1. ECDNA count and presence statistics

Estimation of frequency of samples containing ECDNA:

There is a wide variation on the number of ECDNA across different samples and within metaphases of the same sample. We want to estimate and compare the frequency of samples containing ECDNA for each sample type. We label a sample as being *EC-positive* by using the pathology standard: a sample is deemed to be EC-positive if we observe ≥ 2 ECDNA in ≥ 2 images out of 20 metaphase images. Therefore, we ensure that every sample contains at least 20 metaphases.

We define indicator variable $X_{ij} = 1$ if metaphase image j in sample i has ≥ 2 ECDNA; $X_{ij} = 0$ otherwise. Let n_i be the number of metaphase images acquired from sample i. We assume that X_{ij} is the outcome of the j-th Bernoulli trial, where the probability of success p_i is drawn at random from a beta distribution with parameters determined by $\sum_j X_{ij}$. Formally,

$$p_i|\alpha_i, \beta_i \sim \text{Beta}(\alpha_i = \max\{\epsilon, \sum_j X_{ij}\}, \beta_i = \max\{\epsilon, n_i - \alpha_i\}). \tag{1.1}$$

We model the likelihood of observing k successes in $n = 20$ trials using the binomial density function as:

$$k|p_i \sim \text{Binom}(p_i, n = 20) \tag{1.2}$$

Finally, the *predictive* distribution $p(k)$, is computed using the product of the Binomial likelihood and Beta prior, modeled as a "beta-binomial distribution" [1].

$$\begin{aligned} p(k) = \mathbb{E}_{p_i}[k|p_i] &= \int_0^1 k|p_i \cdot p_i|\alpha_i, \beta_i \, \mathrm{d}p_i \\ &= \int_0^1 \binom{n}{k} p_i^k (1-p_i)^{n-k} \cdot \frac{1}{\mathrm{B}(\alpha_i, \beta_i)} p_i^{\alpha_i - 1} (1-p_i)^{\beta_i - 1} \, \mathrm{d}p_i \\ &= \binom{n}{k} \frac{1}{\mathrm{B}(\alpha_i, \beta_i)} \int_0^1 p_i^{k+\alpha_i-1} (1-p_i)^{n-k+\beta_i-1} \, \mathrm{d}p_i \\ &= \binom{n}{k} \frac{\mathrm{B}(k+\alpha_i, n-k+\beta_i)}{\mathrm{B}(\alpha_i, \beta_i)} \end{aligned} \tag{1.3}$$

We model the probability for sample i being EC-positive with the random variable Y_i such that:

$$\begin{aligned} Y_i &= 1 - Pr(\text{sample } i \text{ is EC-negative}) \\ &= 1 - (k = 1|p_i) - (k = 0|p_i) \end{aligned} \tag{1.4}$$

The expected value of Y_i is:

$$\begin{aligned}\mathbb{E}_{p_i}(Y_i) &= 1 - p(k=1) - p(k=0) \\ &= 1 - \binom{20}{1}\frac{\mathrm{B}(1+\alpha_i, 19+\beta_i)}{\mathrm{B}(\alpha_i,\beta_i)} - \binom{20}{0}\frac{\mathrm{B}(\alpha_i, 20+\beta_i)}{\mathrm{B}(\alpha_i,\beta_i)}\end{aligned} \tag{1.5}$$

The variance of Y_i is:

$$\mathrm{Var}(Y_i) = \mathrm{Var}(k=1|p_i) + \mathrm{Var}(k=0|p_i) + 2\mathrm{Cov}(k=1|p_i, k=0|p_i), \tag{1.6}$$

where,

$$\begin{aligned}&\mathrm{Var}(k|p_i) = \mathbb{E}_{p_i}[(k|p_i)^2] - \mathbb{E}_{p_i}[k|p_i]^2 \\ &= \int_0^1 (k|p_i)^2 \cdot p_i|\alpha_i,\beta_i \,\mathrm{d}p_i - (\int_0^1 k|p_i \cdot p_i|\alpha_i,\beta_i)^2 \,\mathrm{d}p_i \\ &= \binom{n}{k}\binom{n}{k}\frac{1}{\mathrm{B}(\alpha_i,\beta_i)}\int_0^1 p_i^{2k+\alpha_i-1}(1-p_i)^{2n-2k+\beta_i-1}\,\mathrm{d}p_i - \binom{n}{k}\binom{n}{k}\frac{\mathrm{B}(k+\alpha_i, n-k+\beta_i)^2}{\mathrm{B}(\alpha_i,\beta_i)^2} \\ &= \binom{n}{k}\binom{n}{k}\frac{1}{\mathrm{B}(\alpha_i,\beta_i)}[\mathrm{B}(2k+\alpha_i, 2n-2k+\beta_i) - \frac{\mathrm{B}(k+\alpha_i, n-k+\beta_i)^2}{\mathrm{B}(\alpha_i,\beta_i)}],\end{aligned} \tag{1.7}$$

and

$$\begin{aligned}&\mathrm{Cov}(k=1|p_i, k=0|p_i) = \mathbb{E}_{p_i}[k=1|p_i \cdot k=0|p_i] - \mathbb{E}_{p_i}[k=0|p_i]\,\mathbb{E}_{p_i}[k=1|p_i] \\ &= \binom{n}{0}\binom{n}{1}\frac{1}{\mathrm{B}(\alpha_i,\beta_i)}[\int_0^1 p_i^{1+\alpha_i-1}(1-p_i)^{2n-1+\beta_i-1}\,\mathrm{d}p_i - \frac{\mathrm{B}(\alpha_i, n+\beta_i)\mathrm{B}(1+\alpha_i, n-1+\beta_i)}{\mathrm{B}(\alpha_i,\beta_i)} \\ &= \binom{n}{0}\binom{n}{1}\frac{1}{\mathrm{B}(\alpha_i,\beta_i)}[\mathrm{B}(1+\alpha_i, 2n-1+\beta_i) - \frac{\mathrm{B}(\alpha_i, n+\beta_i)\mathrm{B}(1+\alpha_i, n-1+\beta_i)}{\mathrm{B}(\alpha_i,\beta_i)}].\end{aligned} \tag{1.8}$$

Let T be the set of samples belonging to a certain sample type t, e.g. immortalized samples. We define

$$Y_T = \frac{\sum_{i\in T} Y_i}{|T|} \tag{1.9}$$

We estimate the frequency of samples under sample t containing ECDNA (bar heights on Figures 2C and 2D) as

$$\mathbb{E}[Y_T] = \frac{\sum_{i\in T}\mathbb{E}[Y_i]}{|T|} \tag{1.10}$$

and error bar heights (Figure 2C and 2D) as:

$$\mathrm{sd}(Y_T) = \frac{(\sum_{i\in T}\mathrm{Var}[Y_i])^{\frac{1}{2}}}{|T|} \tag{1.11}$$

assuming independence among samples $i \in T$. For any α_i or $\beta_i = 0$, we assign them a sufficiently small ϵ.

