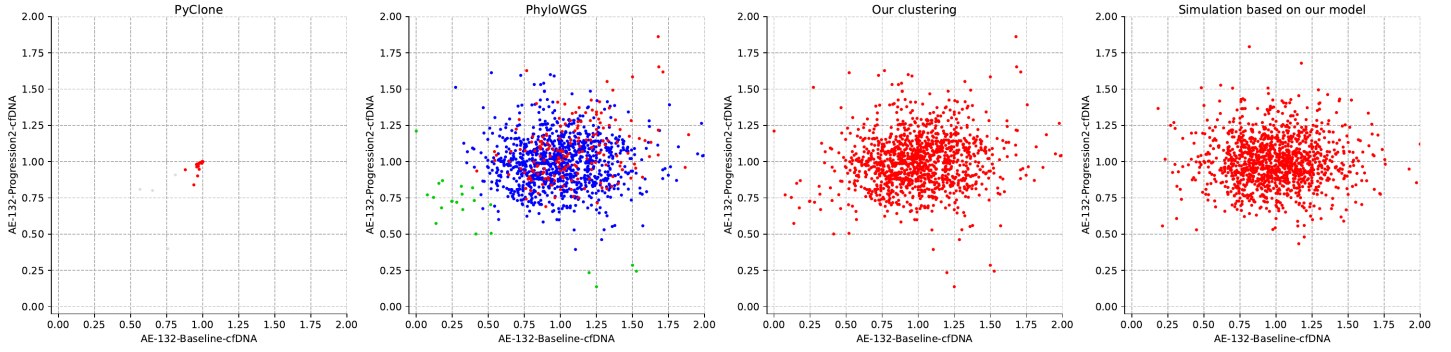
Patient AE-124



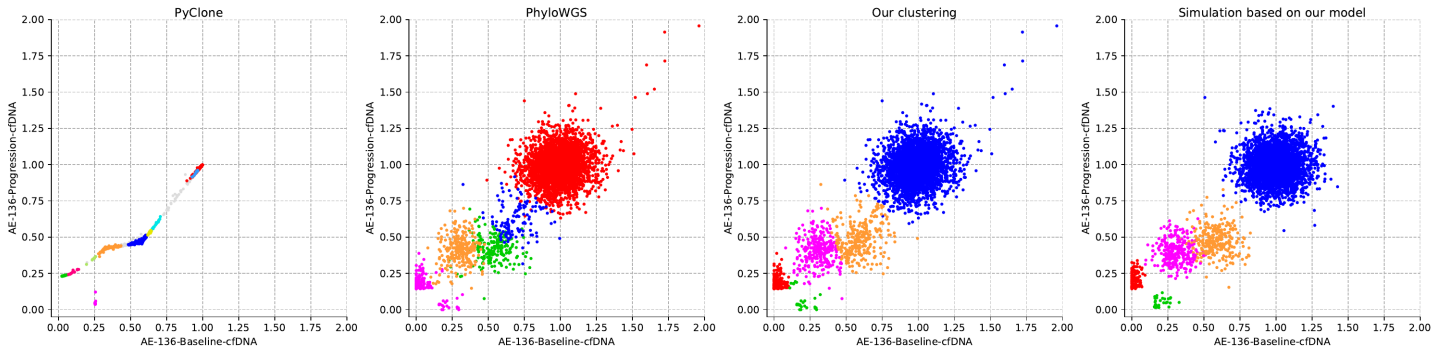
Patient AE-129



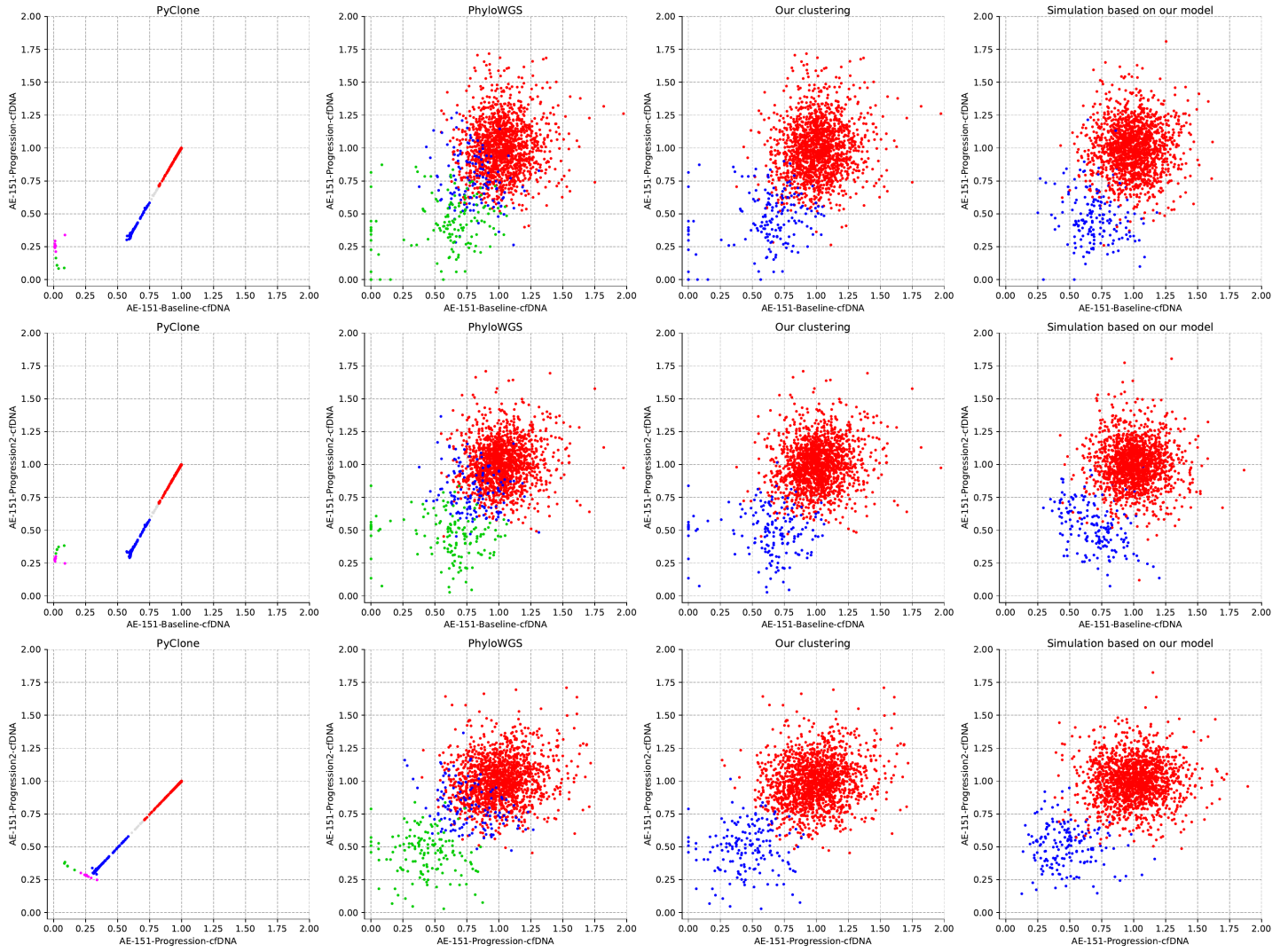
Patient AE-132



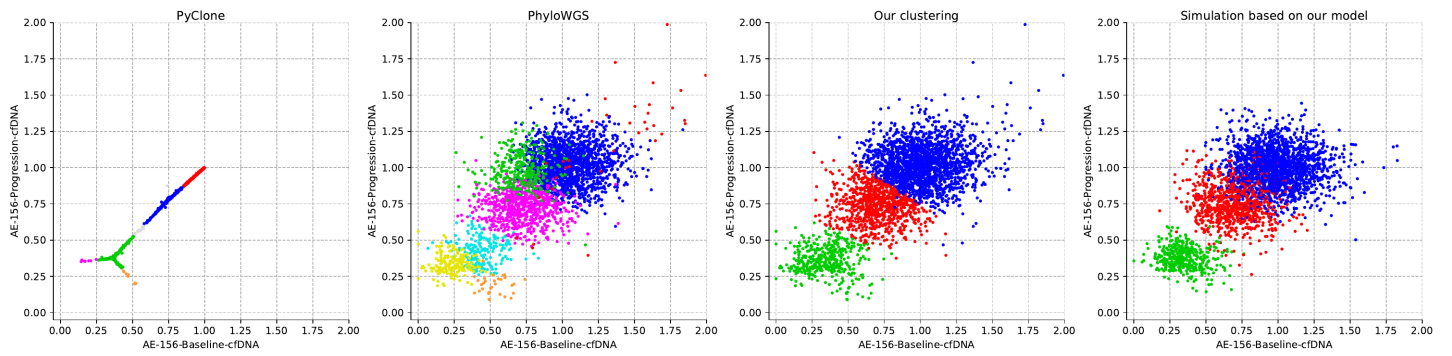
Patient AE-136



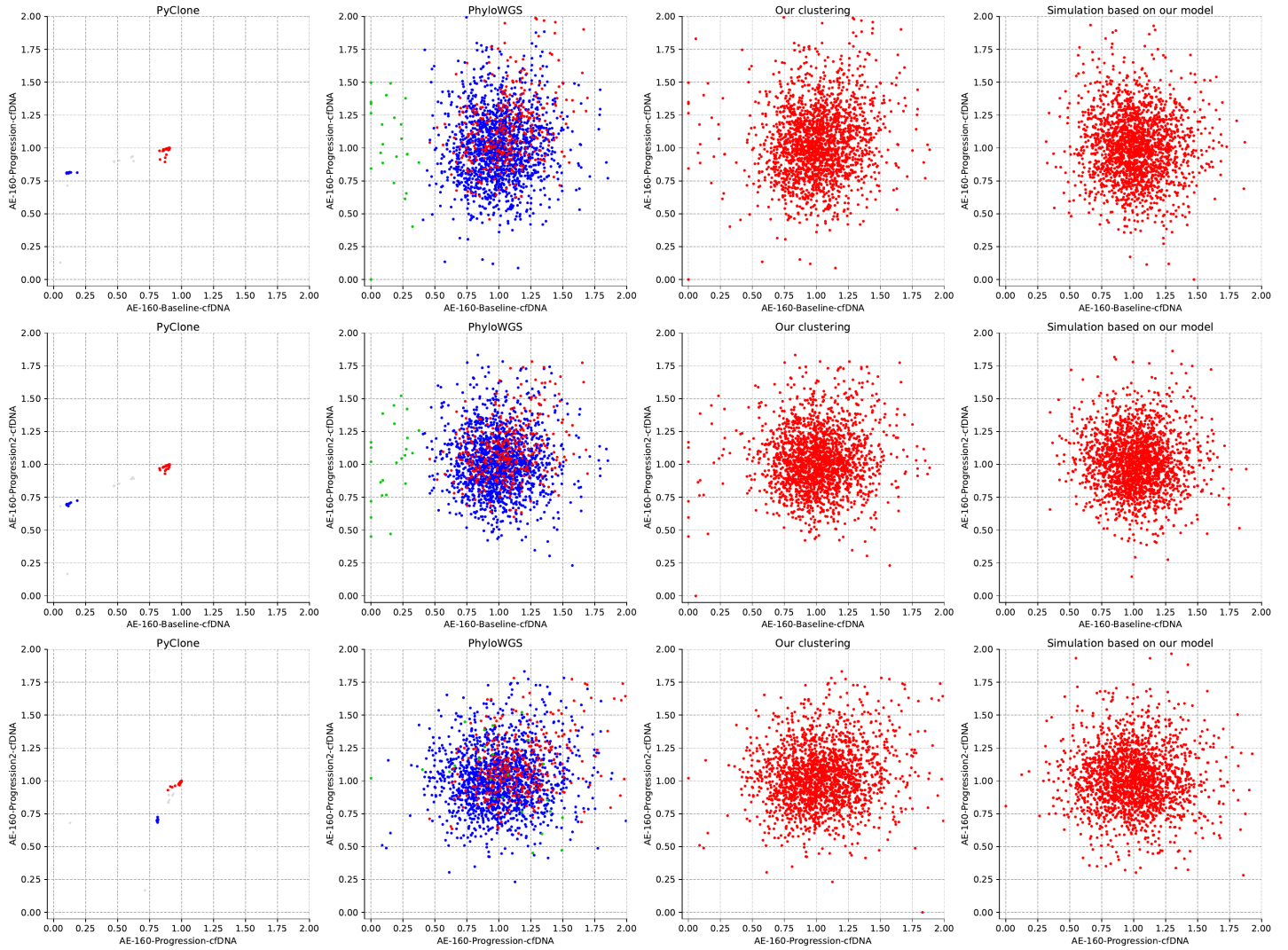
Patient AE-151



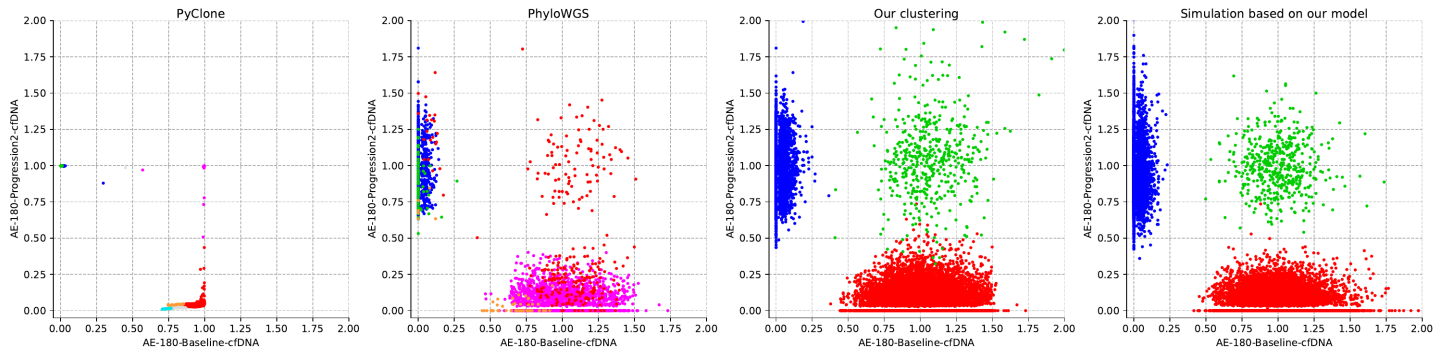
Patient AE-156



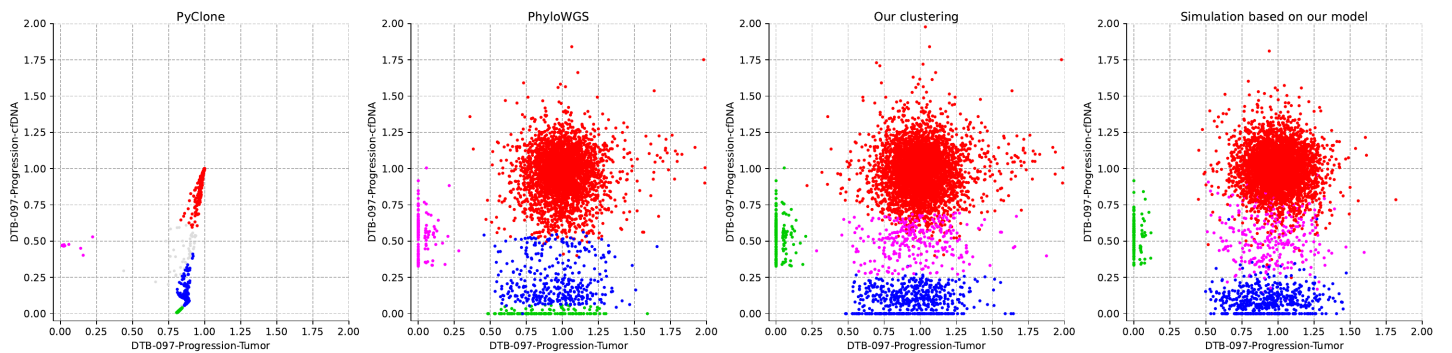
Patient AE-160



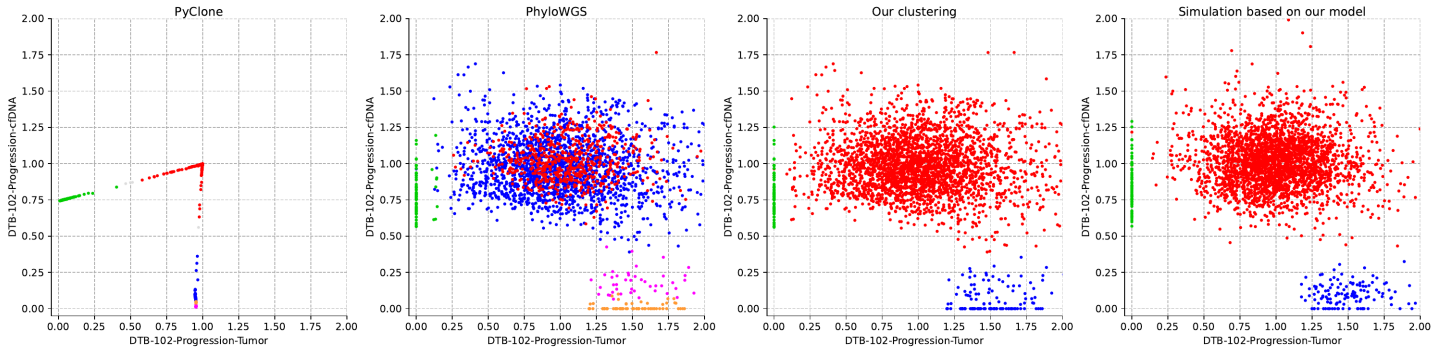
Patient AE-180



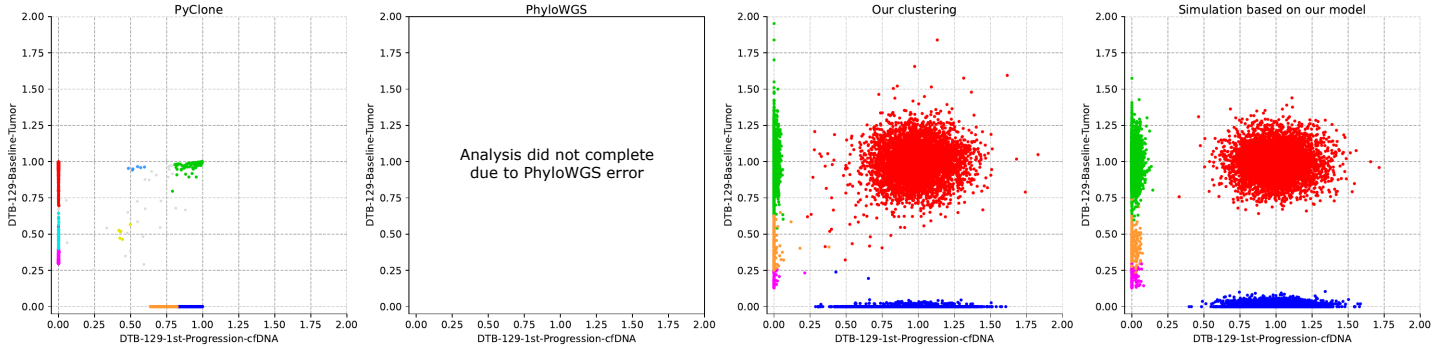
Patient DTB-097



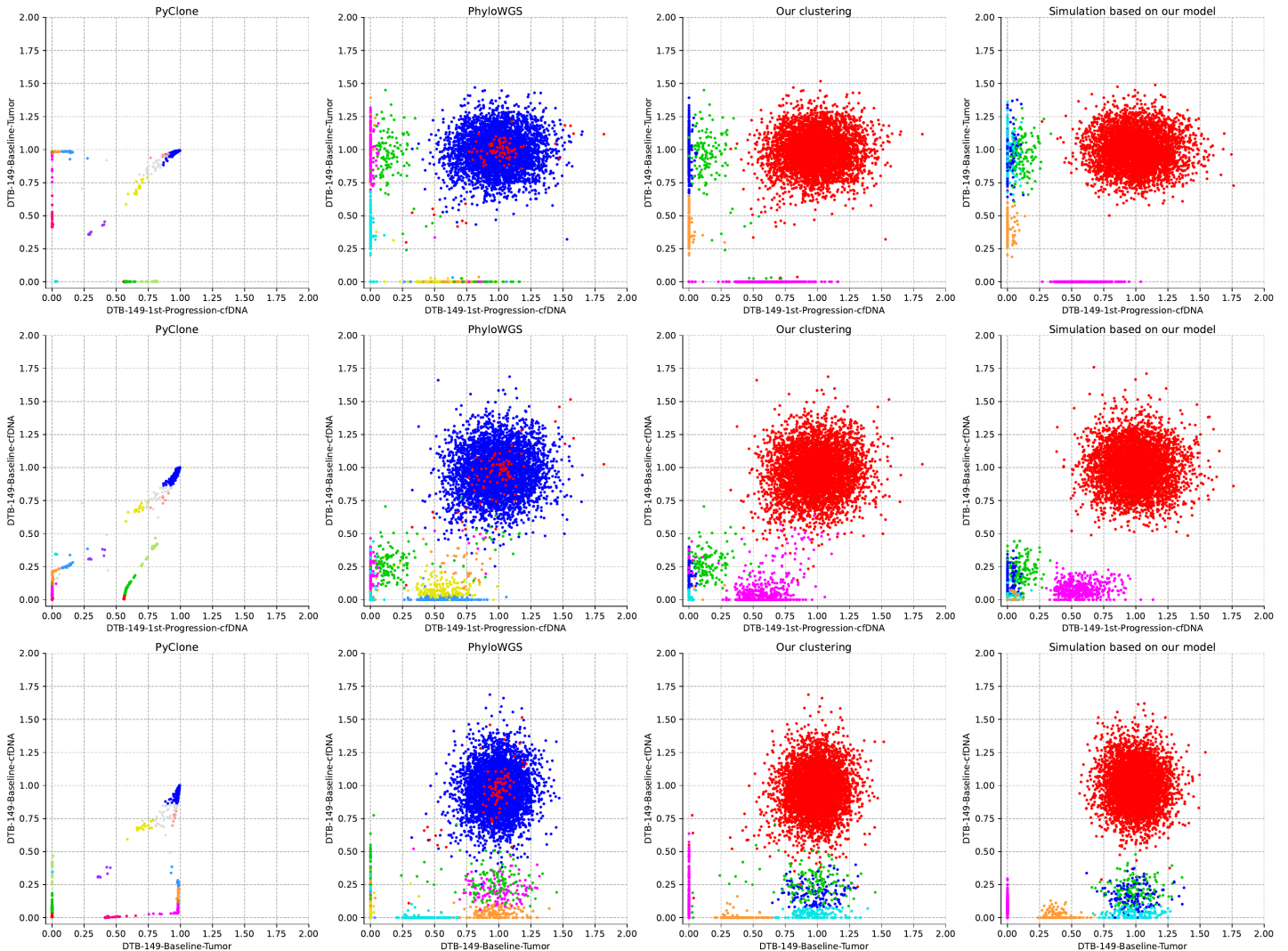
Patient DTB-102



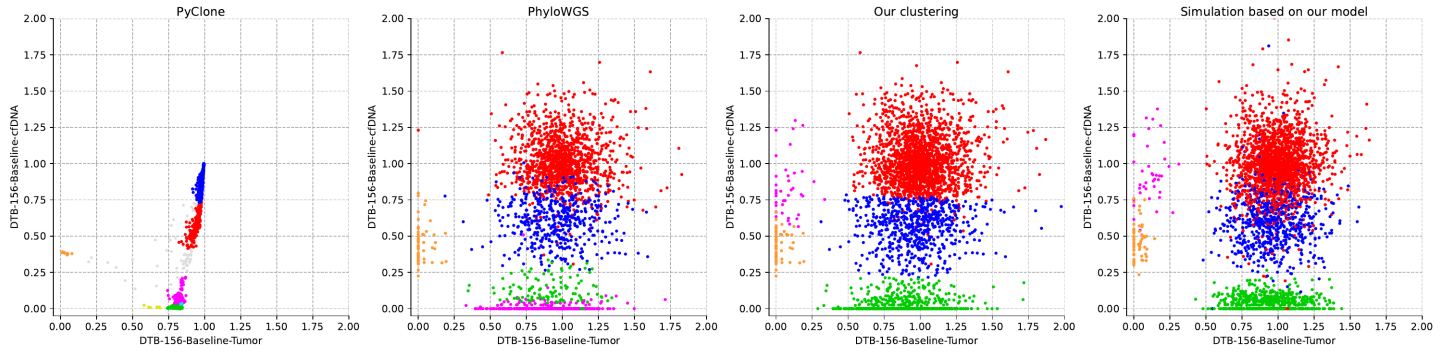
Patient DTB-129



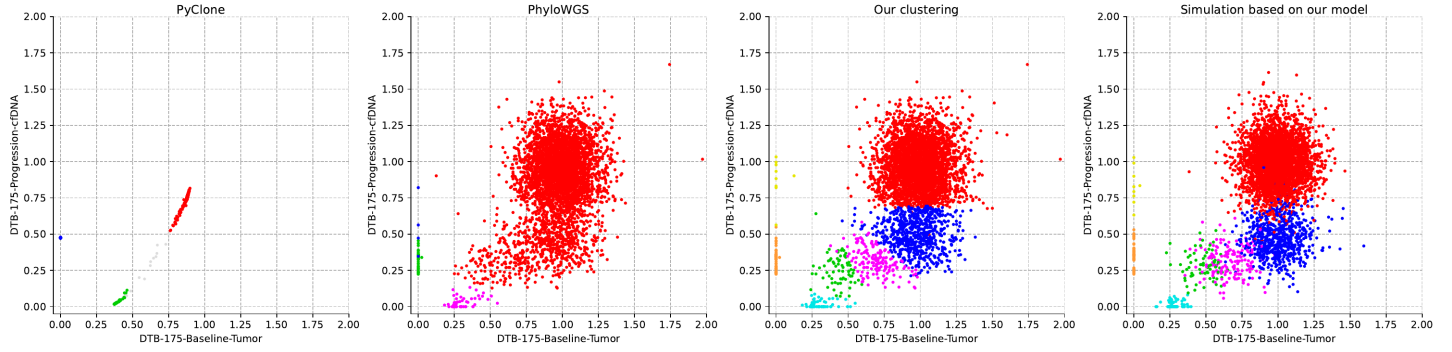
Patient DTB-149



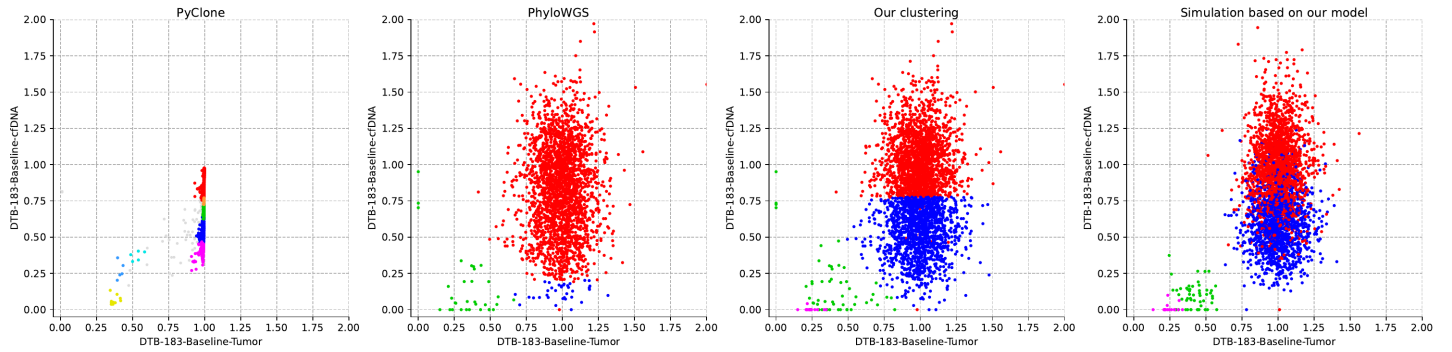
Patient DTB-156



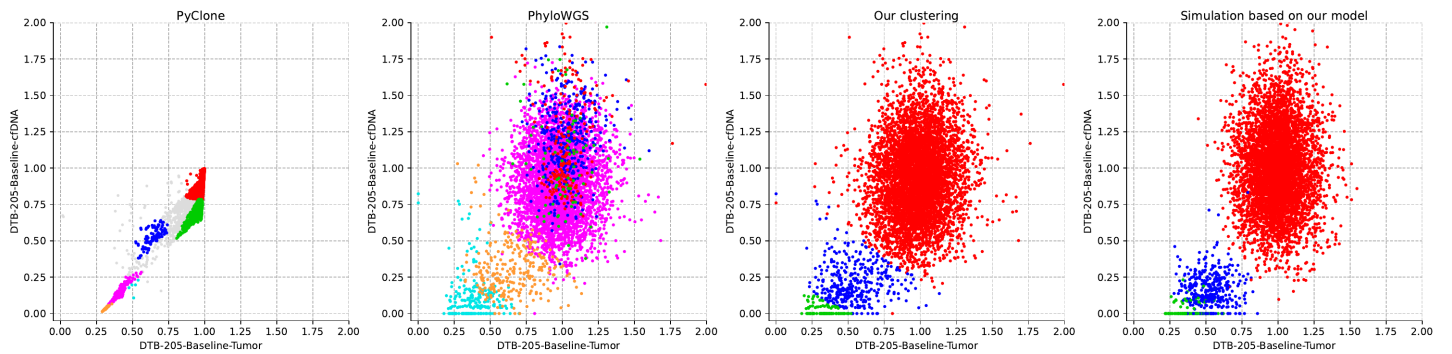
Patient DTB-175