2. ECdetect: Software for detection of extrachromosomal DNA from DAPI staining metaphase images

2.1 Introduction

The DAPI staining metaphase image extrachromosomal DNA (ECDNA) detection software provides a conservative estimation to the number of ECDNA in DAPI staining metaphase images. The software performs a pre-segmentation of the image in order to distinguish chromosomal and non-chromosomal structures, and computes an ECDNA search region of interest (ROI). The designated ROI is displayed on a user interface for the investigator to modify via masking and unmasking desired regions on the image, to correct for potential inaccurate segmentation and/or exclude debris from the ROI. The modifications made on the ROI are saved once verified, and are available for future usage. The output of the software includes the original images with ECDNA detections overlayed, the count of ECDNA found, and their coordinates in the image. ECdetect does not require a pan-centromeric probe, and works on DAPI staining metaphase images only, therefore any detected ECDNA is assumed to not contain a centromere.

2.2 Software

Input

The ECDNA detection software uses Tagged Image File Format (.tiff) DAPI staining metaphase images. In this project we used 2572 images, after checking for duplicates, each at resolution 1392x1040. The investigator needs to provide the parent folder containing all imaging data as input and no other parameter will be required. The software will recursively process every tiff image under the parent folder.

Image pre-segmentation

The software applies an initial coarse adaptive thresholding [2, 3] to detect the major components in the image, with a window size of $150x150$ pixels, and $T = 10\%$. After filling the closed structures, components breaching 3000 pixels and 80% of solidity (the ratio of the area of the component to the area of its convex hull) are masked as non-chromosomal regions in order to remove the intact nuclei regions from subsequent analysis. Small components are also discarded, and the remaining image is accepted as the binary chromosomal image (BCI). The weakly connected components of the BCI are computed to find the separate chromosomal regions. The weakly connected components breaching a cumulative pixel count of 5000

are considered as candidate search regions, and their convex hull with a dilation of 100 pixels are added into the ECDNA search region of interest (ROI).

ROI verification

The software provides a user interface as shown in Figure S2.1, where the original DAPI image is displayed next to its segmentation result, alongside an overview image.

We manually masked any non-chromosomal region that the software failed to discard during the pre-segmentation as shown in Figure S2.2. Similarly, we also unmasked any region that the software mistakenly discarded as non-chromosomal region. The segmentation results are displayed in three colors: teal (chromosomal region qualified to be inside of the search region), dark blue (non-chromosomal/masked region), and green (chromosomal or small components not qualified to be inside of the search region). The color orange shows the current ECDNA search ROI. At the end of every masking/un-masking, the ECDNA search ROI is recomputed based on the newly generated BCI and displayed.

ECDNA detection

Figure S2.3 shows the steps of ECDNA detection. After the verification of the ECDNA search ROI (Figure S2.3a), the software applies a 2-D Gaussian smoothing to the image with standard deviation of 0.5, performs a second finer adaptive thresholding, with a window size of $20x20$ pixels and $T = 7\%$, and fills any closed structures. Components that are greater than 75 pixels are designated as non-ECDNA structures and their 15-pixel neighborhood is removed from the ECDNA search ROI, in order not to mistakenly call chromosomal extensions or other near intact nuclei structures as ECDNA (Figure S2.3b). Any component detected with a size less than or equal to 75 and greater than or equal to 3 pixels inside the final search ROI is returned as ECDNA (Figure S2.3c).

Output

The detected ECDNA elements are shown in the original image with overlayed red circles, as well as their coordinates in a separate file for every image. The total ECDNA count per image is also recorded.

Manual ECDNA marking

For ECDNA detection evaluation purposes, we allowed the investigator to manually select the ECDNA structures while being able to have access to the verified ECDNA search region (including the chromosome region neighborhood) and segmentation results, alongside zooming, if desired. Figure S2.4 shows an example set of marked ECDNA at a specified zooming level.

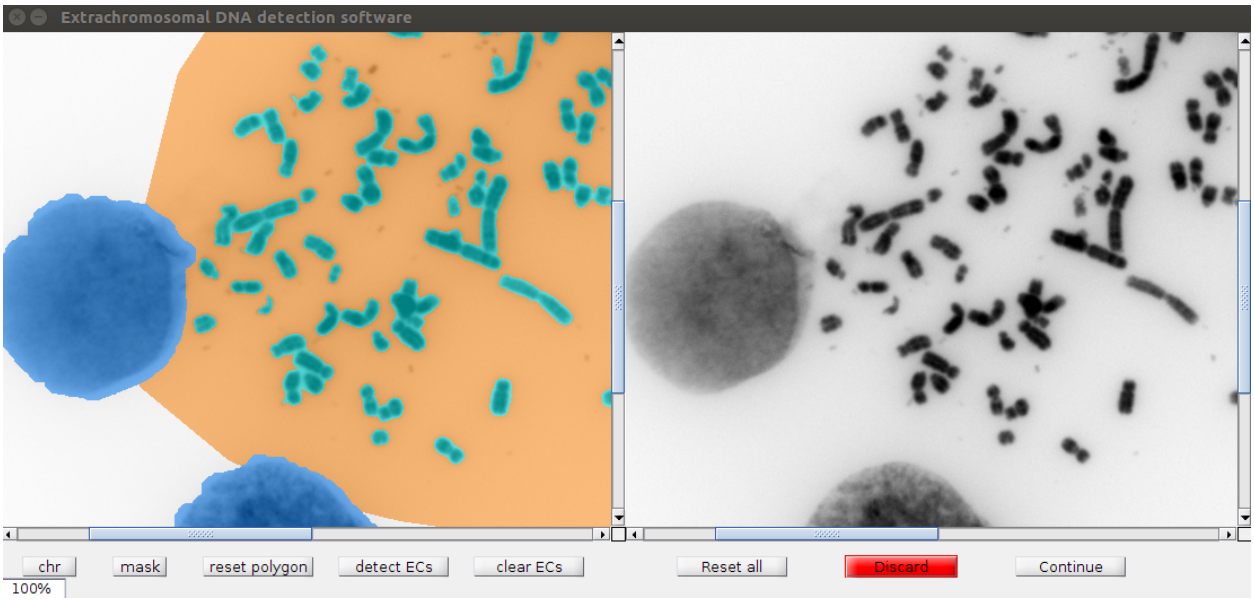
Comparison of software vs. visual inspection

The ECDNA coordinates detected by the software and selected by manual marking are compared and they are accepted to match if the distance between them is no more than 7 pixels. A sample comparison result is shown in Figure S2.10. The green circles show the software detected ECDNA coordinates that agree with manually marked ECDNA, blue circles show manually marked ECDNA that the software missed, and red circles show software detected ECDNA that were not manually marked. Notice that a majority of blue circles appear in the immediate neighborhood of chromosomal structures, which we deliberately removed from the ECDNA search ROI. The red circles appear to have faint pixel intensities, which the visual inspection may have missed or discarded.

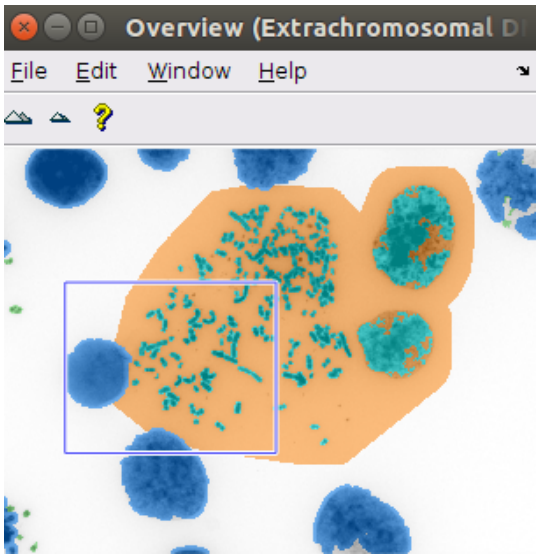
2.3 Results

We arbitrarily chose 28 images, in which we could confidently mark the ECDNA, while also aiming for a large range of ECDNA count across images, from various different tumor cell lines for purposes of robustness. We evaluated the performance of the ECDNA detection software by comparing it with manual ECDNA marking on the aforementioned 28 DAPI metaphase images from various tumor cell lines with varying count of ECDNAs. The comparison results are shown in Figures S2.5-S2.32 for each picture in detail. Out of 406 detected ECDNA, 392 of them (97%) agreed with manually marked ECDNAs, however among the 737 total manually marked ECDNAs, the software missed 345 of them, resulting in a underestimation by 53%. We would like to emphasize, however, that it was by design to discard the regions at the immediate neighborhood of non-ECDNA structures, e.g. chromosomal regions, from the ECDNA search ROI and undercall ECDNAs in order not to accept any questionable structure as extrachromosomal DNA. Indeed, 88% of the ECDNAs missed by the software compared to manual marking resides in the aforementioned discarded region. The software provides a conservative estimate of the total ECDNA signal; it achieves high precision at the expense of sensitivity compared to visual inspection, which may also have imperfections. Figure 1F shows the high correlation (Pearson; $r = 0.98$, $P < 2.2 \times 10^{-16}$) achieved between the ECDNA counts detected by the software and manual marking, suggesting a balanced undercalling of ECDNAs accross images, and a reliable estimation for correlative studies.

We also show the ECDNA count histograms for all samples analyzed by ECdetect in Figures S2.33-S2.41.

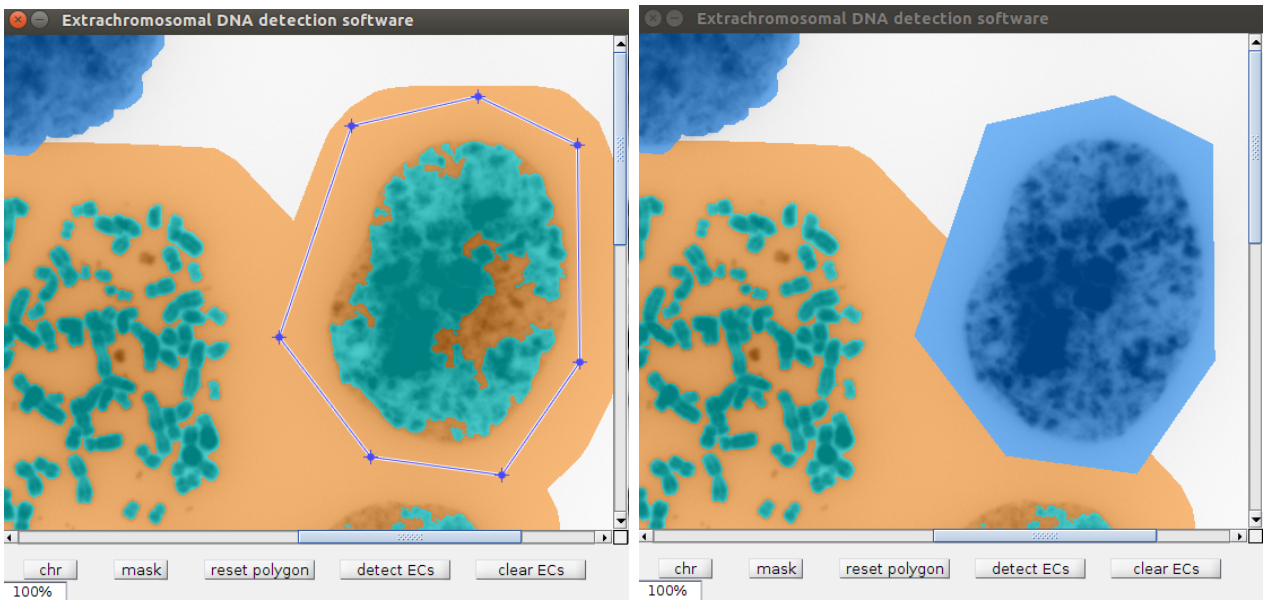


(a) Pre-segmented and original DAPI images.



(b) Overview of pre-segmentation.

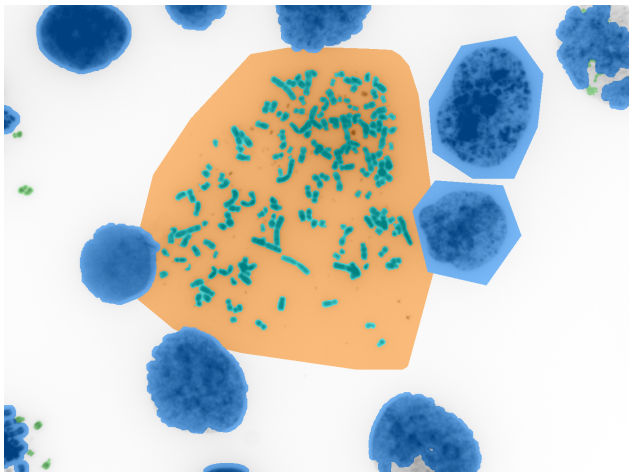
Figure S2.1: User interface for ECDNA search ROI verification.



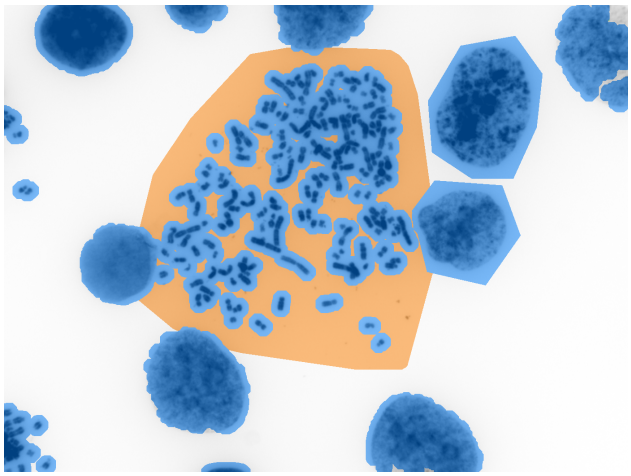
(a) Selection of the undesired region.