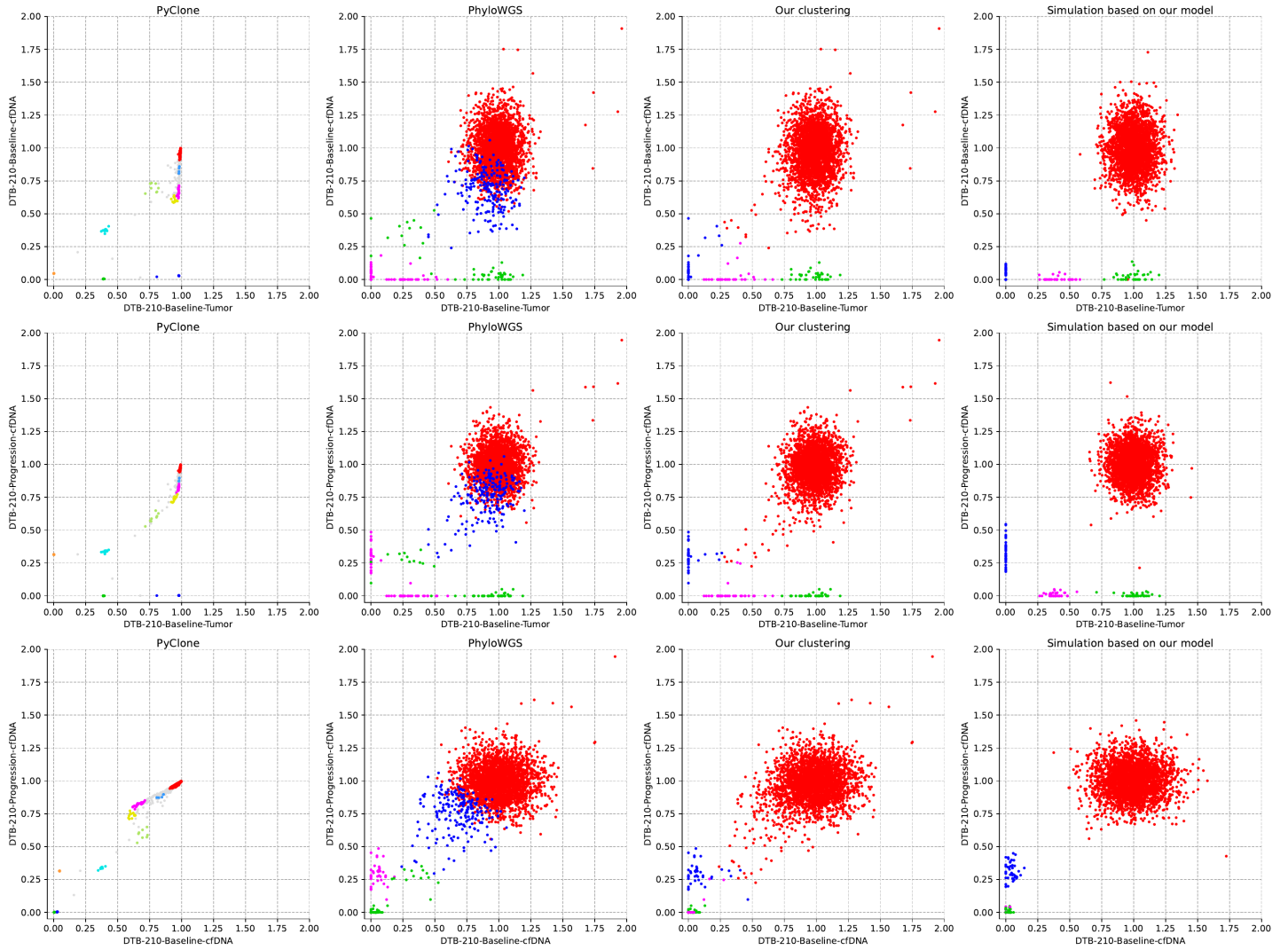
Patient DTB-183



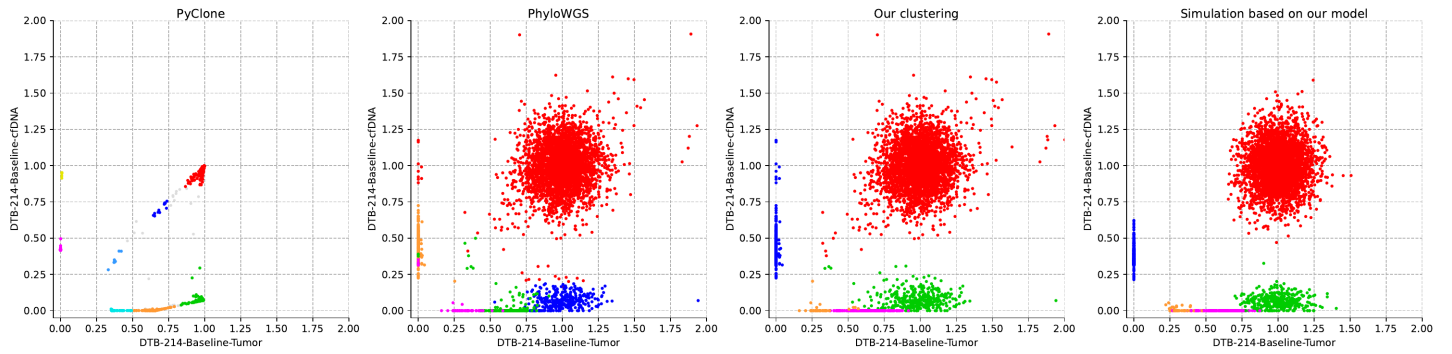
Patient DTB-205



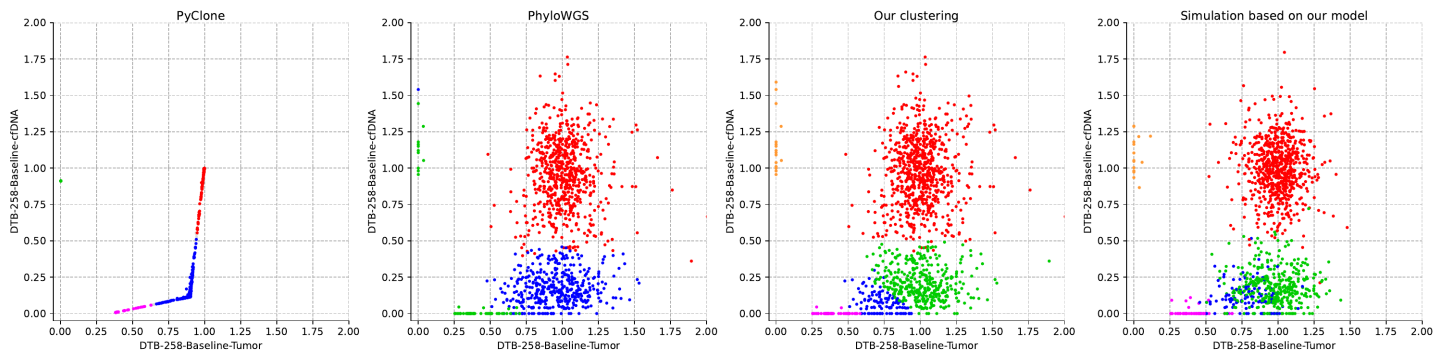
Patient DTB-210



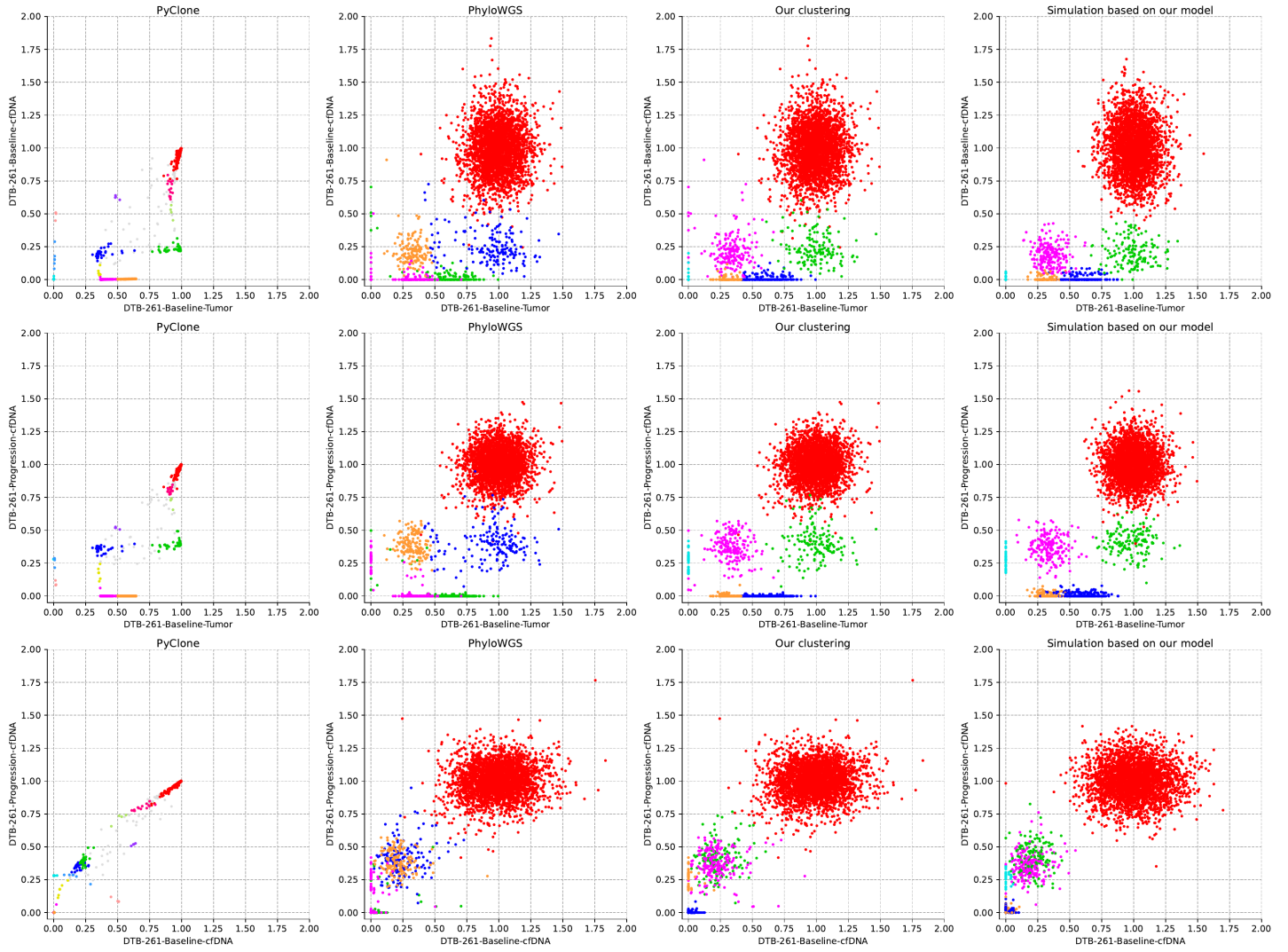
Patient DTB-214



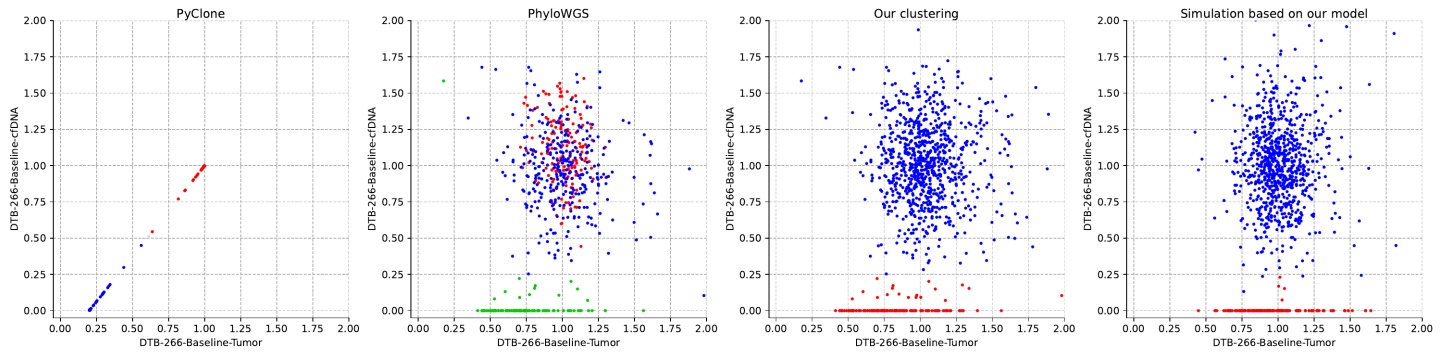
Patient DTB-258



Patient DTB-261



Patient DTB-266

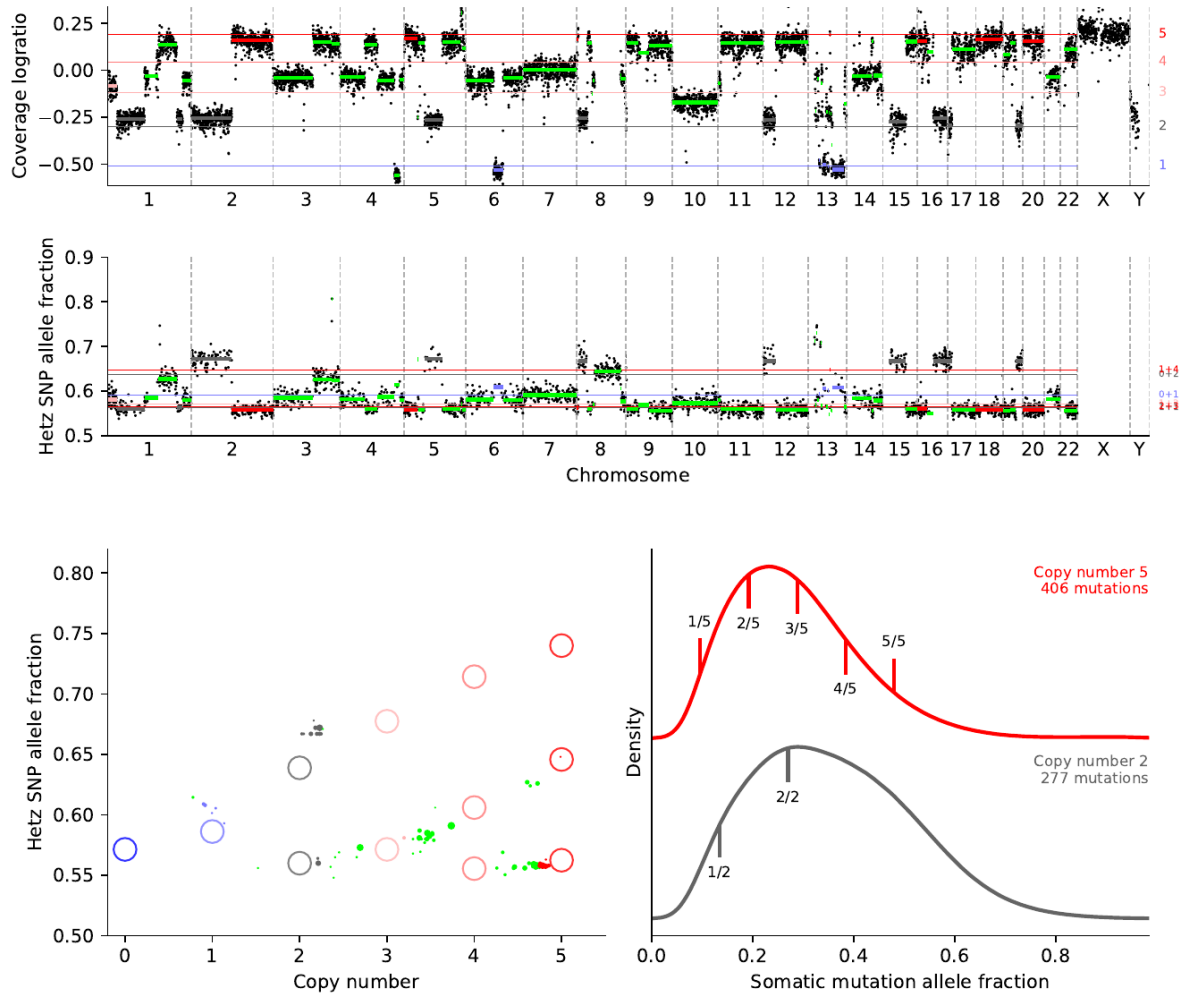


Battenberg copy number models

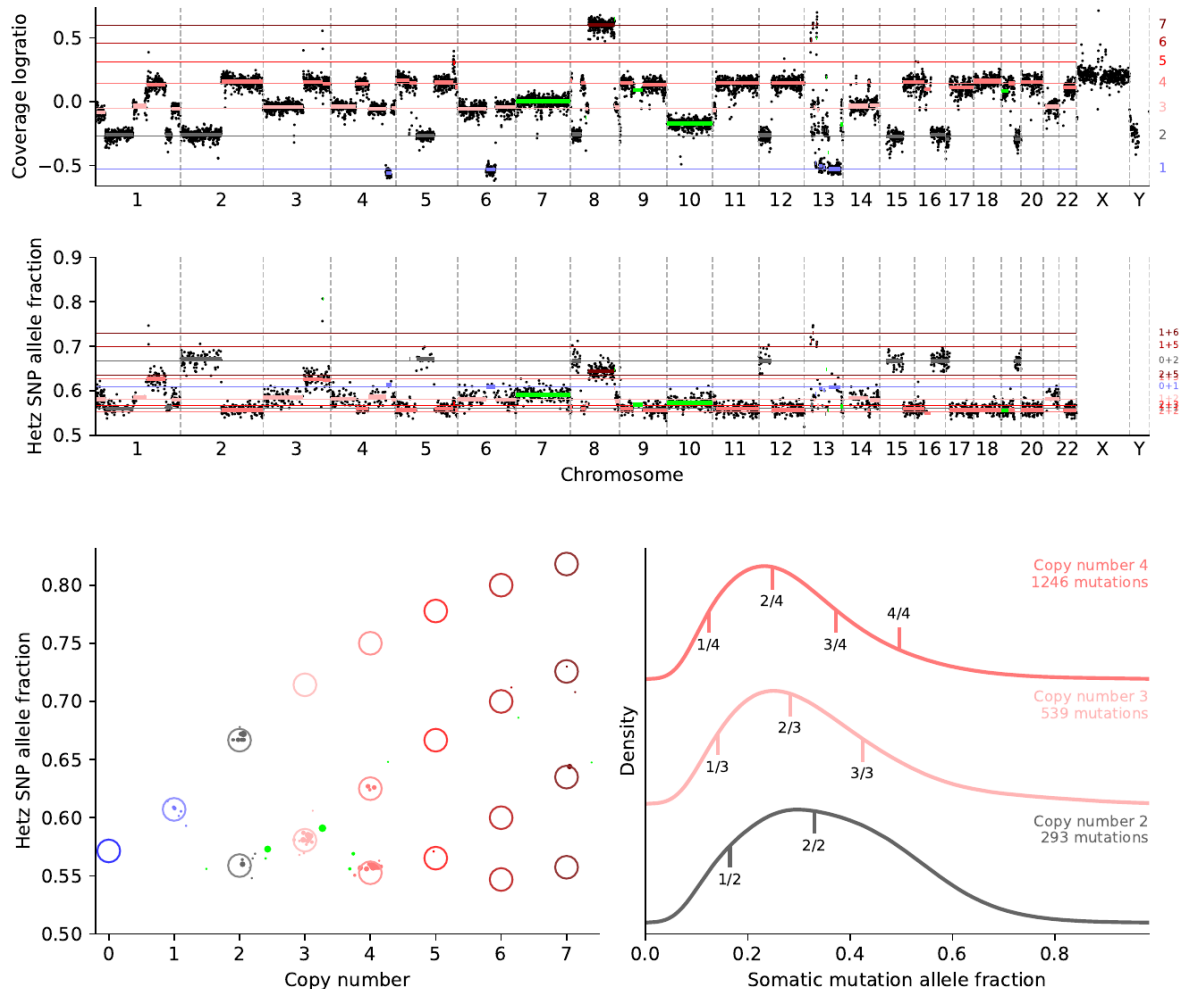
In this section we show side-by-side comparisons between Battenberg copy number models and our copy number models for the 17 cfDNA and tumor samples where Battenberg had inferred a different cancer fraction. On each page, the top half shows the Battenberg model, and the bottom half shows our model. The copy number models are shown using our visual representation, since it allows direct