(b) Masking and removing from the ECDNA search ROI.

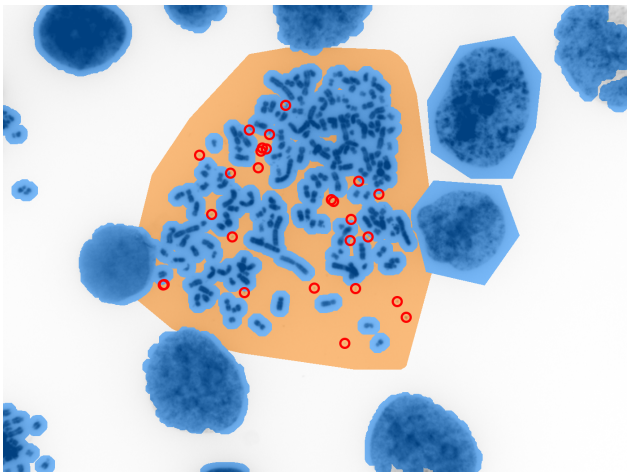
Figure S2.2: Non-chromosomal region masking.



(a) Step 1: Verified ECDNA search ROI.



(b) Step 2: 15-pixel neighborhood of any larger than ECDNA structure is removed.



(c) Step 3: ECDNA detection on final search ROI.

Figure S2.3: ECDNA detection steps.

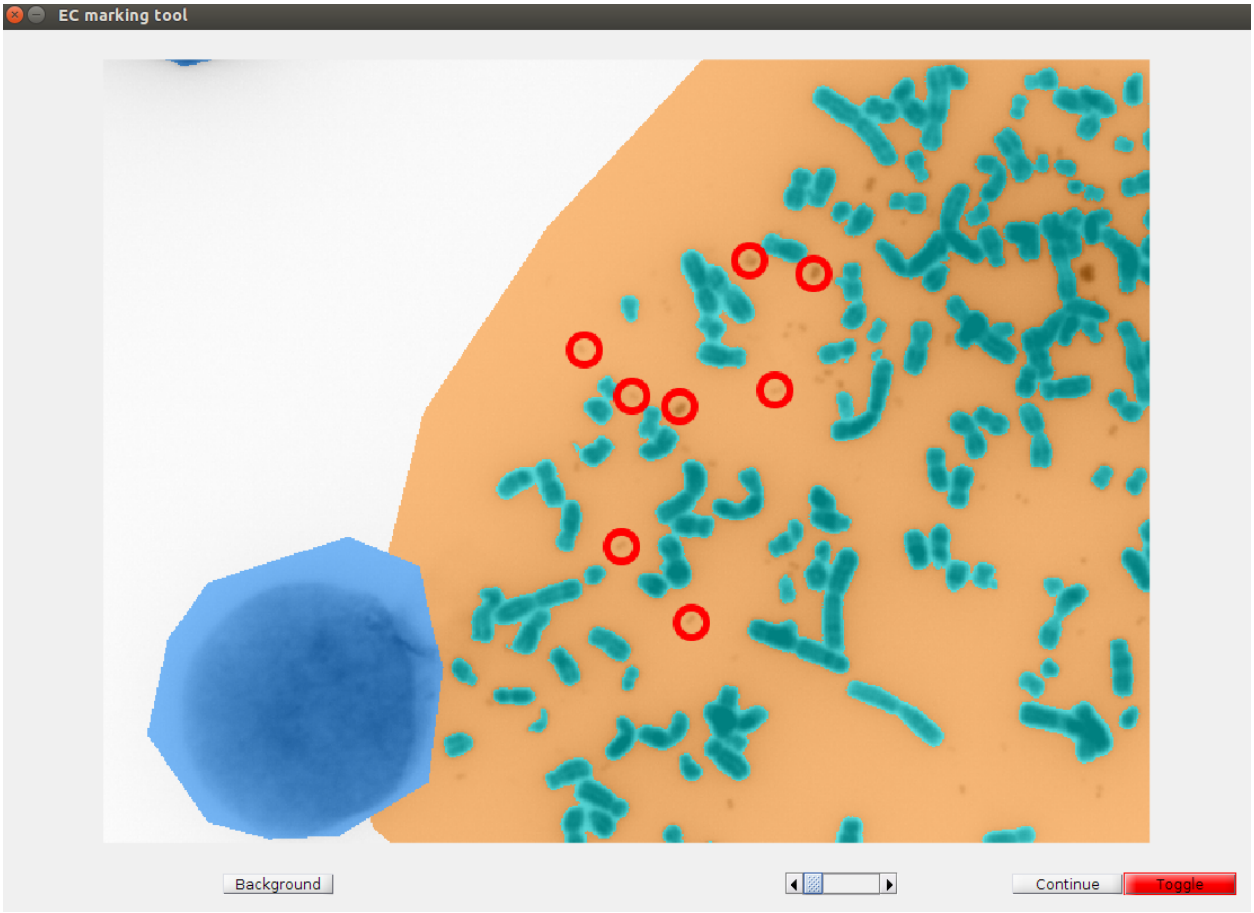


Figure S2.4: Manual marking of ECDNA.