investigation of how well the observed coverage logratio and HSAF data align with their expected values under the copy number model.

AE-035-Baseline-cfDNA

Battenberg model - Cancer fraction: 27%, Average ploidy: 3.78, Diploid level: -0.3

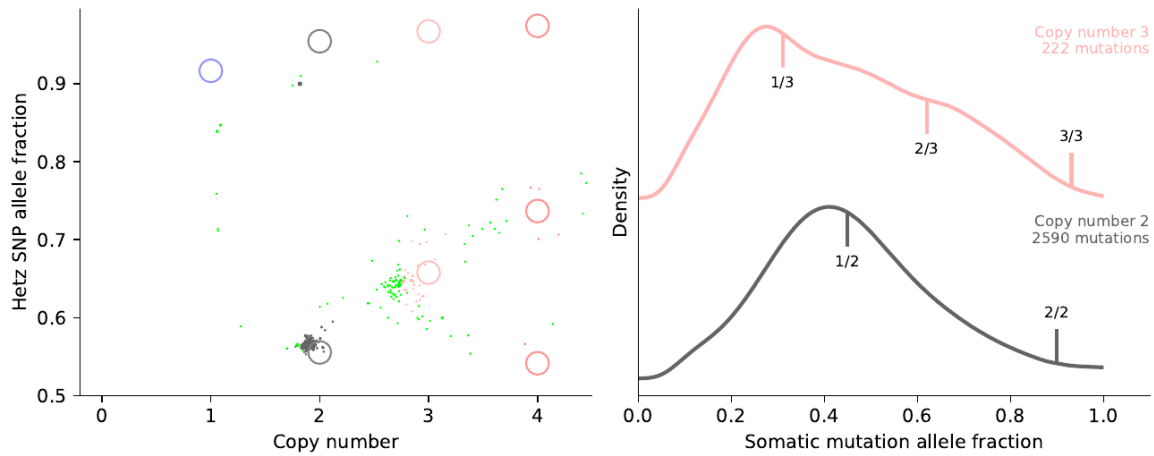
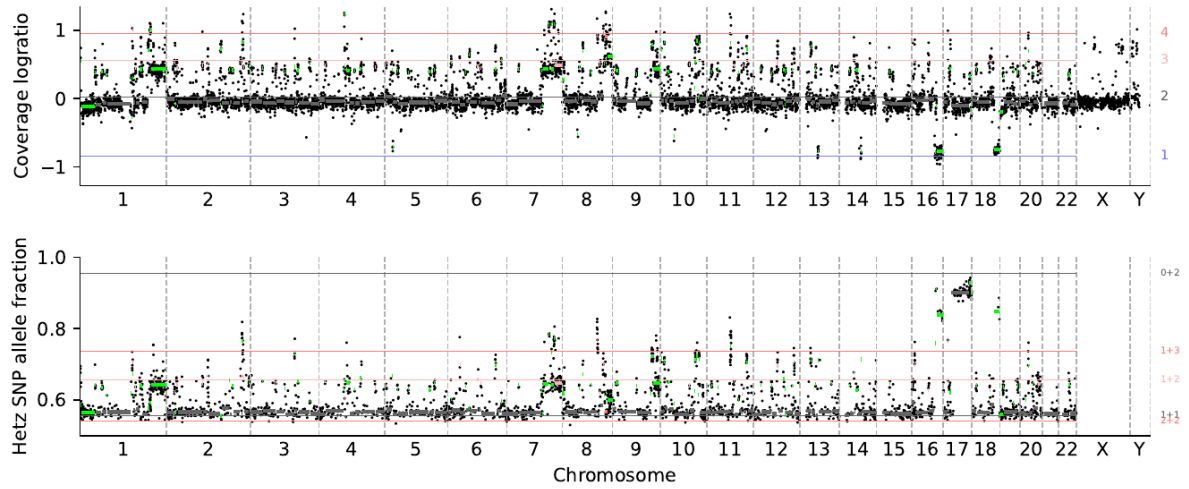


Our model - Cancer fraction: 33%, Average ploidy: 3.39, Diploid level: -0.27

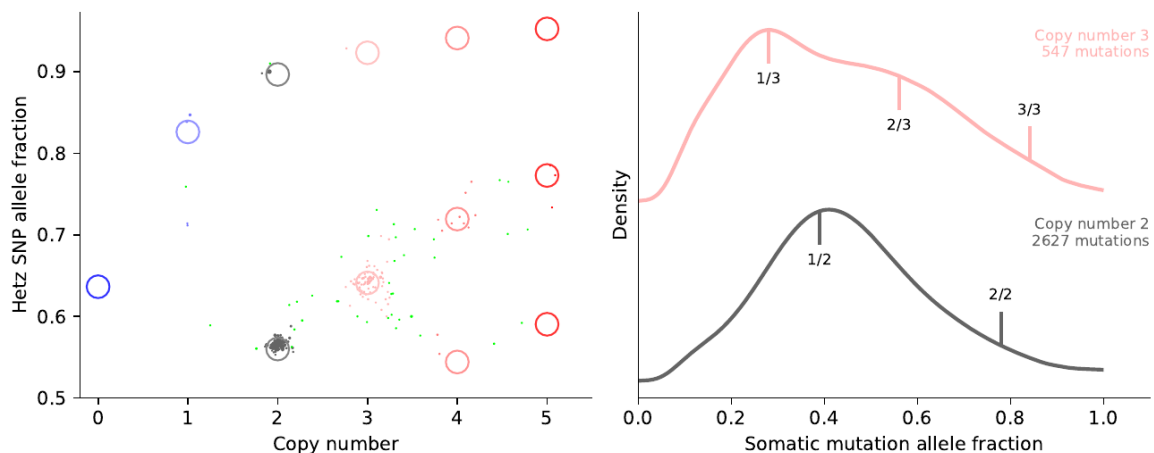
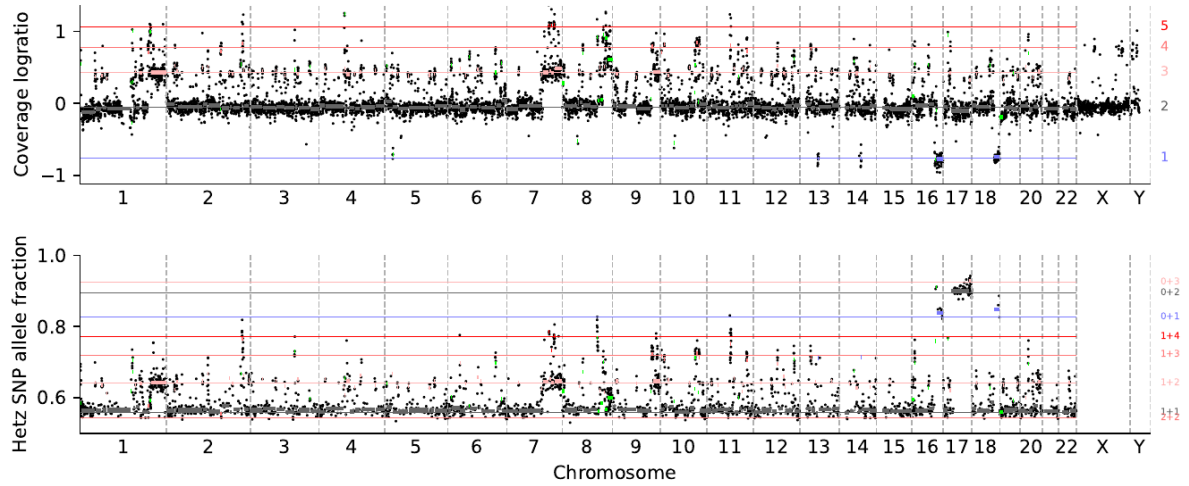


AE-092-Baseline-cfDNA

Battenberg model - Cancer fraction: 90%, Average ploidy: 2.05, Diploid level: 0.02

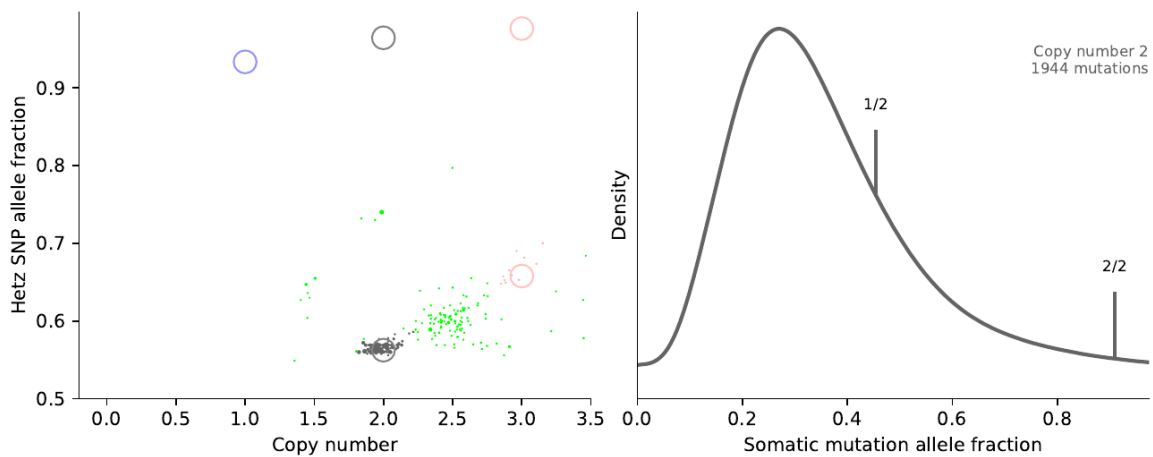
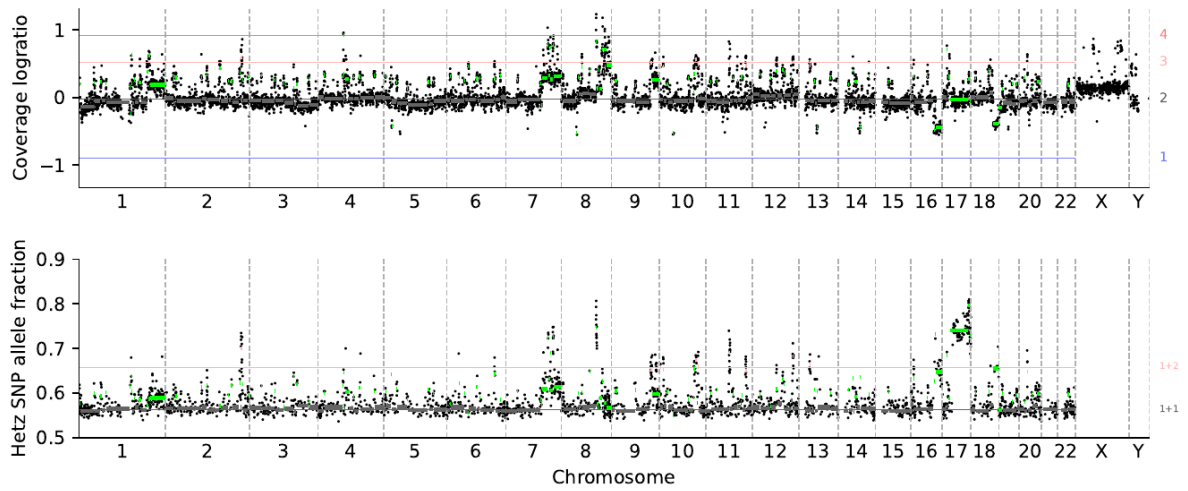


Our model - Cancer fraction: 78%, Average ploidy: 2.25, Diploid level: -0.05

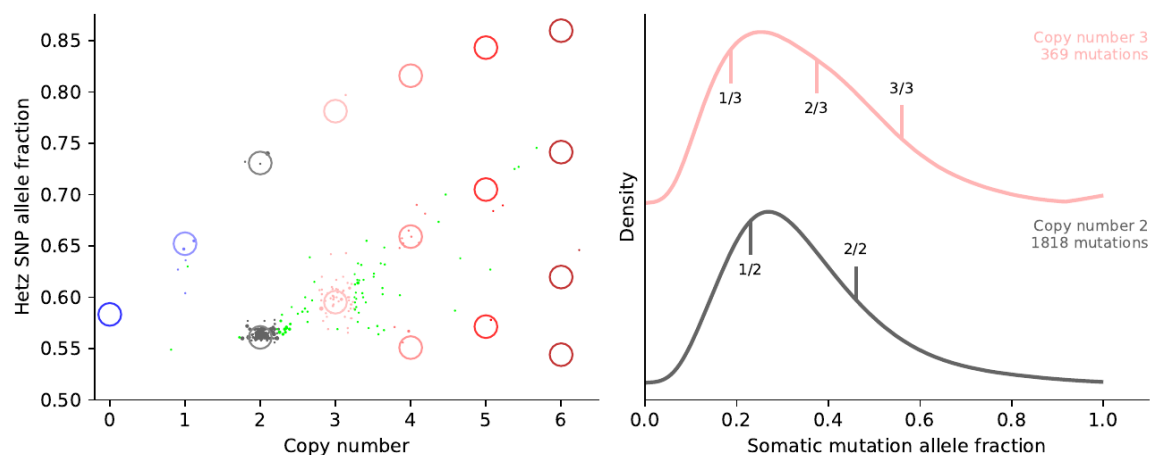
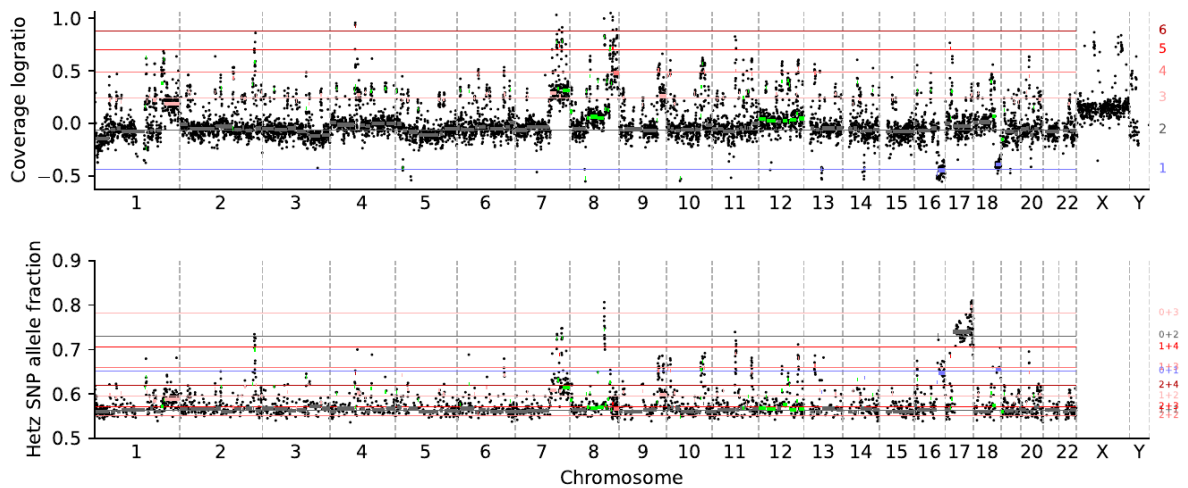