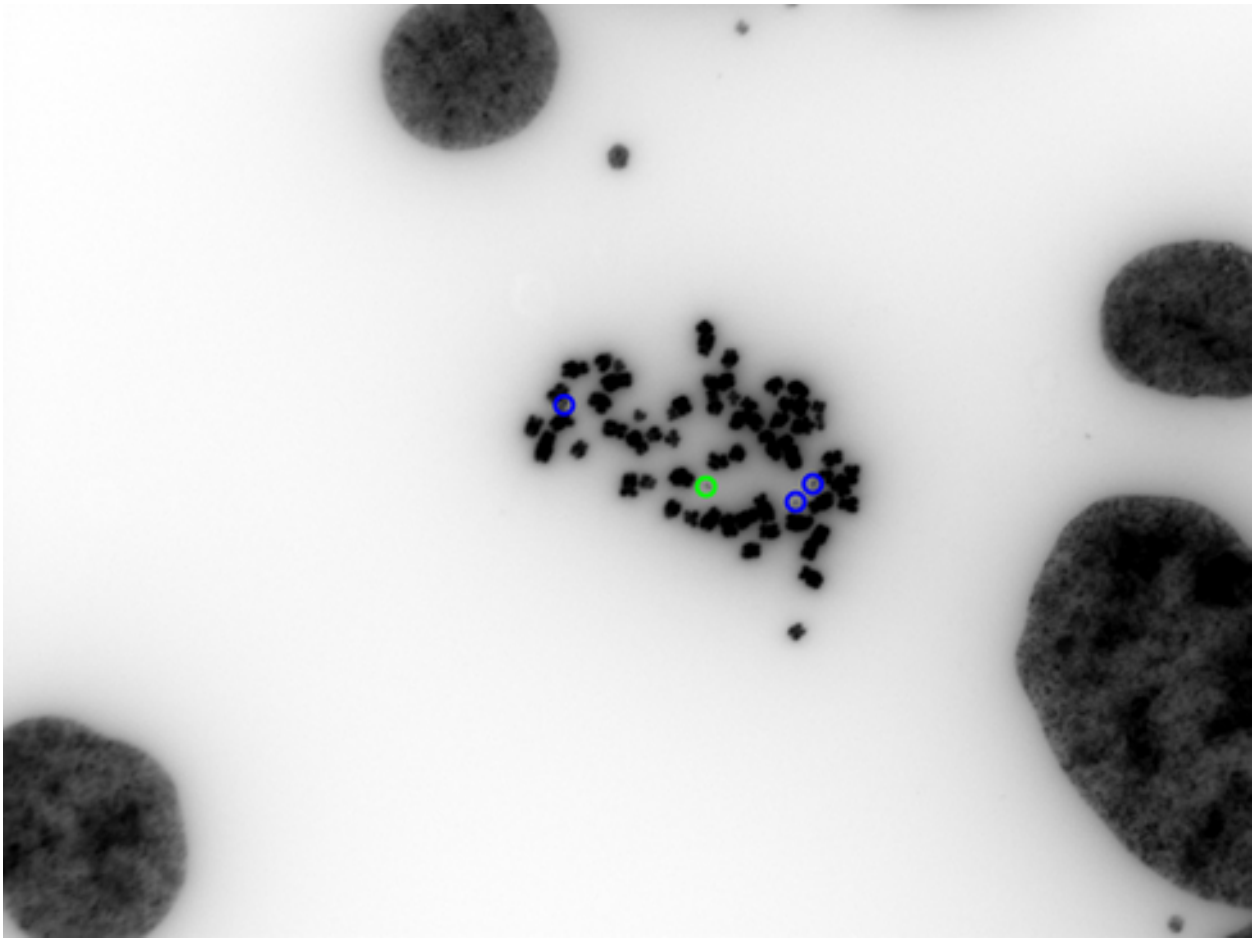


Figure S2.5: RXF623 - 003

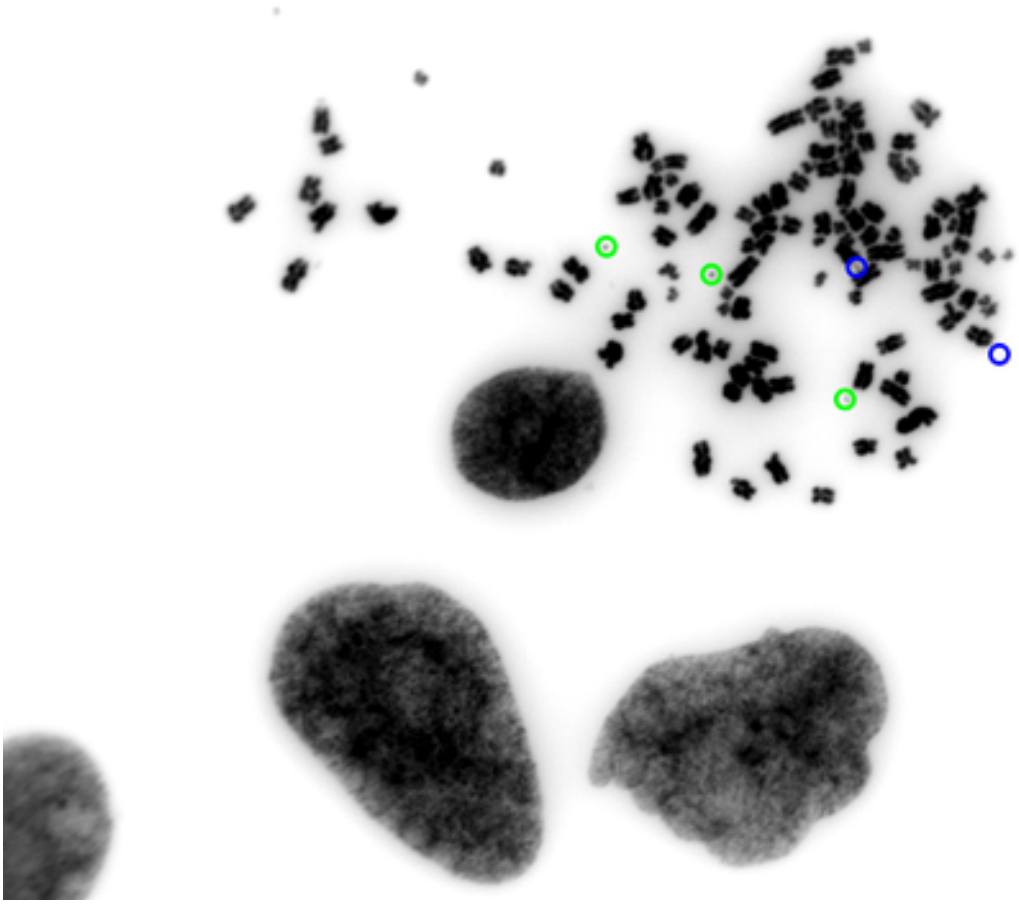


Figure S2.6: OVCAR3 - 013

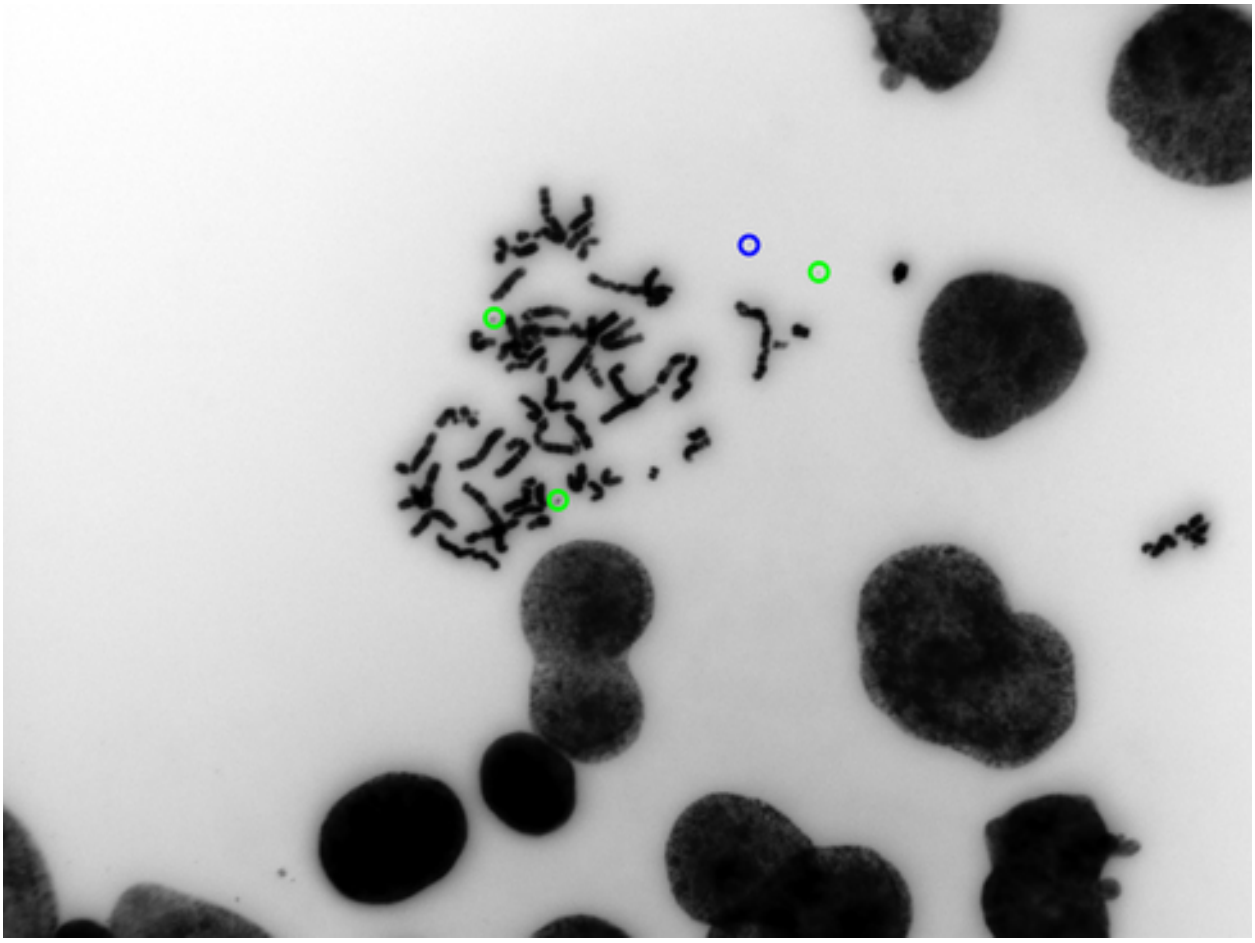


Figure S2.7: H23 - 032

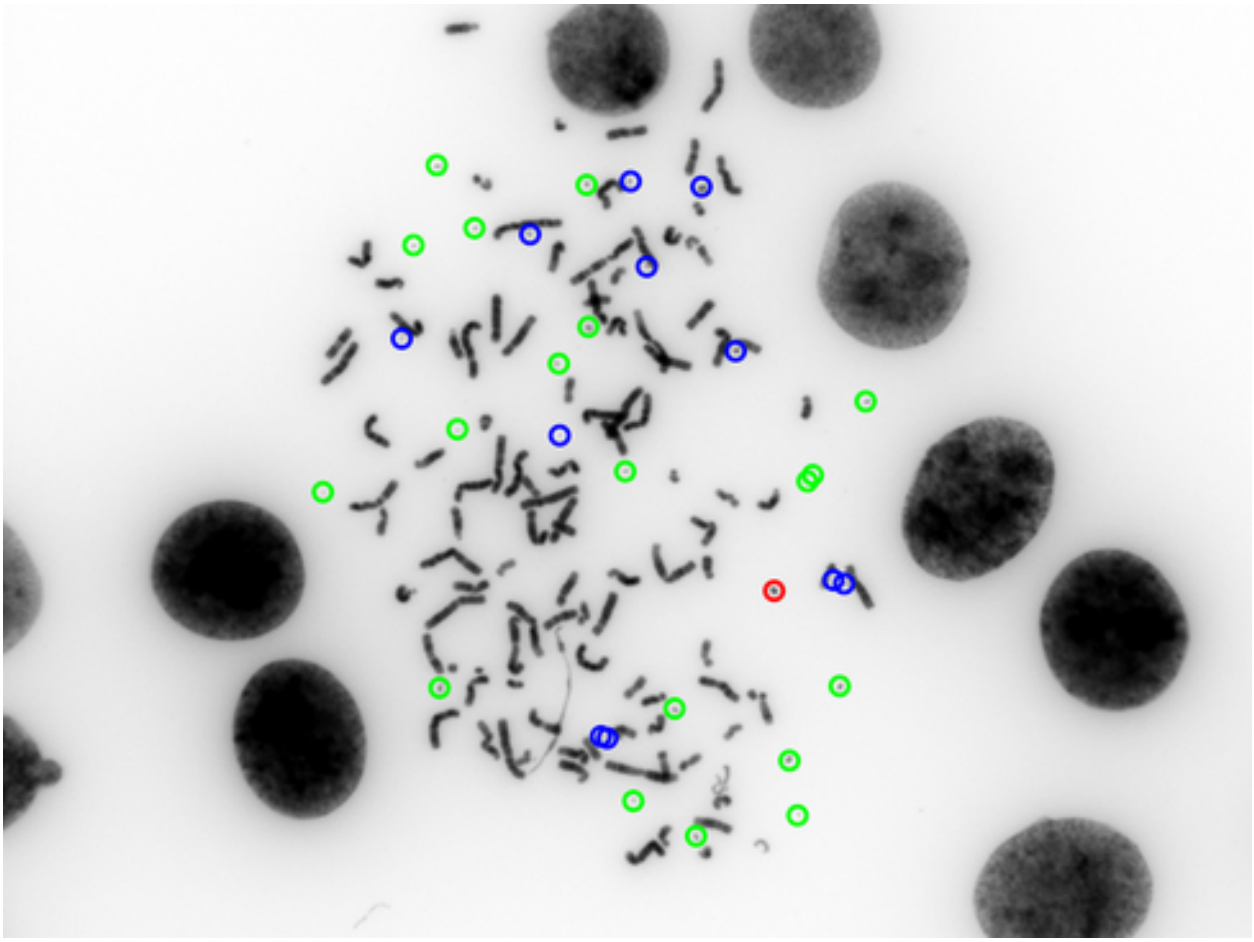


Figure S2.8: M14 - 042