AE-092-Progression-cfDNA

Battenberg model - Cancer fraction: 91%, Average ploidy: 2.06, Diploid level: -0.02

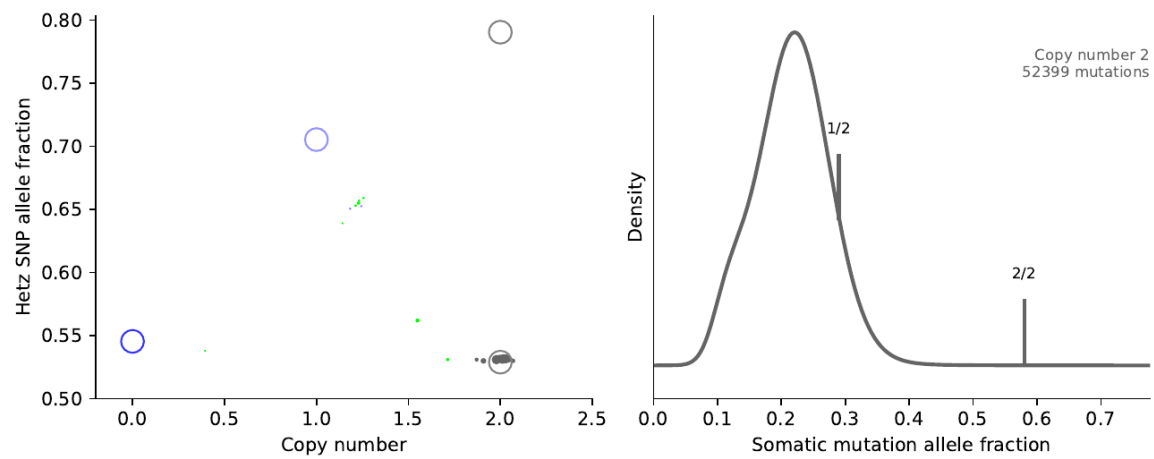
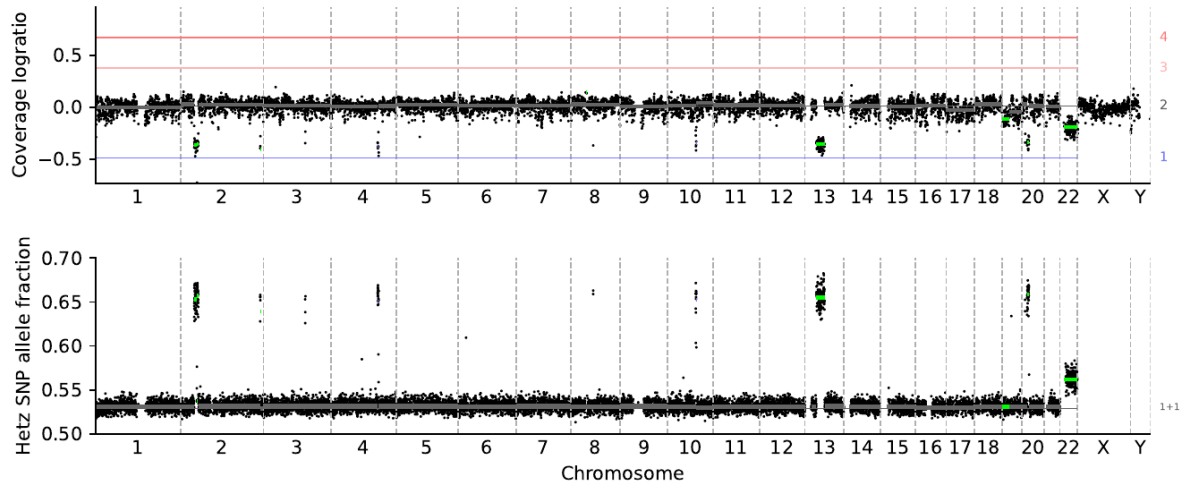


Our model - Cancer fraction: 46%, Average ploidy: 2.35, Diploid level: -0.06

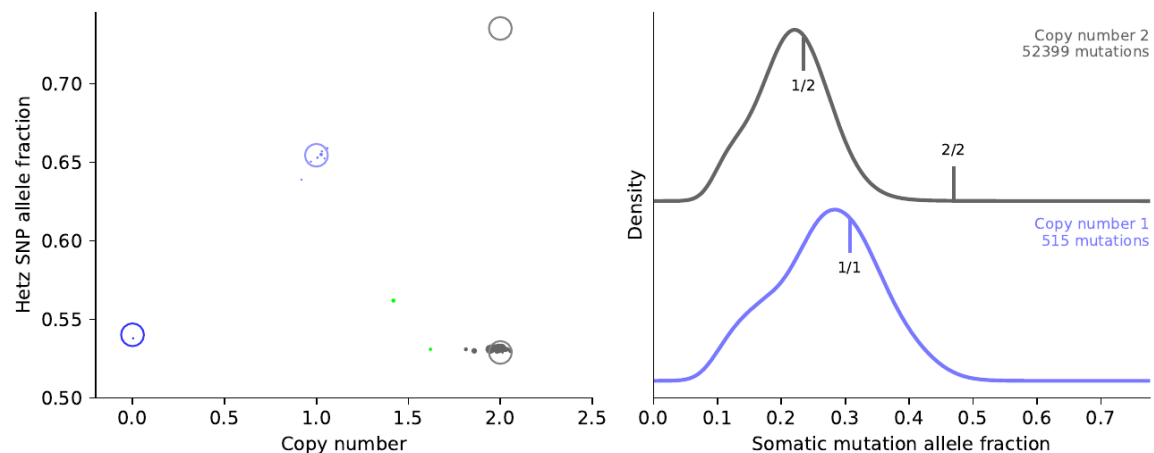
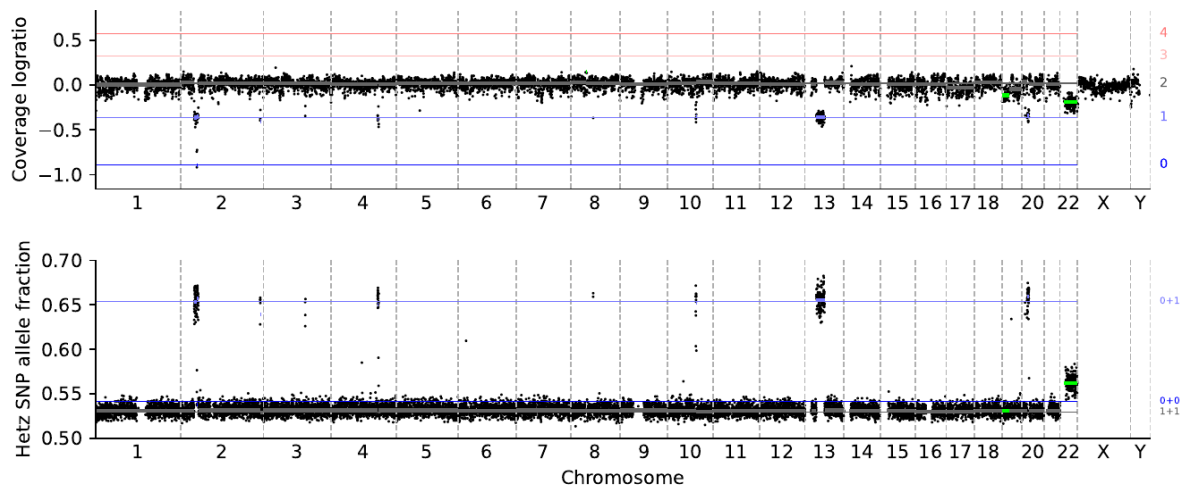


AE-098-Baseline-cfDNA

Battenberg model - Cancer fraction: 58%, Average ploidy: 1.98, Diploid level: 0.01

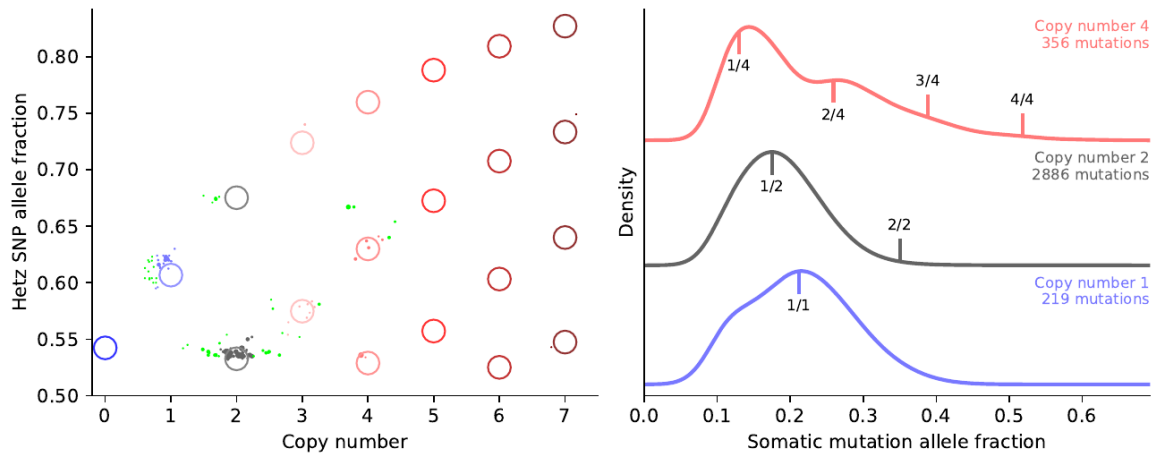
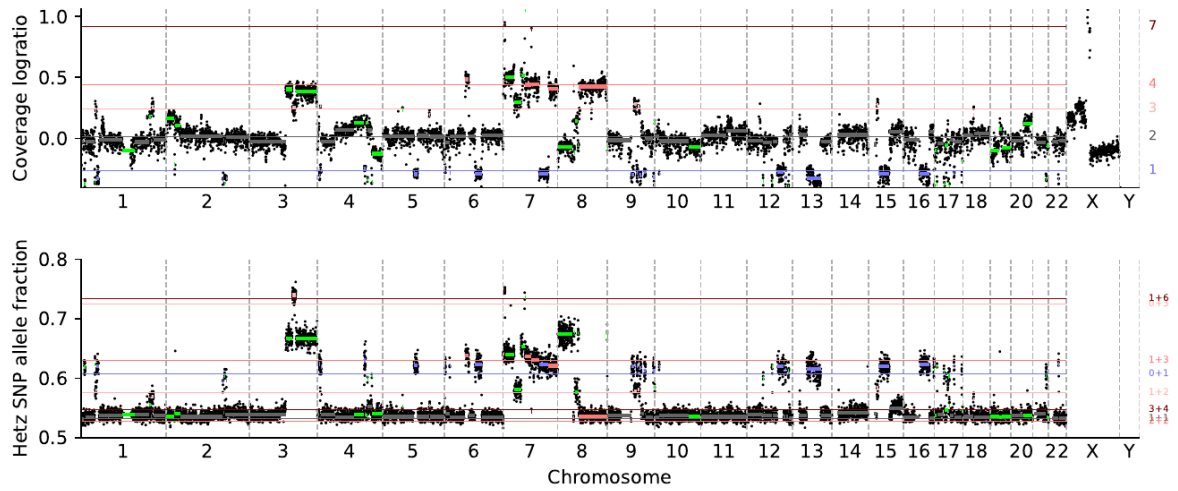


Our model - Cancer fraction: 47%, Average ploidy: 1.95, Diploid level: 0.02

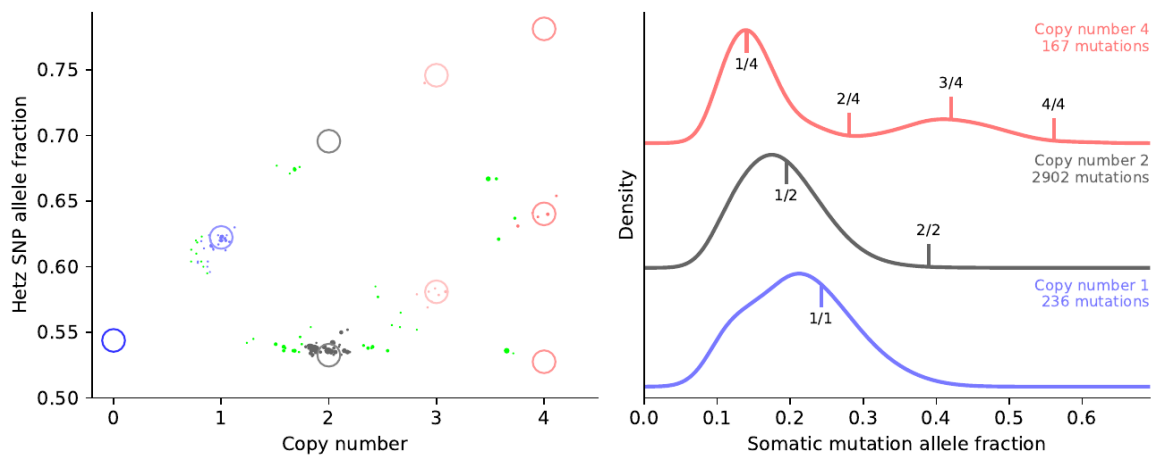
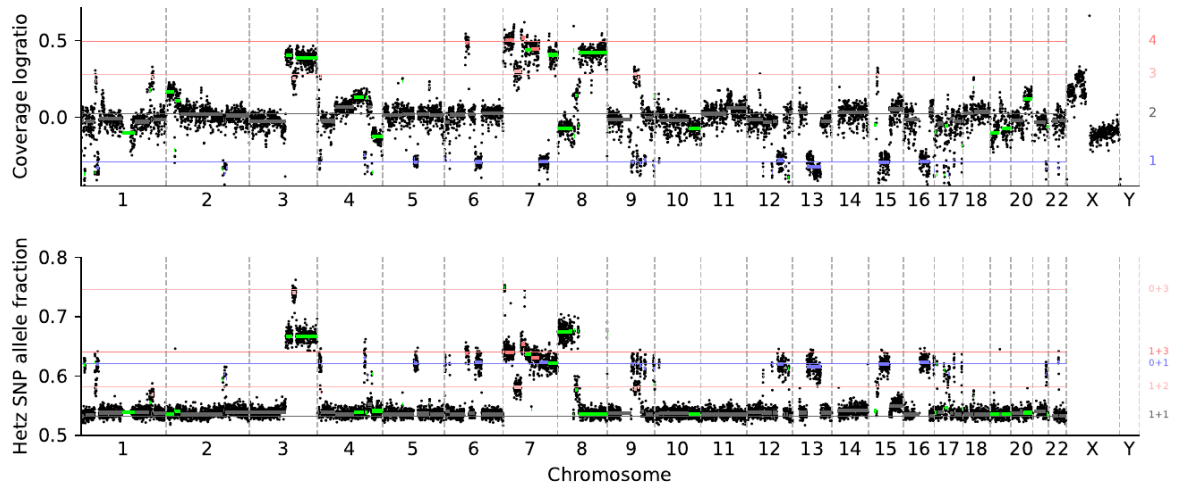


AE-110-Baseline-cfDNA

Battenberg model - Cancer fraction: 35%, Average ploidy: 2.1, Diploid level: 0.01

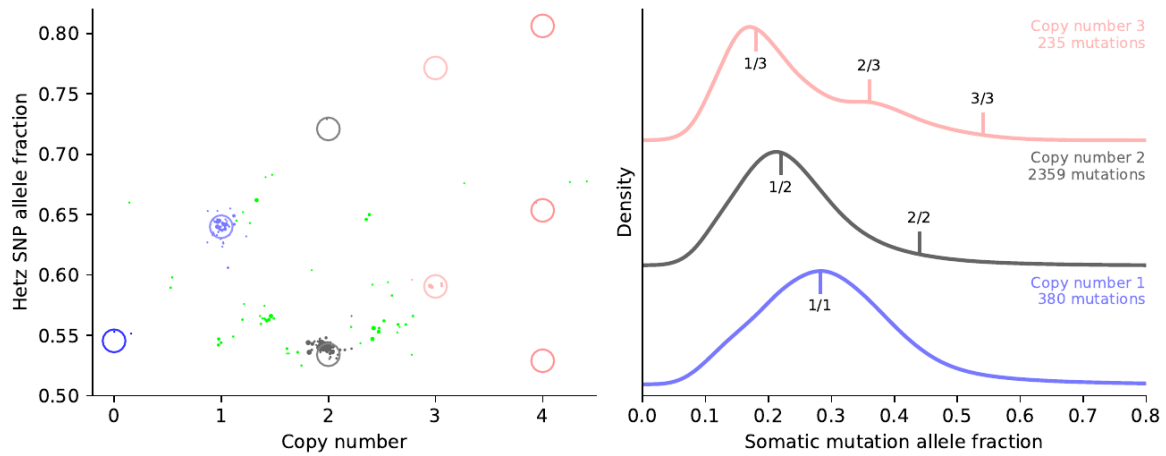
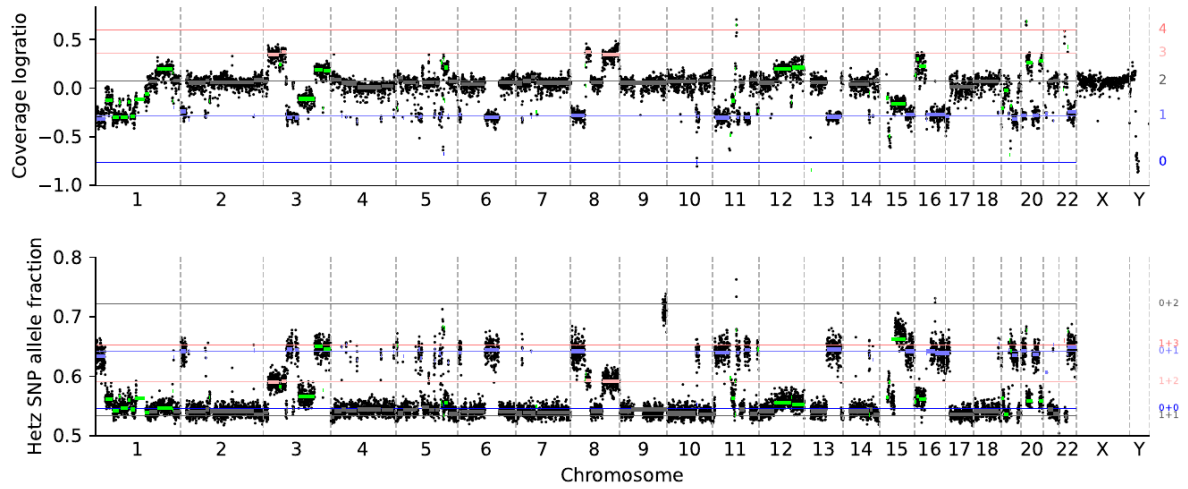


Our model - Cancer fraction: 39%, Average ploidy: 2.03, Diploid level: 0.02

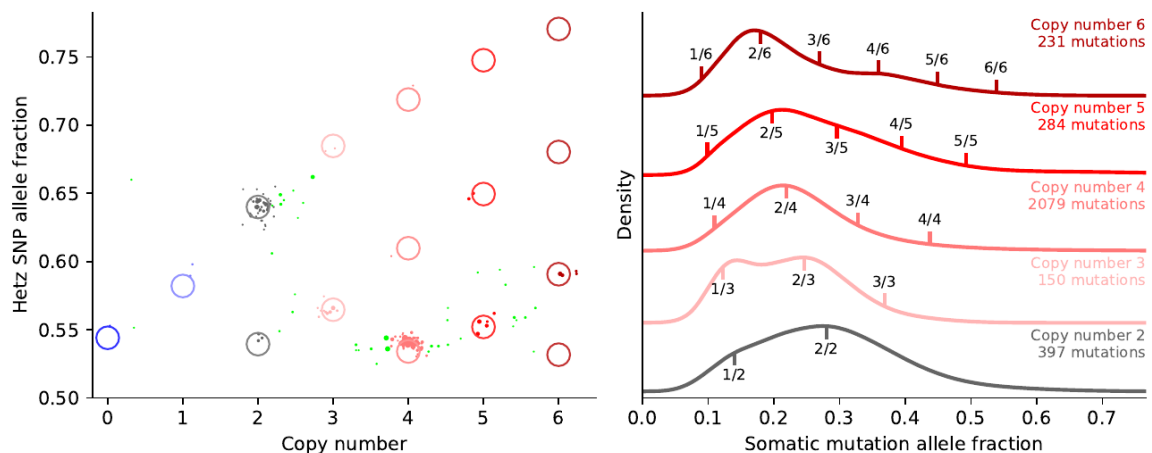
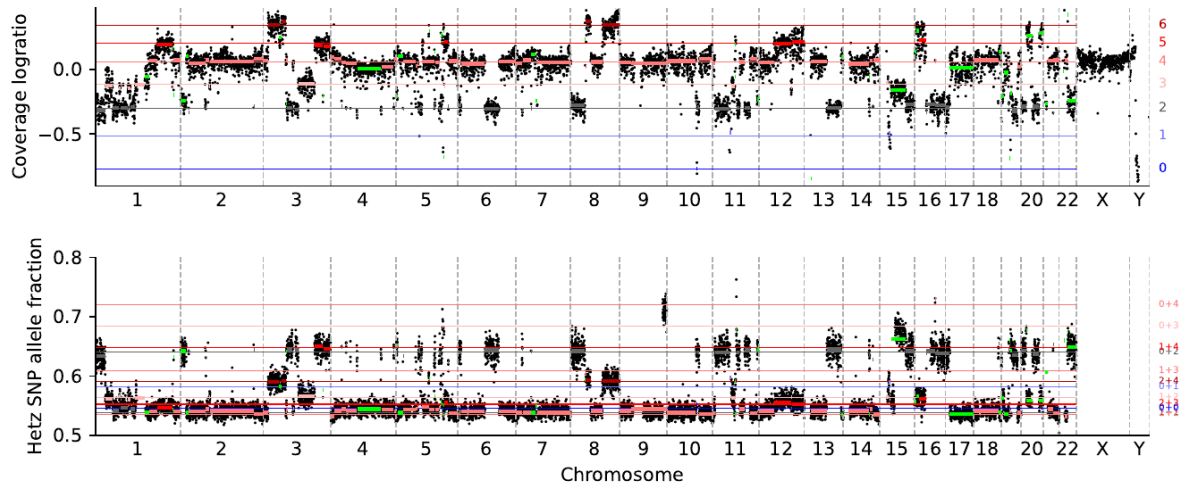