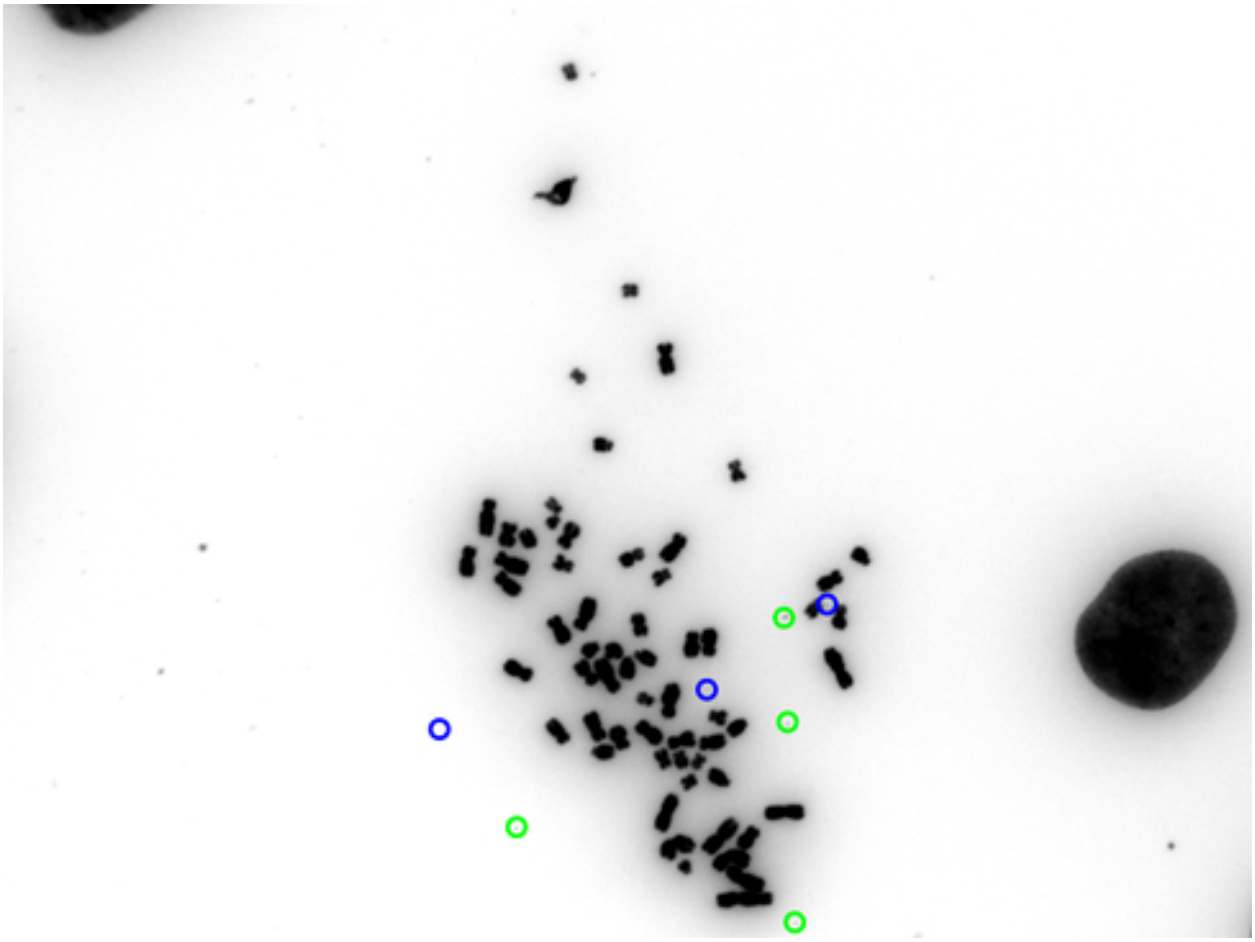


Figure S2.9: A549 - 029

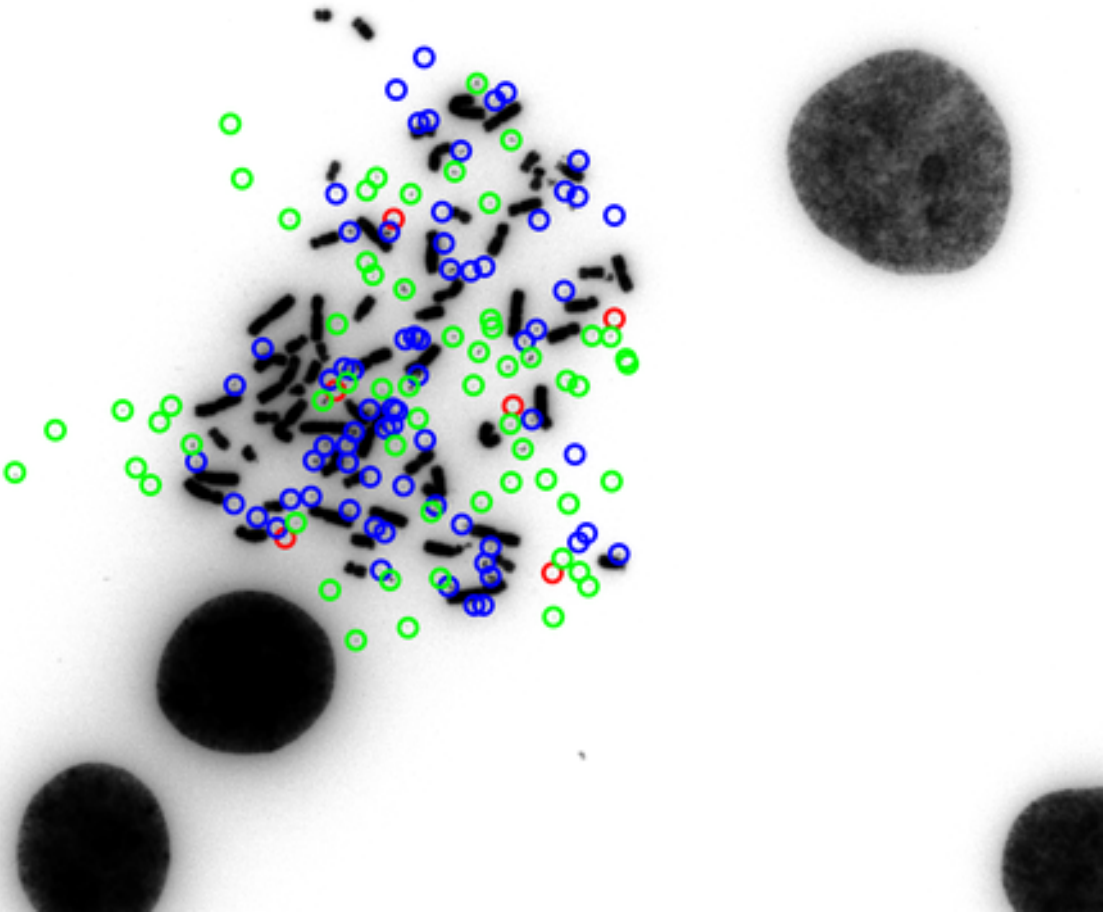


Figure S2.10: M14 - 004