AE-129-Baseline-cfDNA

Battenberg model - Cancer fraction: 44%, Average ploidy: 1.81, Diploid level: 0.07

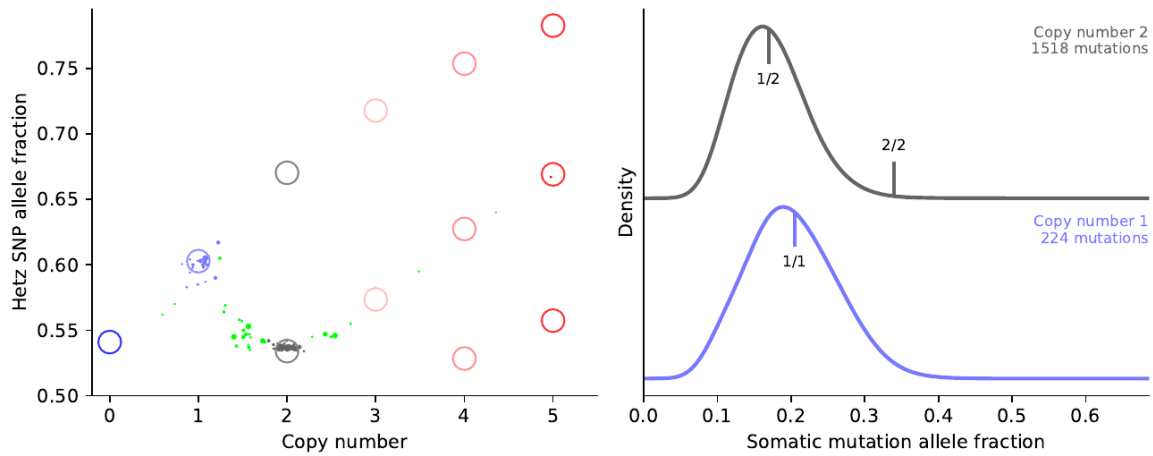
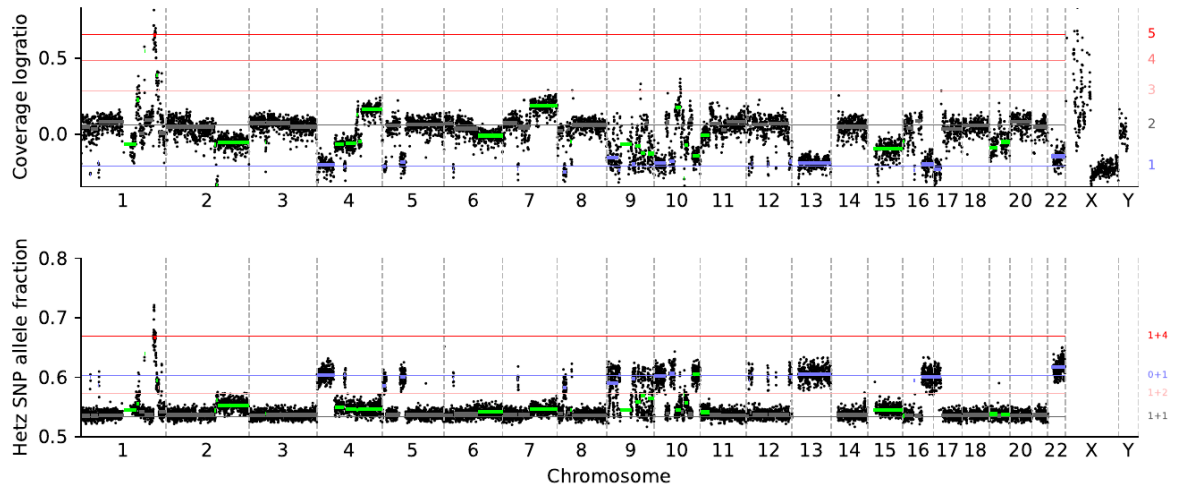


Our model - Cancer fraction: 28%, Average ploidy: 3.71, Diploid level: -0.3

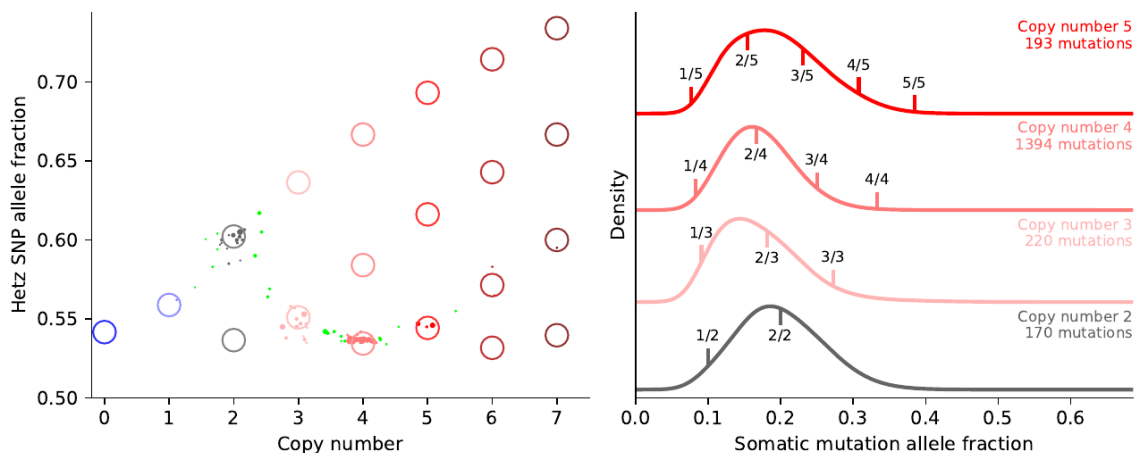
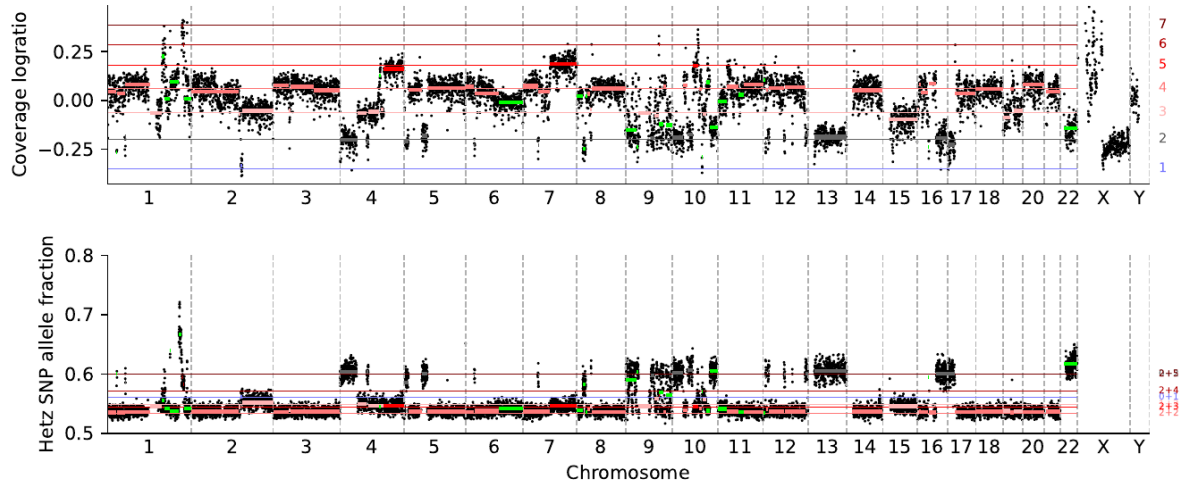


AE-132-Baseline-cfDNA

Battenberg model - Cancer fraction: 34%, Average ploidy: 1.8, Diploid level: 0.06

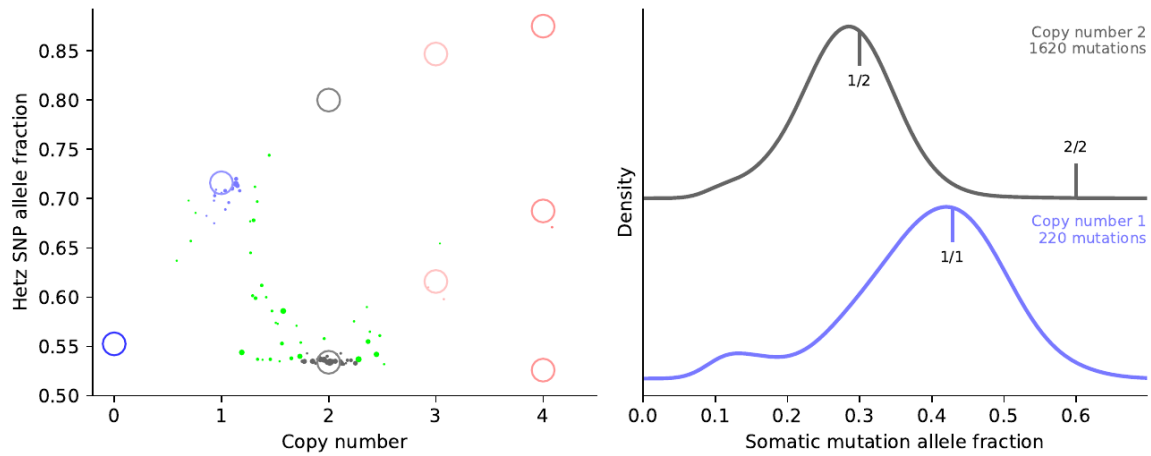
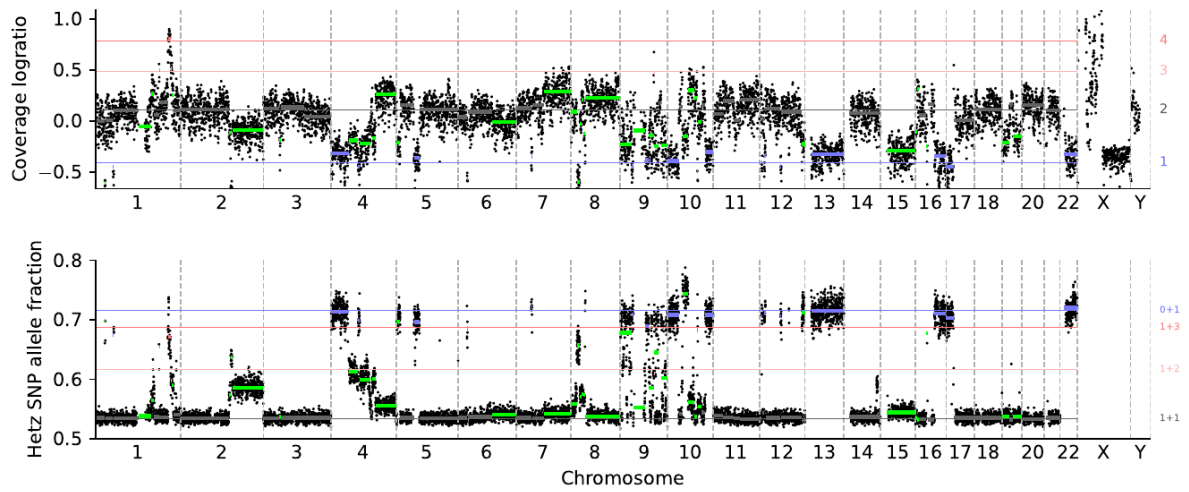


Our model - Cancer fraction: 20%, Average ploidy: 3.7, Diploid level: -0.2

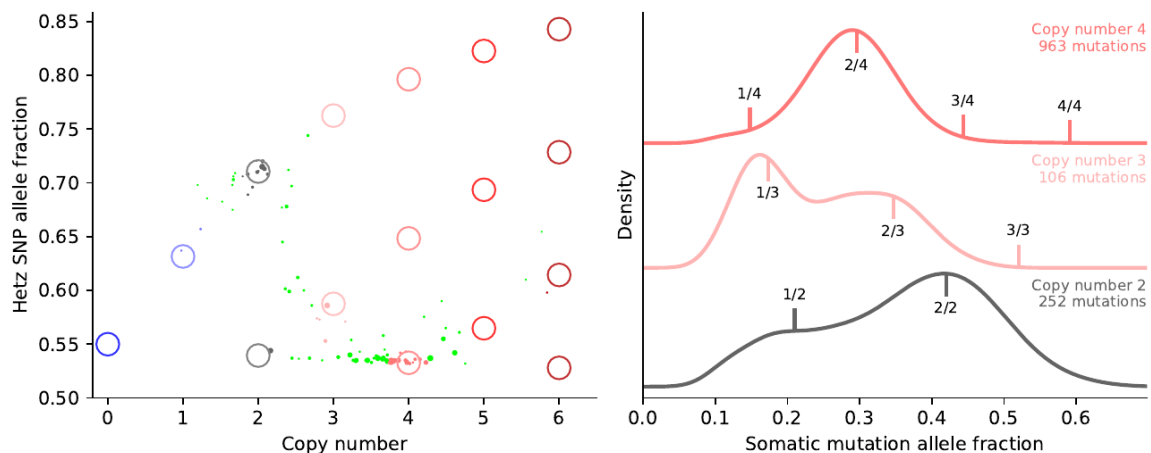
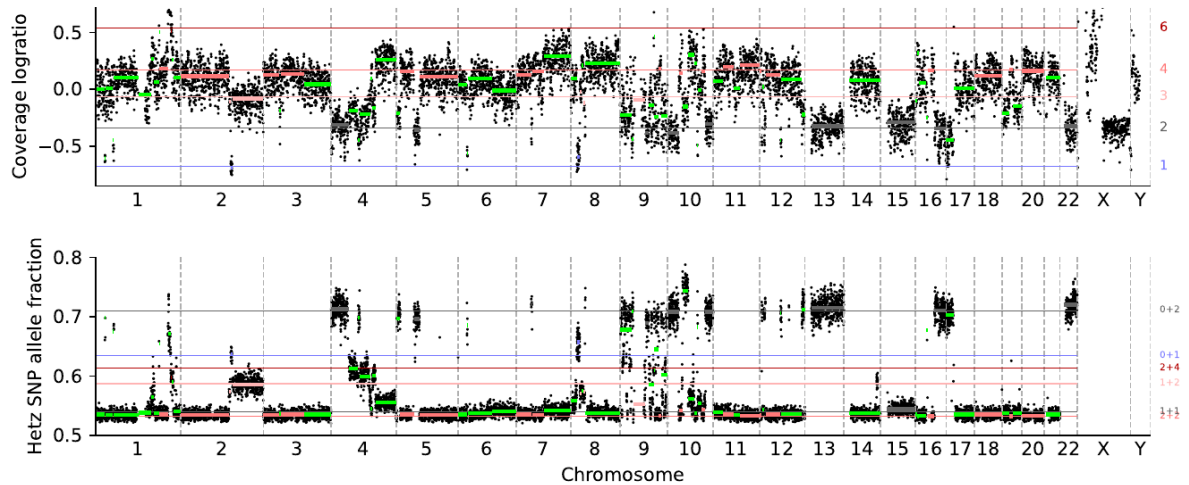


AE-132-Progression2-cfDNA

Battenberg model - Cancer fraction: 60%, Average ploidy: 1.81, Diploid level: 0.11

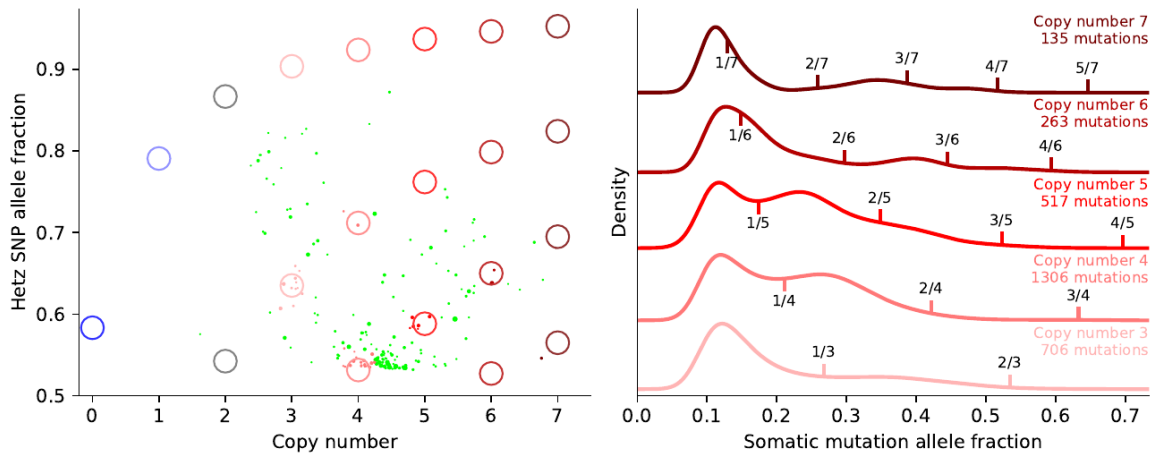
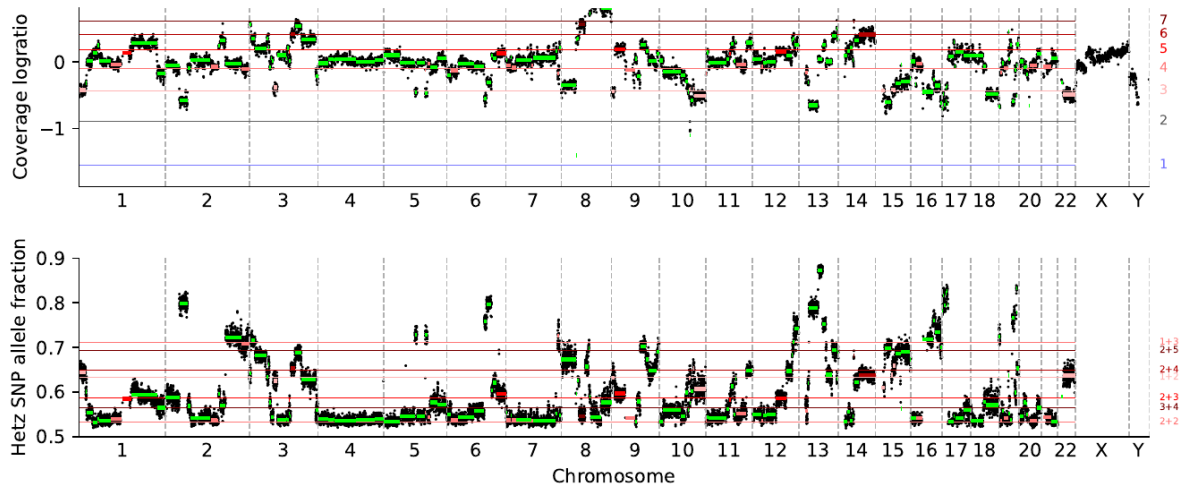


Our model - Cancer fraction: 42%, Average ploidy: 3.52, Diploid level: -0.34

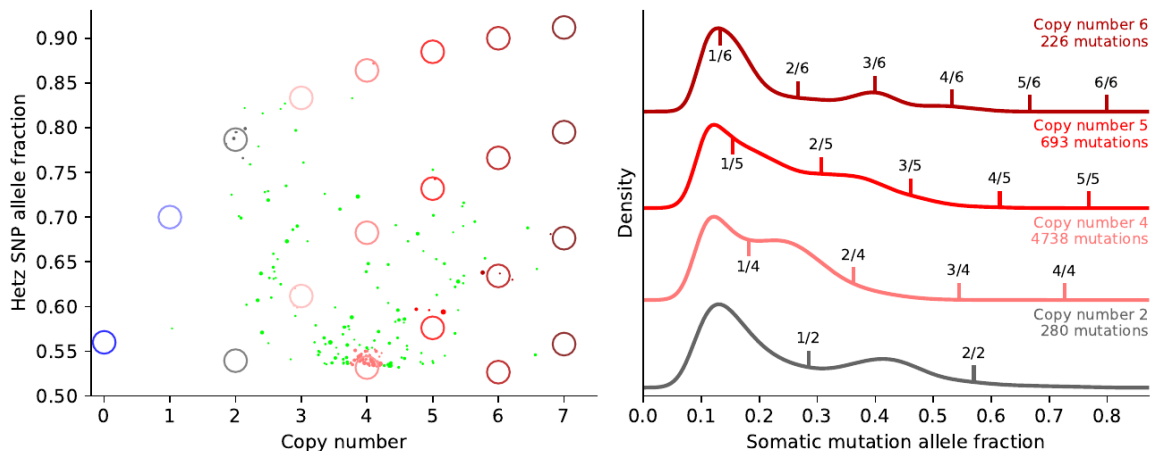
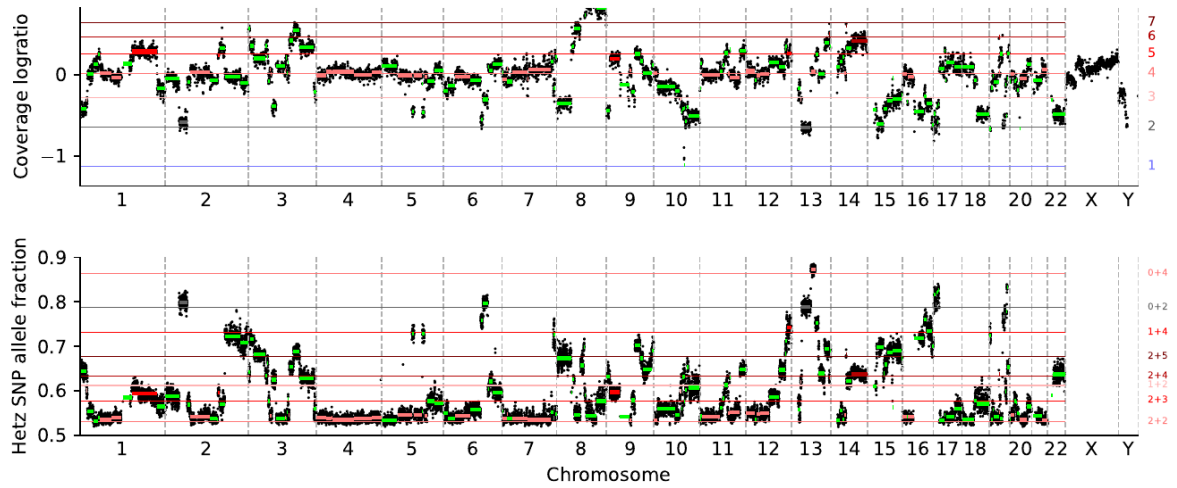


AE-180-Progression-cfDNA

Battenberg model - Cancer fraction: 73%, Average ploidy: 4.44, Diploid level: -0.89

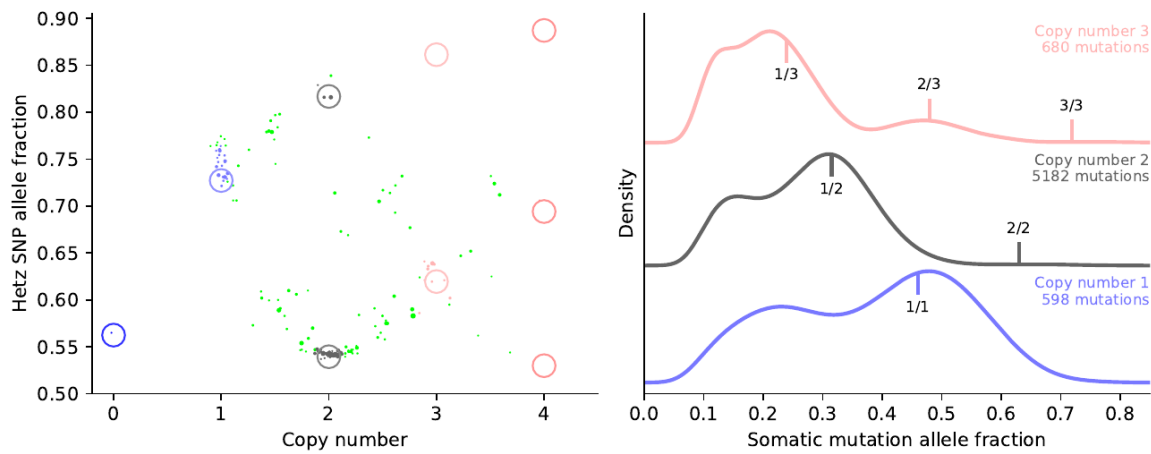
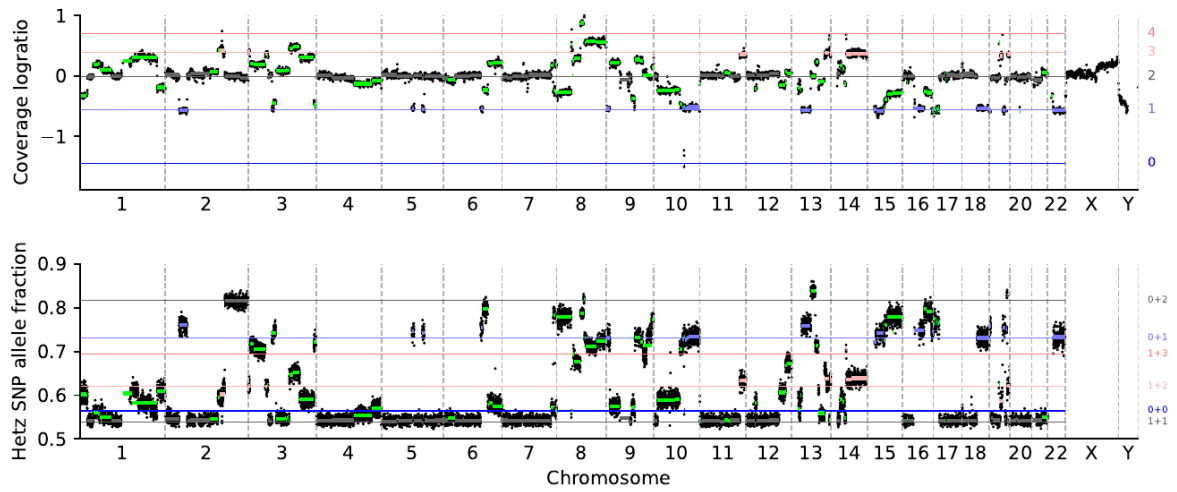


Our model - Cancer fraction: 57%, Average ploidy: 4.05, Diploid level: -0.64

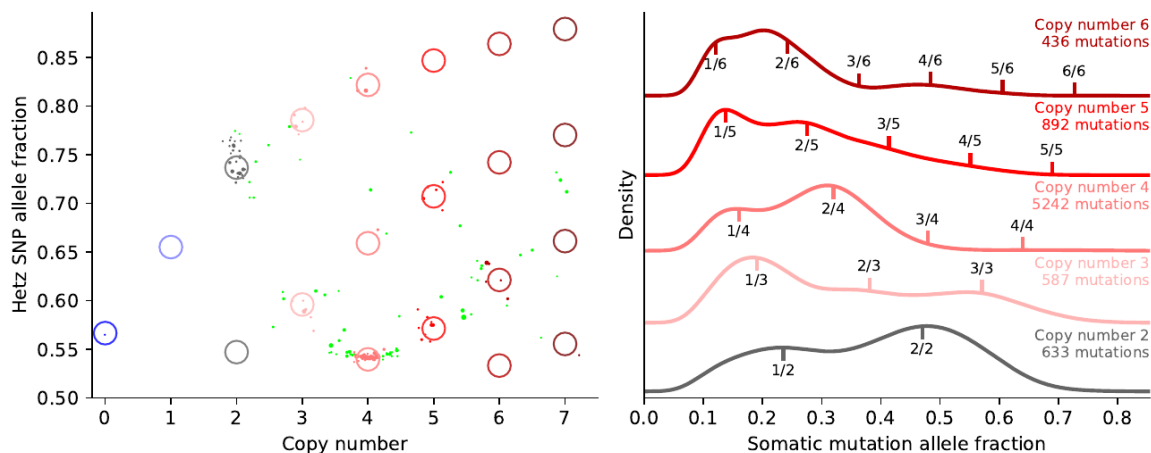
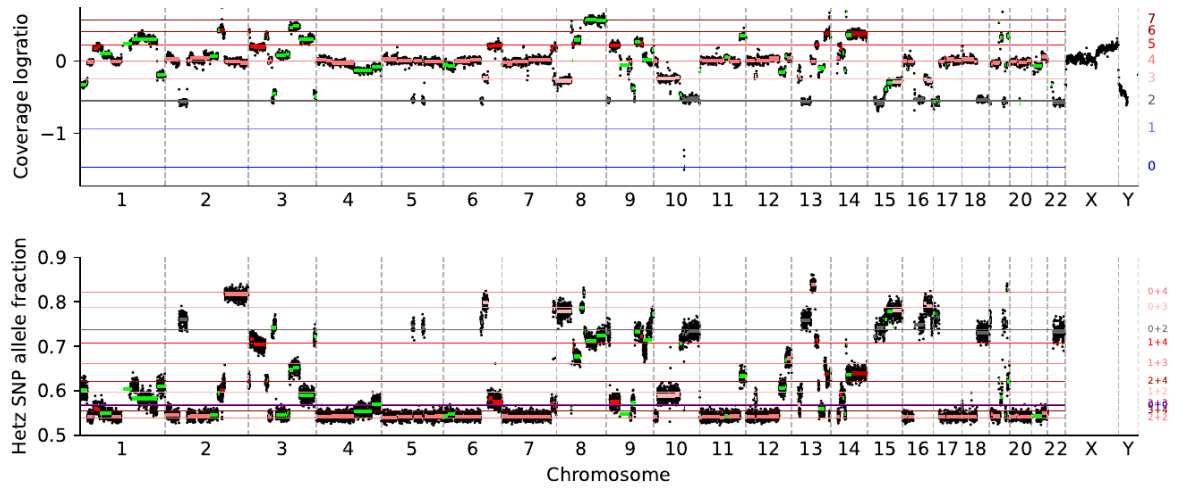


AE-180-Progression2-cfDNA

Battenberg model - Cancer fraction: 63%, Average ploidy: 2.04, Diploid level: -0.01

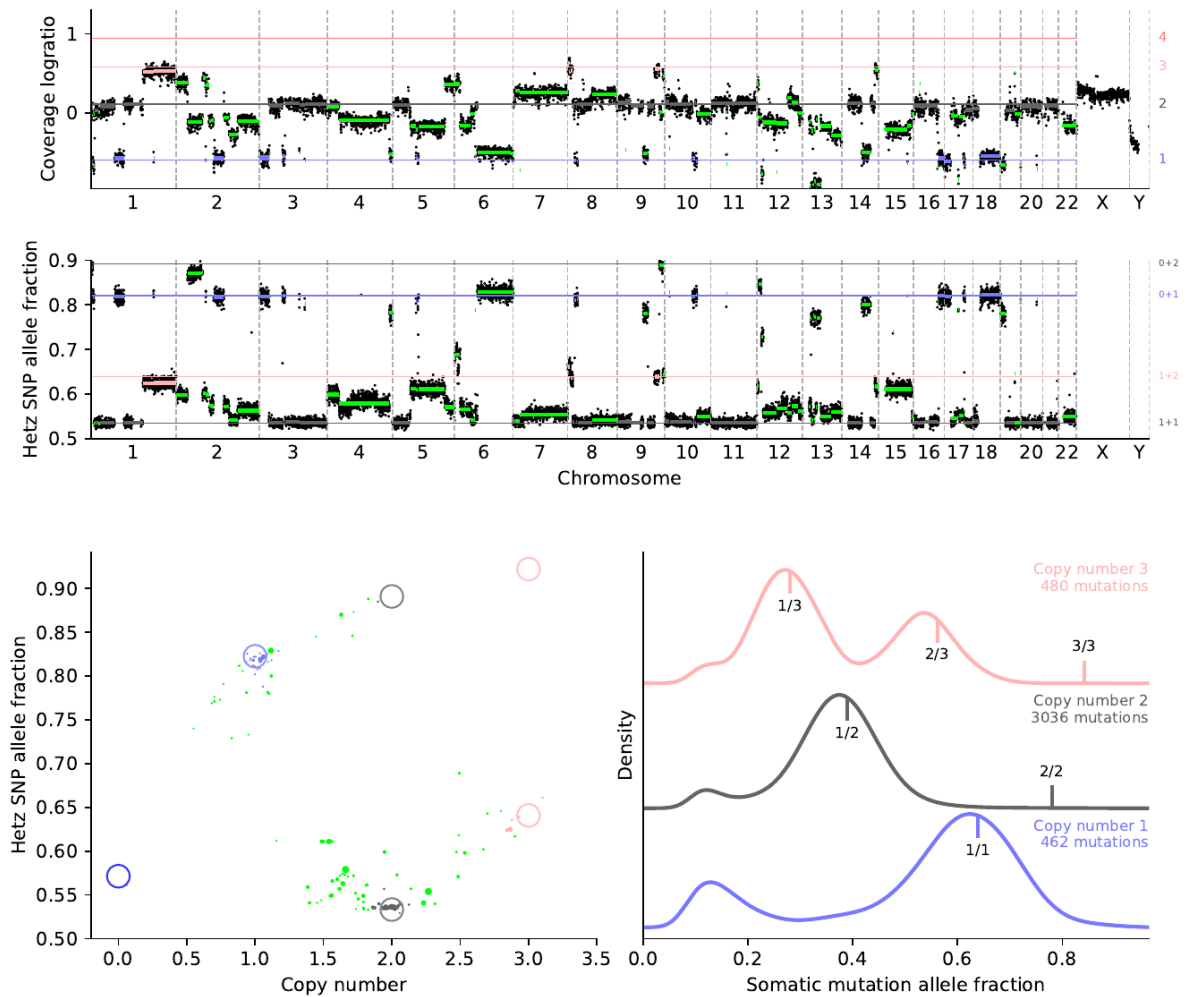


Our model - Cancer fraction: 47%, Average ploidy: 3.99, Diploid level: -0.55

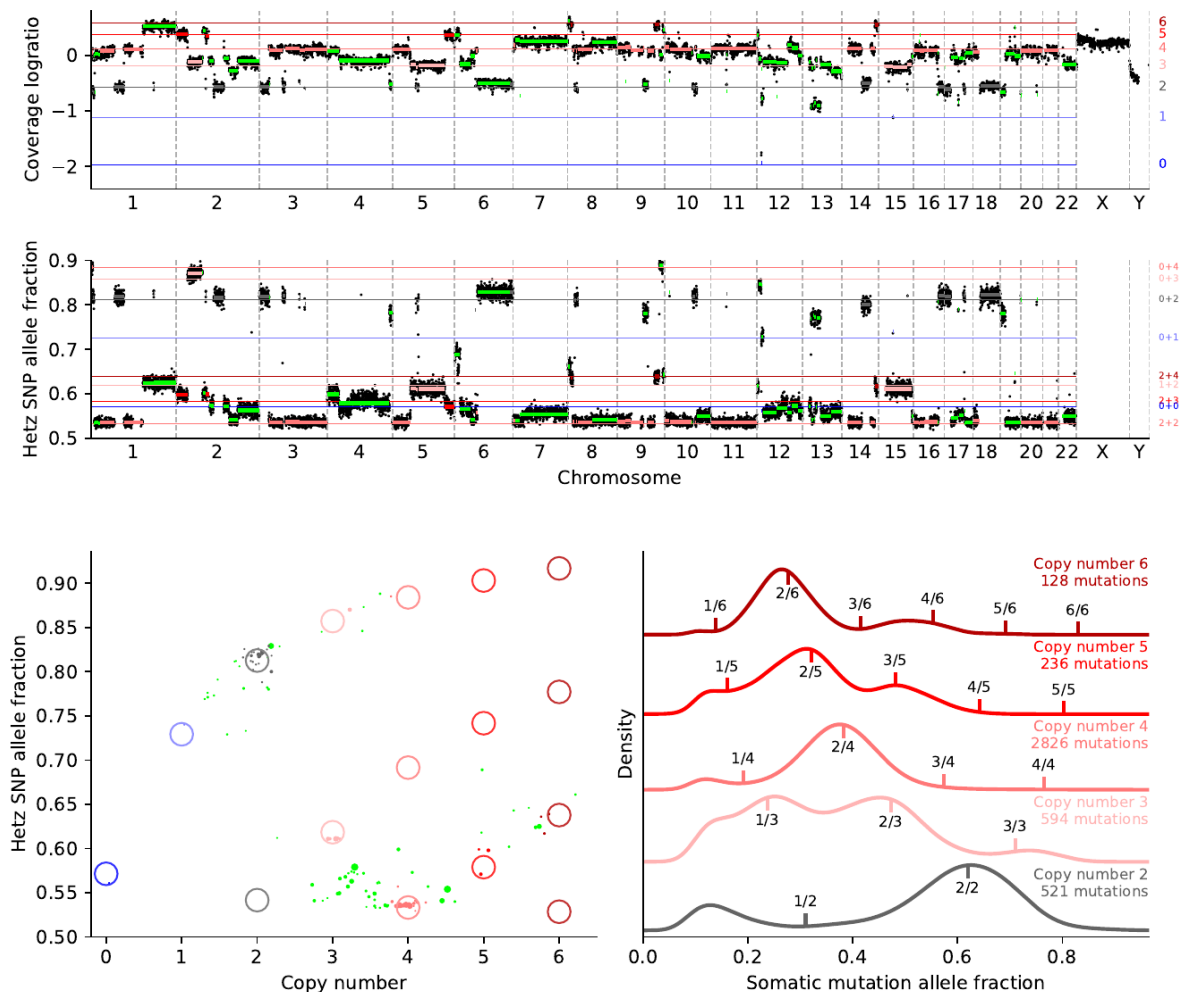


DTB-097-Progression-cfDNA

Battenberg model - Cancer fraction: 78%, Average ploidy: 1.82, Diploid level: 0.11

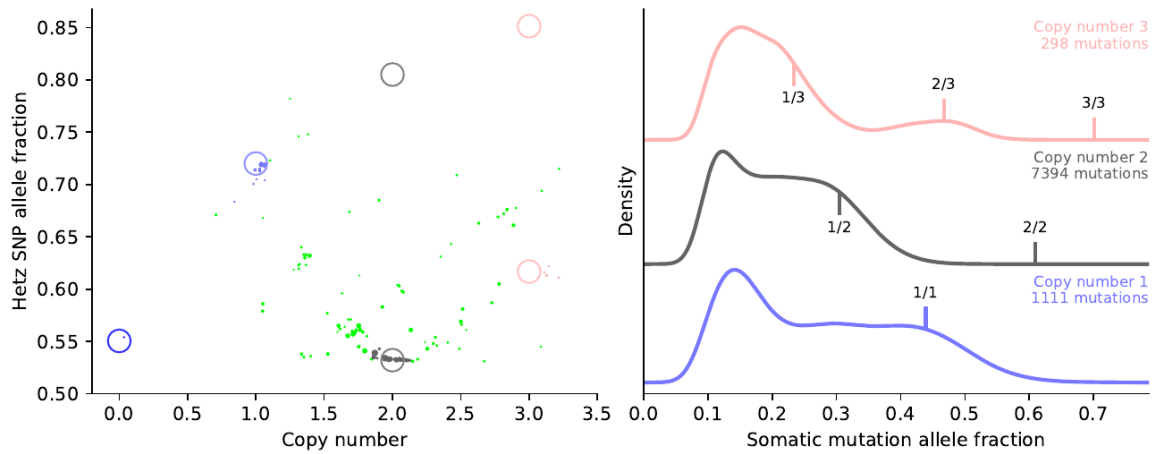
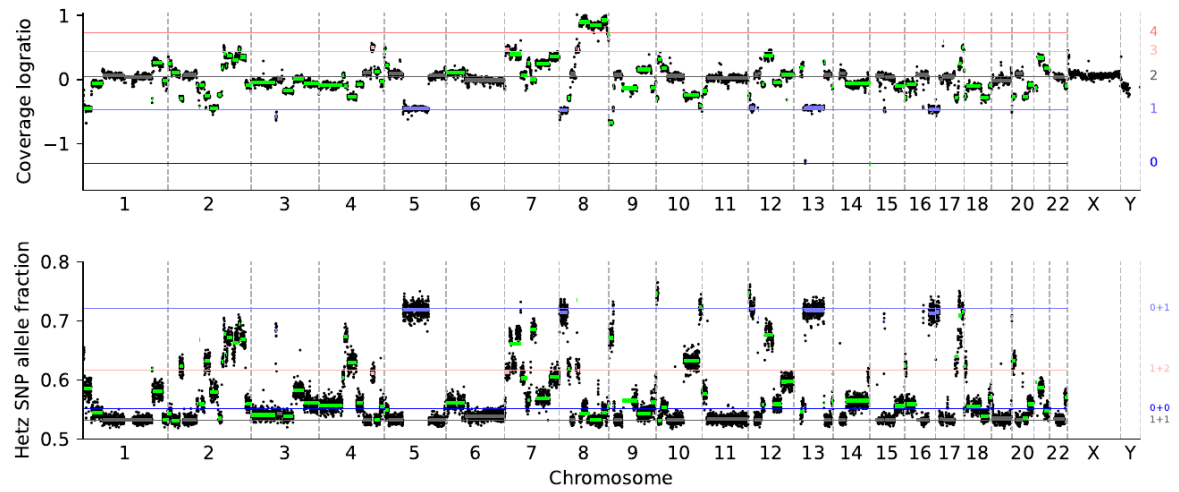


Our model - Cancer fraction: 62%, Average ploidy: 3.64, Diploid level: -0.58

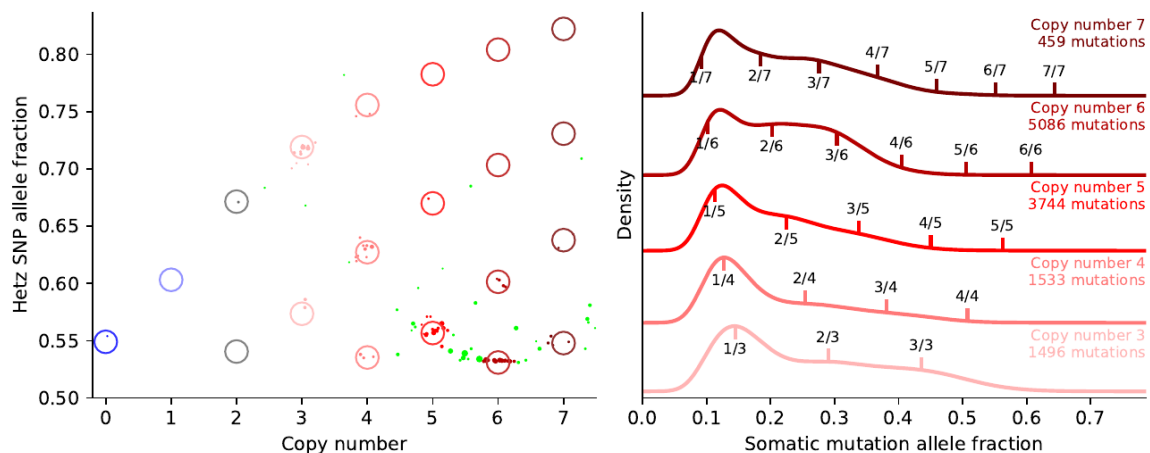
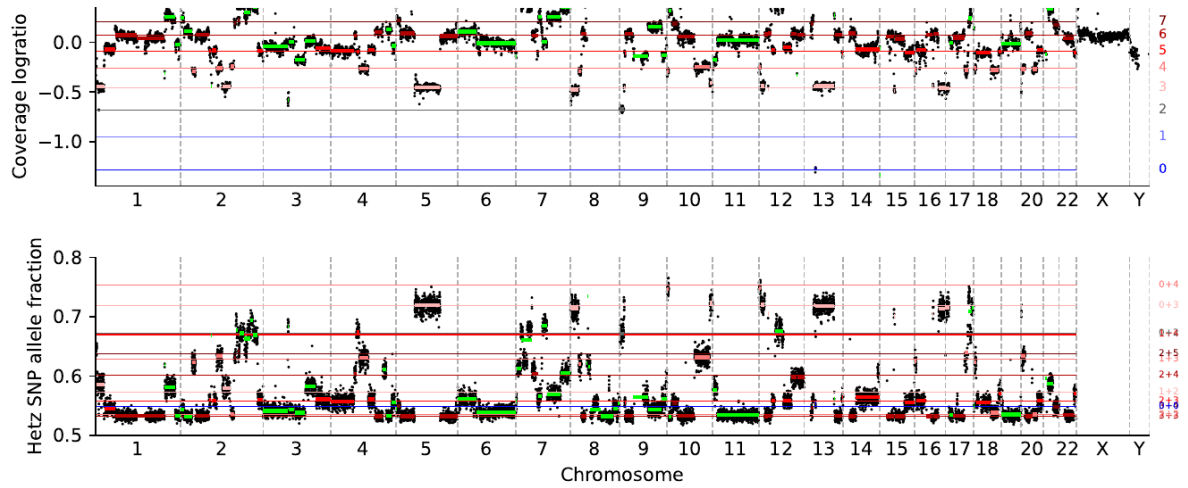


DTB-174-Baseline-cfDNA

Battenberg model - Cancer fraction: 61%, Average ploidy: 1.96, Diploid level: 0.05

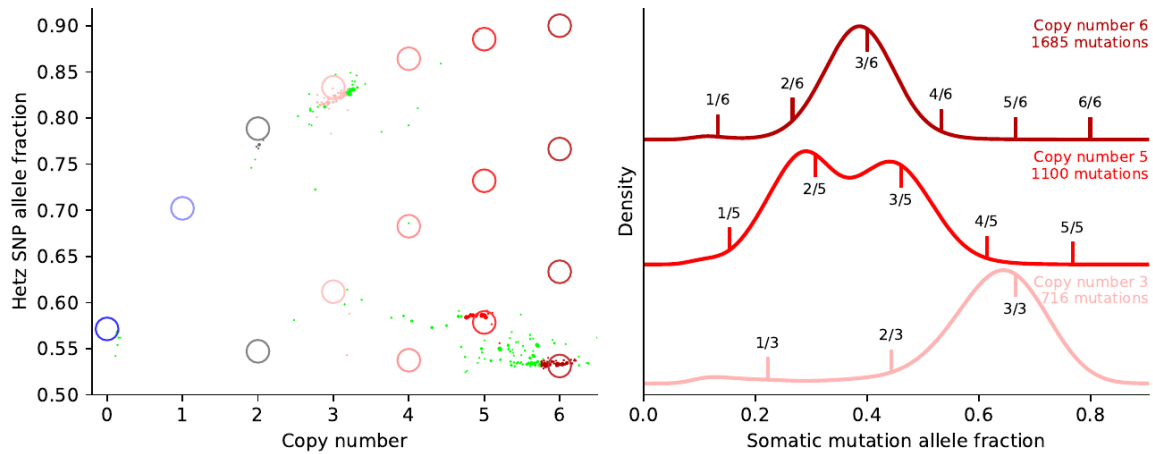
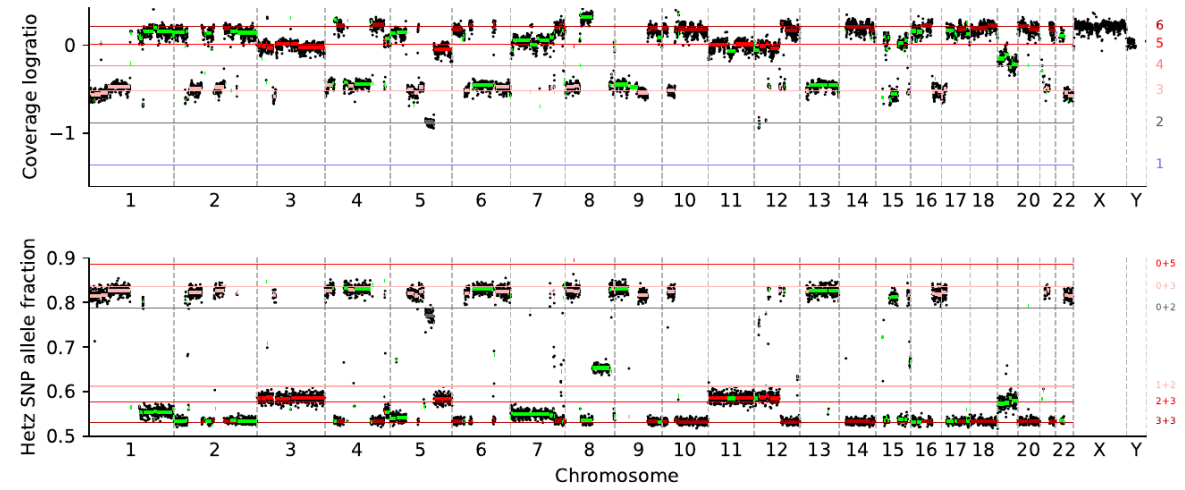


Our model - Cancer fraction: 34%, Average ploidy: 5.76, Diploid level: -0.68

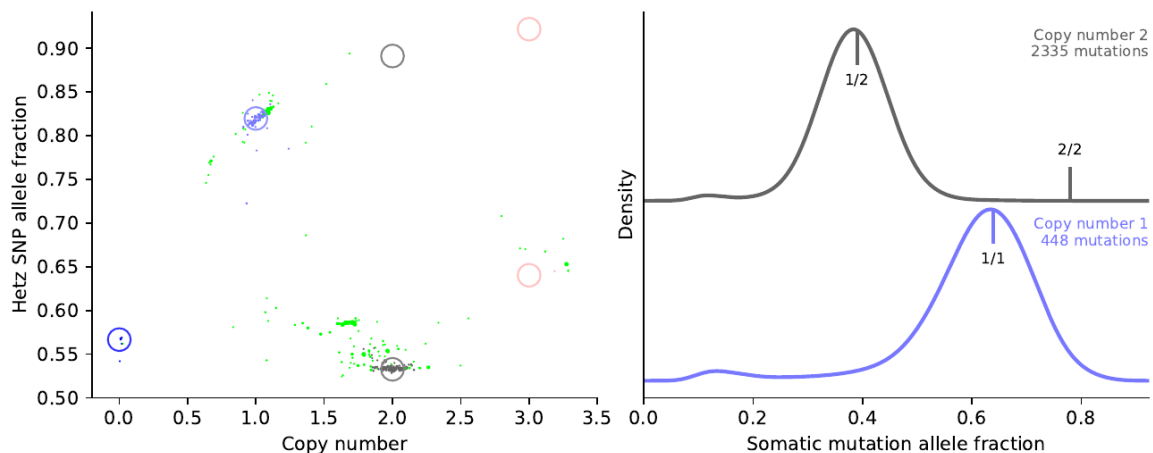
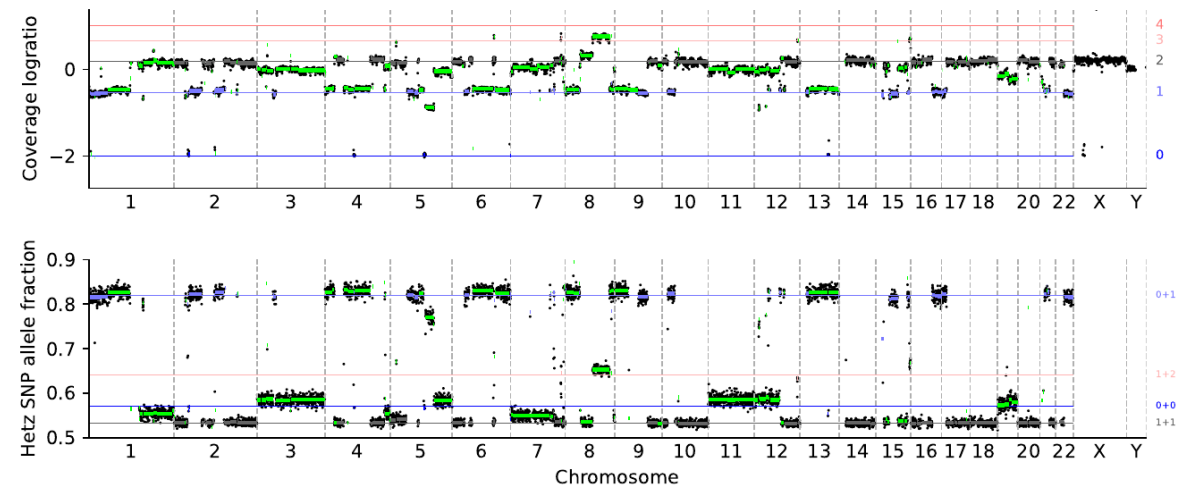


DTB-210-Progression-cfDNA

Battenberg model - Cancer fraction: 57%, Average ploidy: 4.74, Diploid level: -0.88

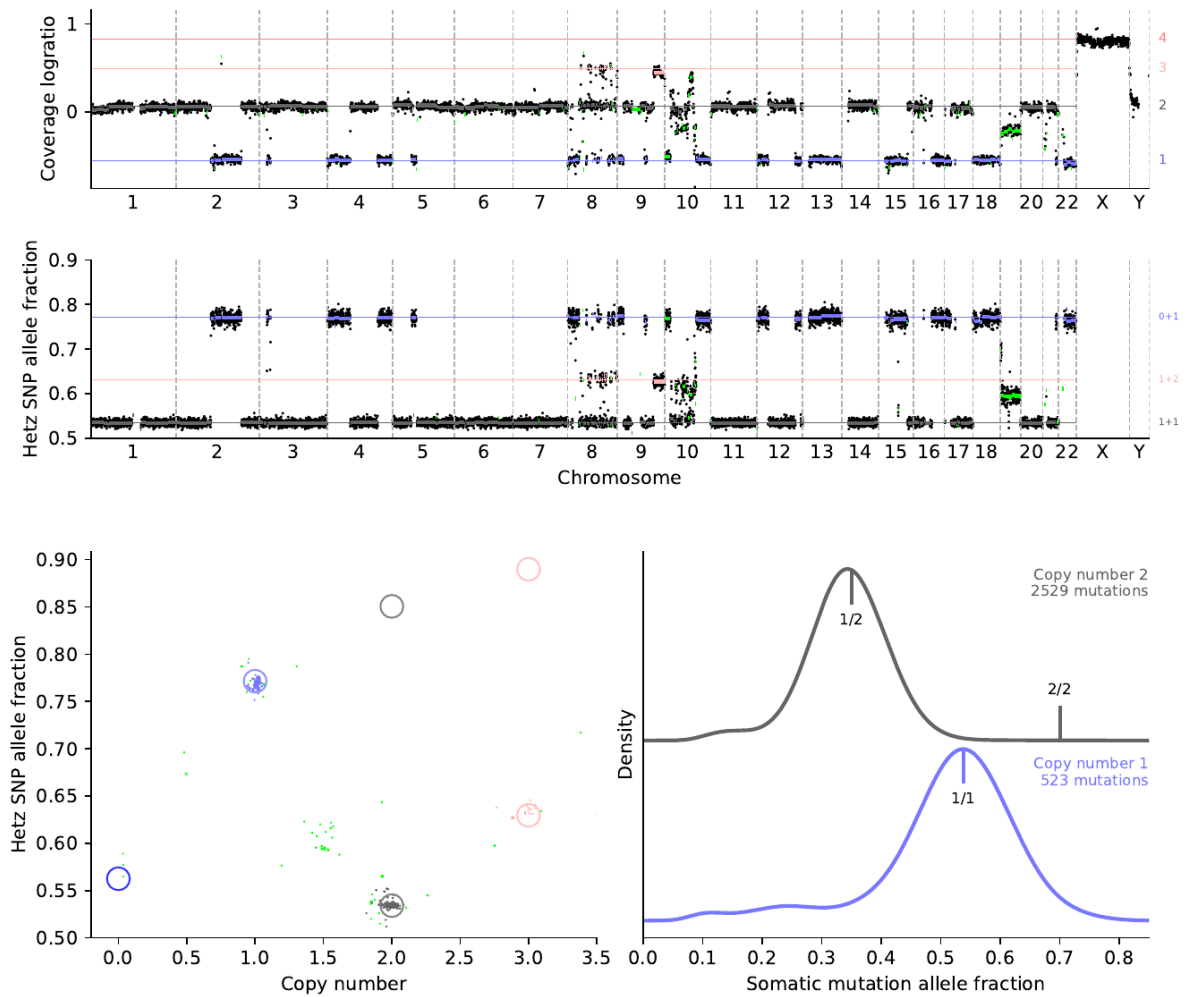


Our model - Cancer fraction: 78%, Average ploidy: 1.65, Diploid level: 0.18

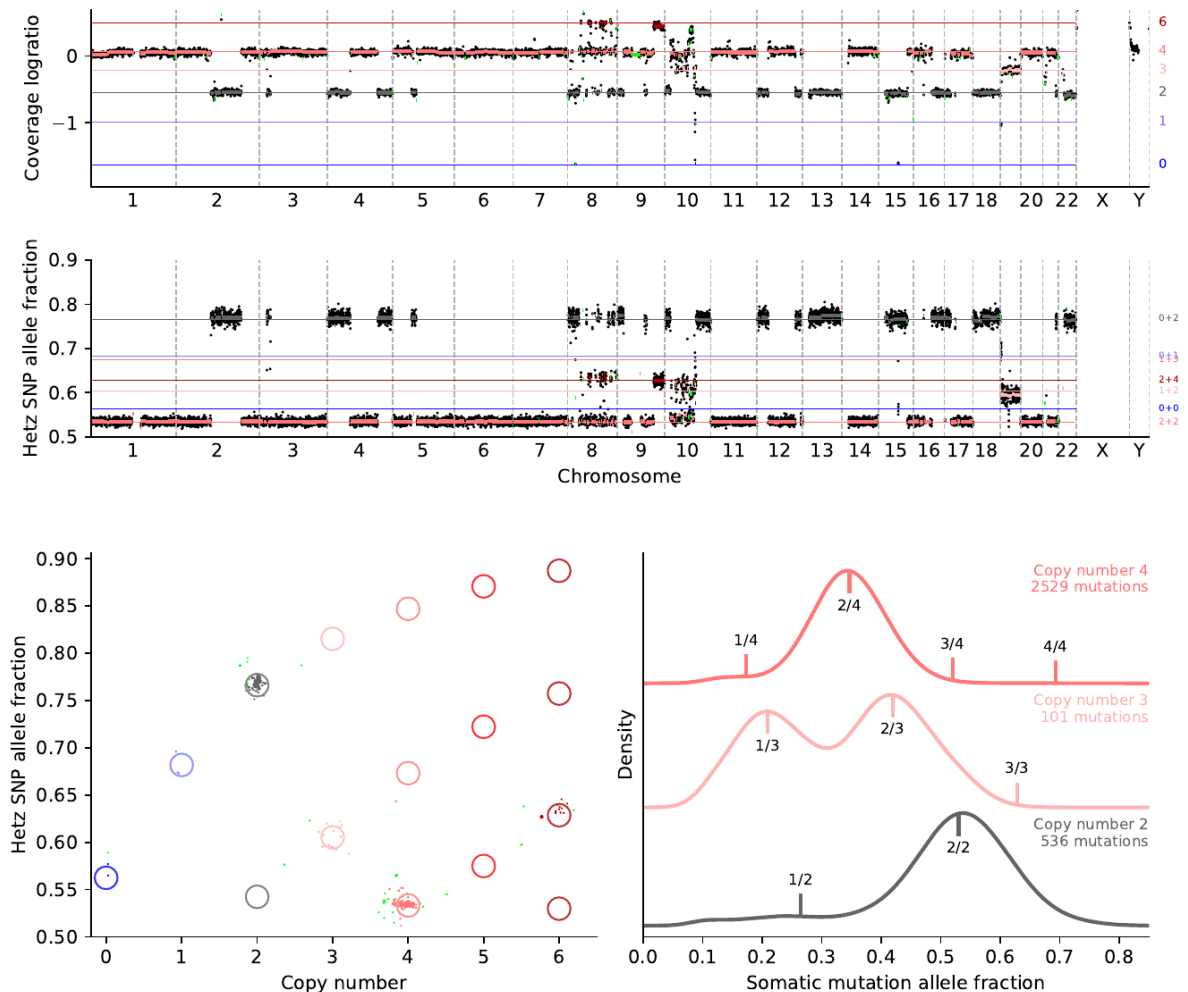


DTB-216-Baseline-cfDNA

Battenberg model - Cancer fraction: 70%, Average ploidy: 1.74, Diploid level: 0.06

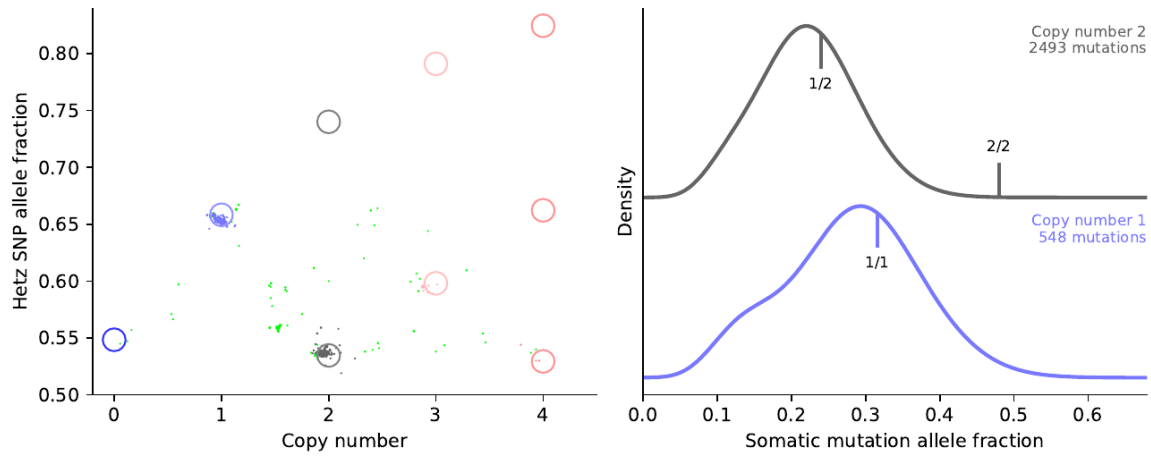
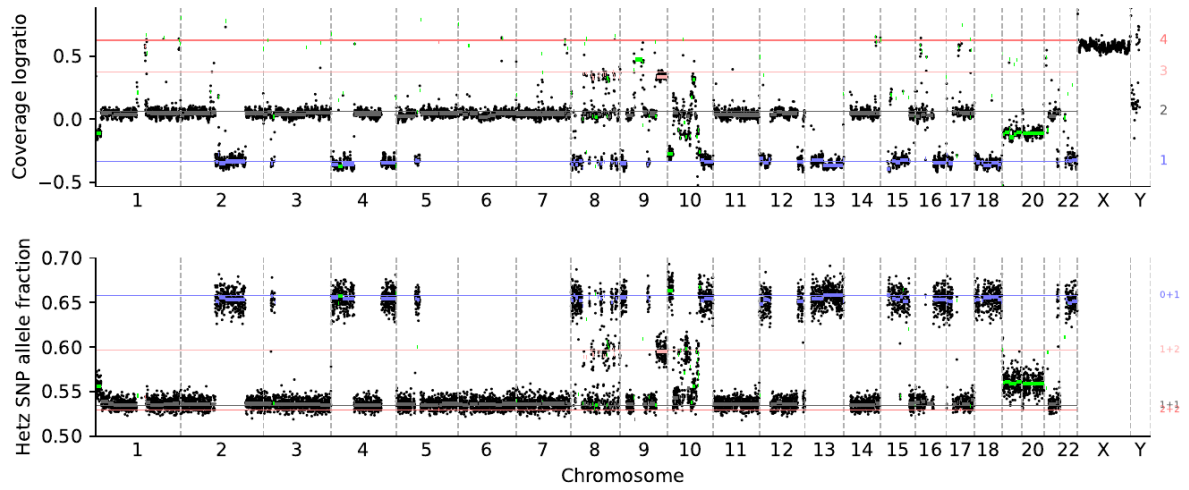


Our model - Cancer fraction: 53%, Average ploidy: 3.7, Diploid level: -0.55

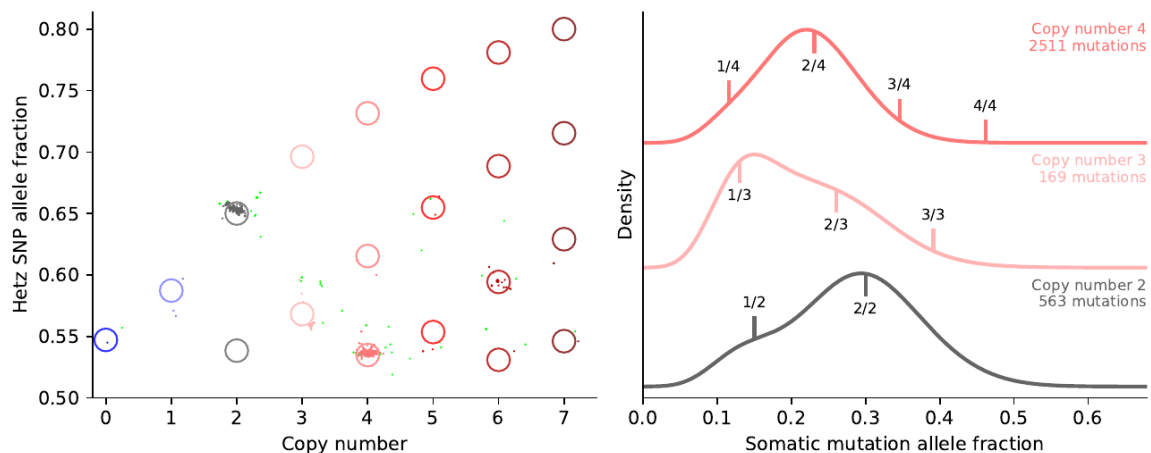
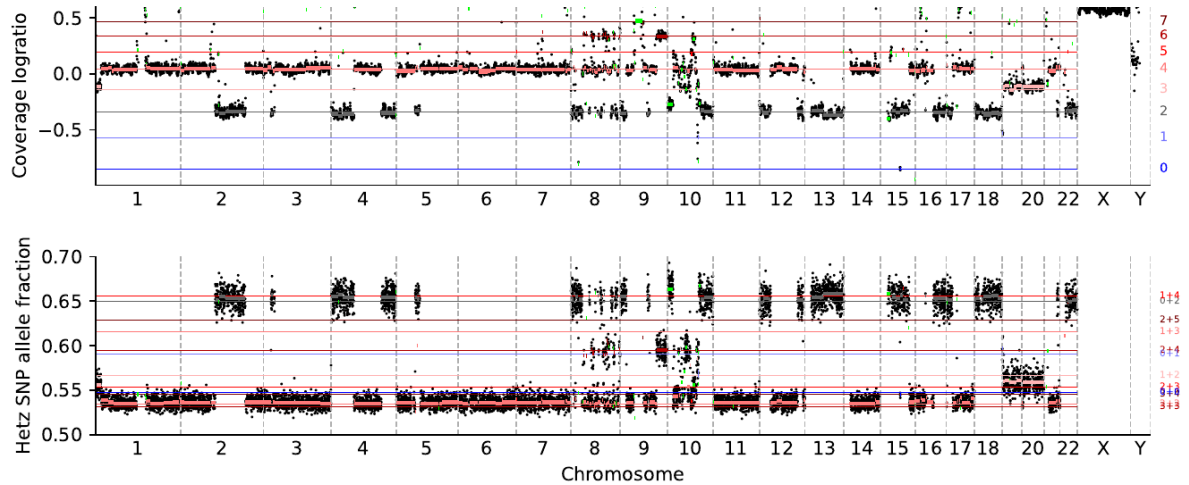


DTB-216-Progression-Tumor

Battenberg model - Cancer fraction: 48%, Average ploidy: 1.74, Diploid level: 0.06

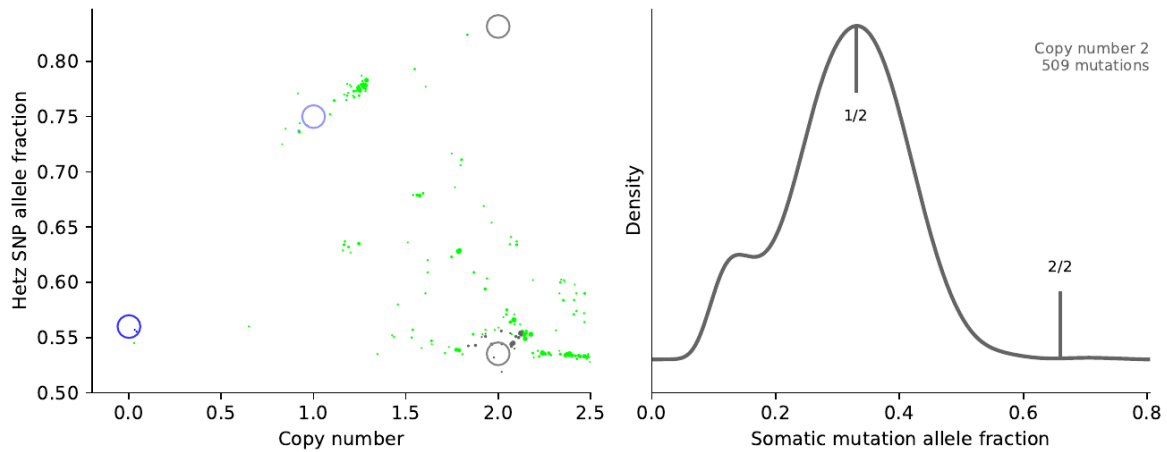
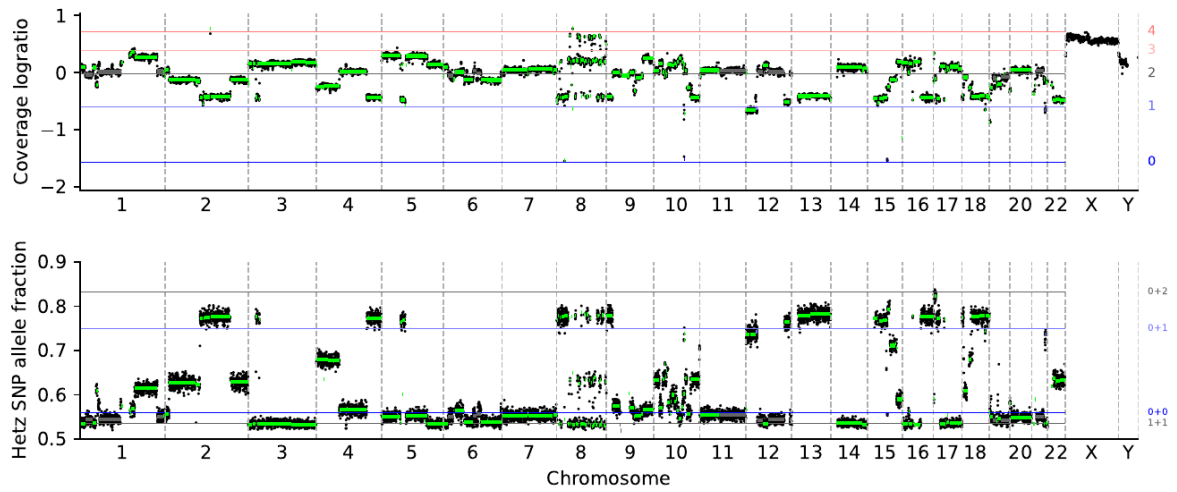


Our model - Cancer fraction: 30%, Average ploidy: 3.83, Diploid level: -0.34

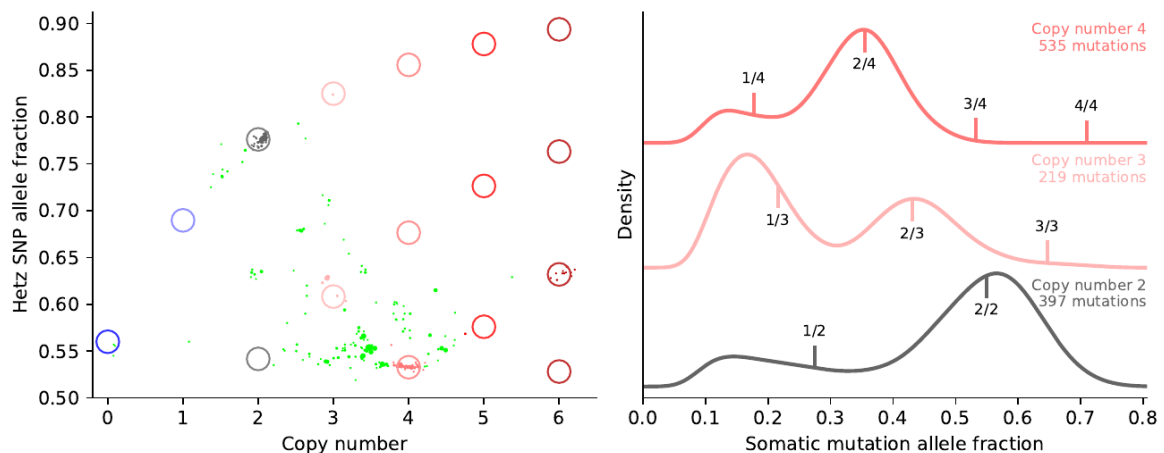
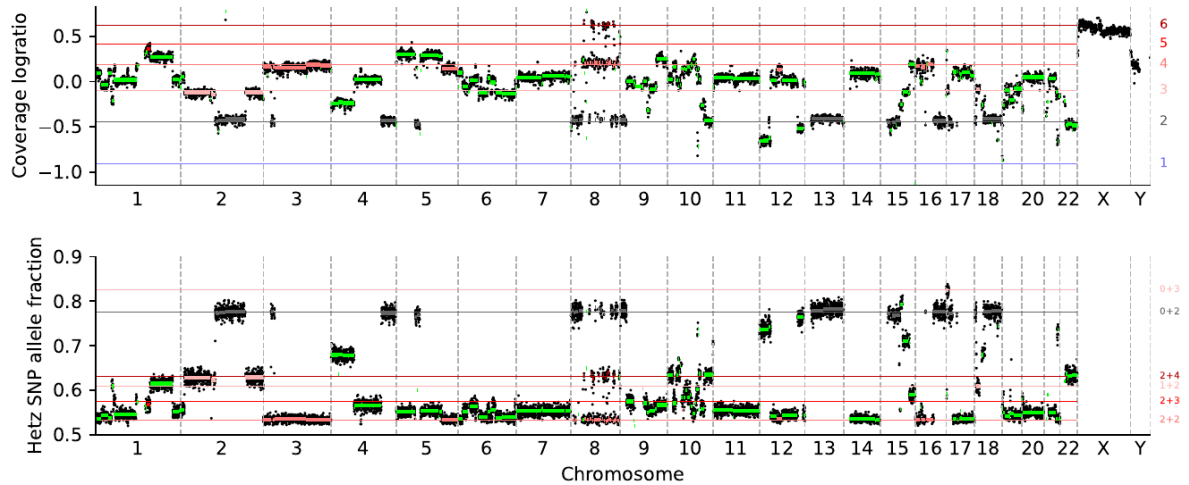


DTB-216-Progression-cfDNA

Battenberg model - Cancer fraction: 66%, Average ploidy: 1.98, Diploid level: -0.02

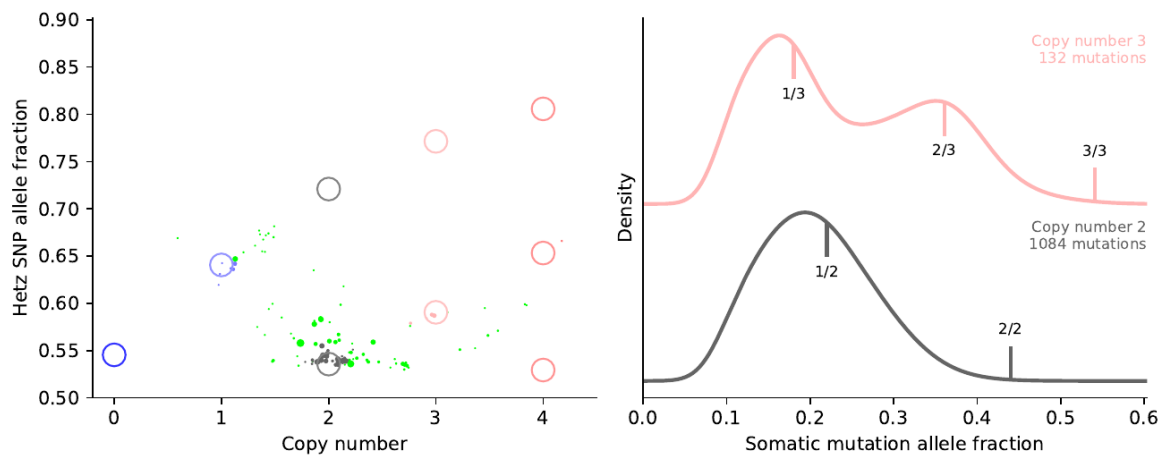
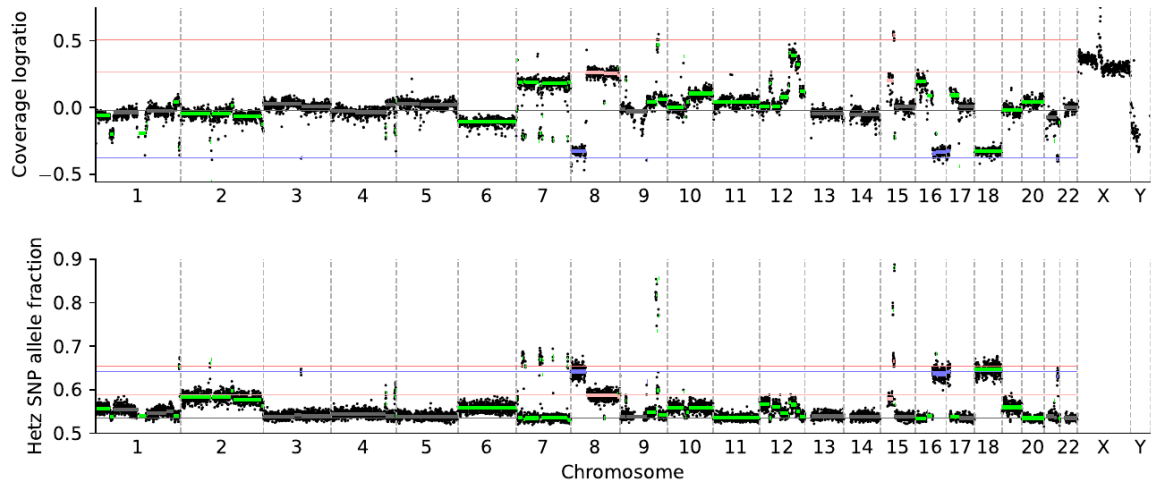


Our model - Cancer fraction: 55%, Average ploidy: 3.37, Diploid level: -0.45

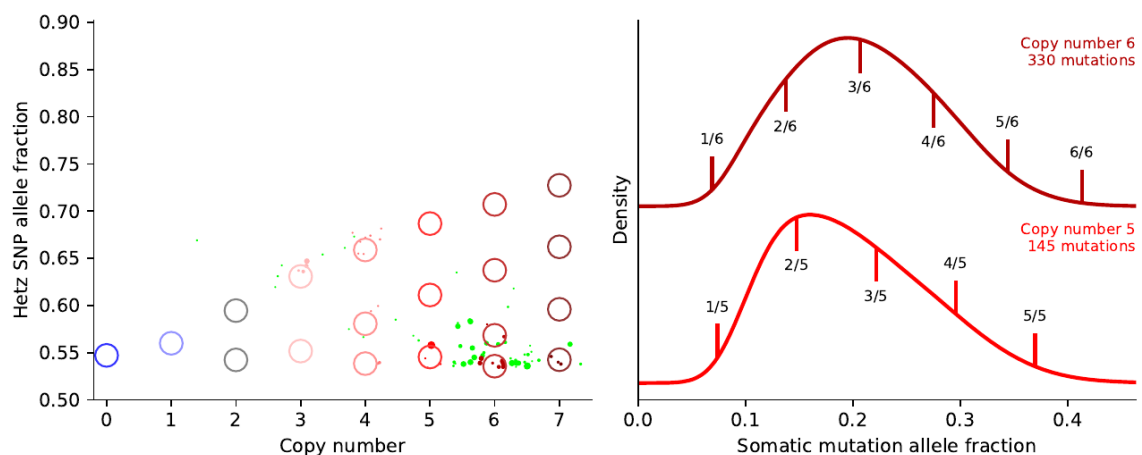
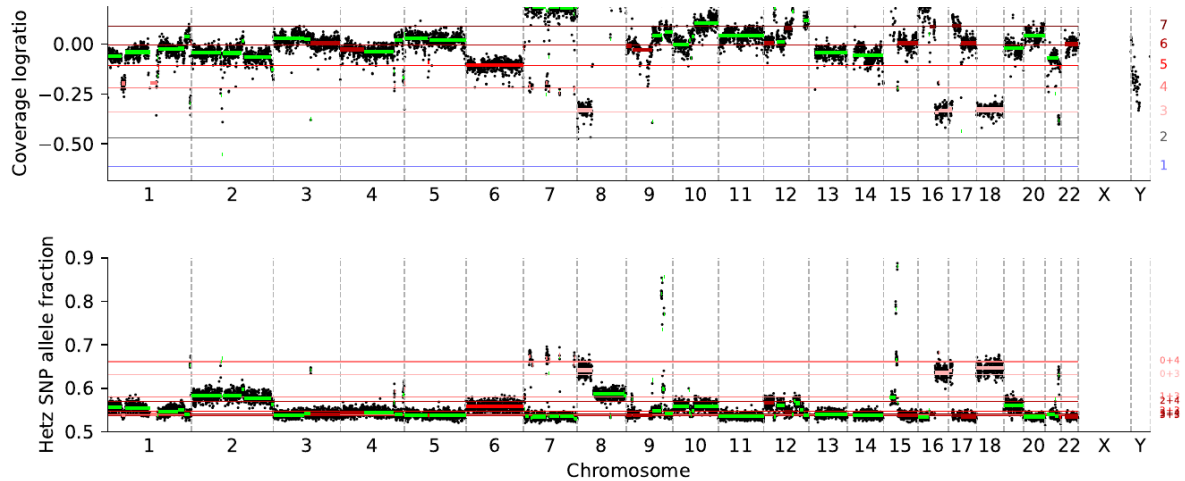


DTB-266-Baseline-cfDNA

Battenberg model - Cancer fraction: 44%, Average ploidy: 2.08, Diploid level: -0.02



Our model - Cancer fraction: 19%, Average ploidy: 6.3, Diploid level: -0.47



Supplementary references

1. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
2. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14**, R80 (2013).
3. Beerwinkler, N., Schwarz, R. F., Gerstung, M. & Markowitz, F. Cancer Evolution: Mathematical Models and Computational Inference. *Systematic Biology* vol. 64 e1–e25 (2015).
4. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* vol. 26 841–842 (2010).
5. van Dijk, E. *et al.* Chromosomal copy number heterogeneity predicts survival rates across cancers. *Nat. Commun.* **12**, 3188 (2021).
6. Tarabichi, M. *et al.* A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat. Methods* **18**, 144–155 (2021).
7. Steffen, M. A simple method for monotonic interpolation in one dimension. *Astron. Astrophys.* **239**, 443 (1990).
8. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
9. Annala, M. *et al.* Evolution of Castration-Resistant Prostate Cancer in ctDNA during Sequential Androgen Receptor Pathway Inhibition. *Clin. Cancer Res.* **27**, 4610–4623 (2021).
10. Annala, M. *et al.* Circulating Tumor DNA Genomics Correlate with Resistance to Abiraterone and Enzalutamide in Prostate Cancer. *Cancer Discov.* **8**, 444–457 (2018).
11. Warner, E. *et al.* BRCA2, ATM, and CDK12 Defects Differentially Shape Prostate Tumor Driver Genomics and Clinical Aggression. *Clinical Cancer Research* vol. 27 1650–1662 (2021).
12. Wu, Y.-M. *et al.* Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. *Cell* **173**, 1770–1782.e14 (2018).
13. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
14. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
16. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
17. Cameron, D. L. *et al.* GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**, 202 (2021).
18. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature Methods* vol. 11 396–398 (2014).
19. Caravagna, G. *et al.* Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet.* **52**, 898–907 (2020).
20. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* vol. 578 94–101 (2020).
21. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
22. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **175**, 889 (2018).
23. Harrigan, C. F., Rubanova, Y., Morris, Q. & Selega, A. TrackSigFreq: subclonal reconstructions based on mutation signatures and allele frequencies. *Pac. Symp. Biocomput.* **25**, 238–249 (2020).
24. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* vol. 30 1006–1007 (2014).
25. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).