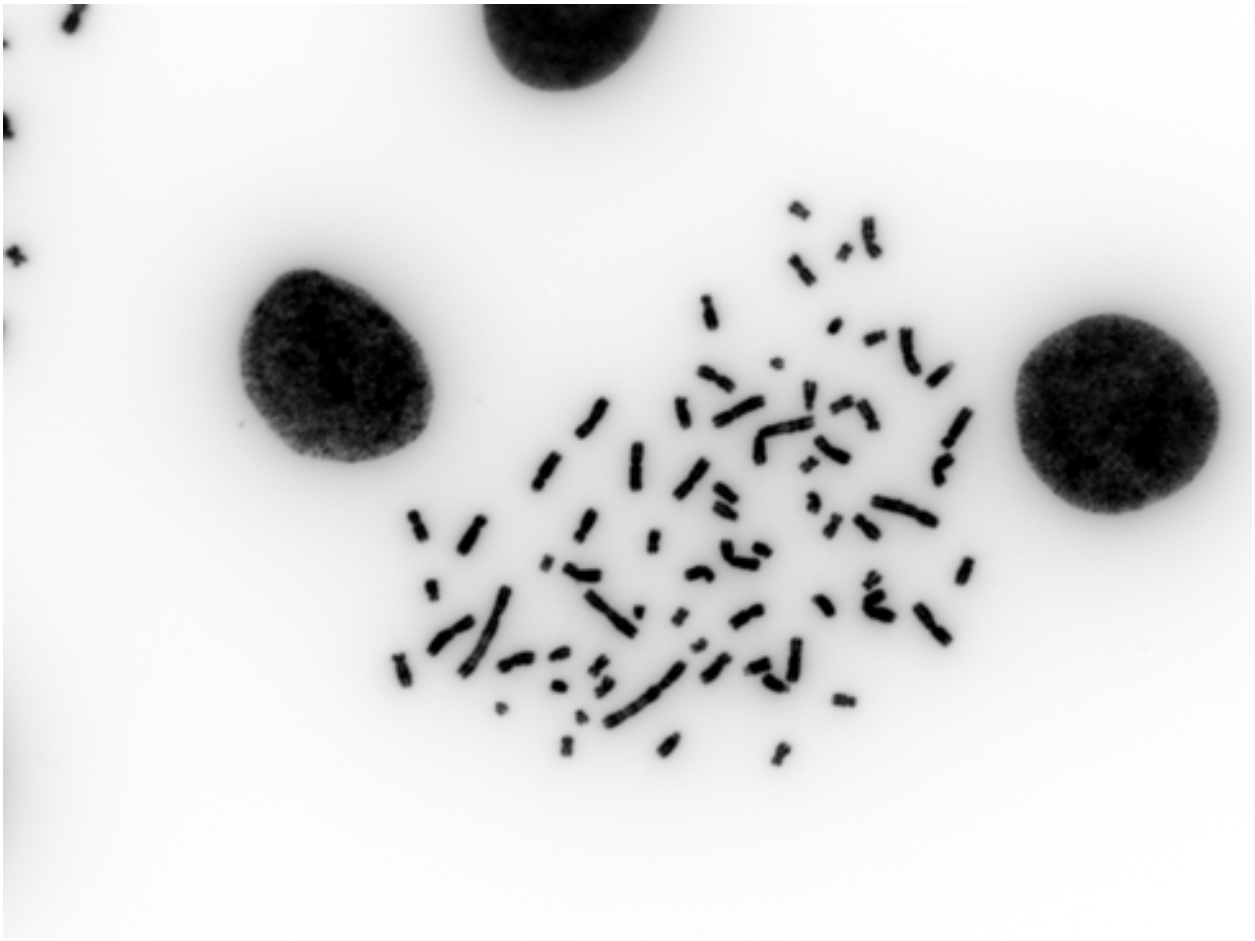


Figure S2.11: TK10 - 030

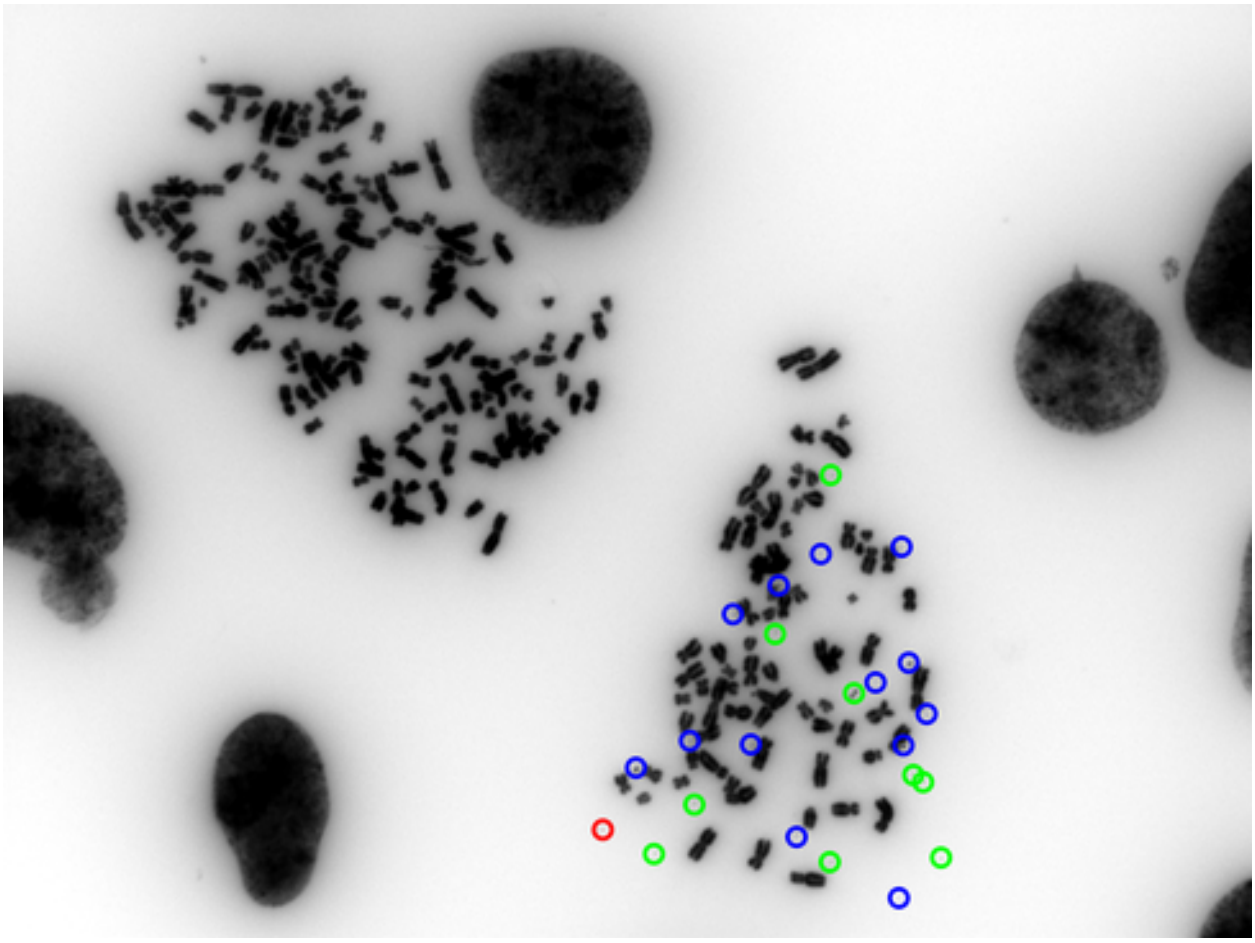


Figure S2.12: SF295 - 002

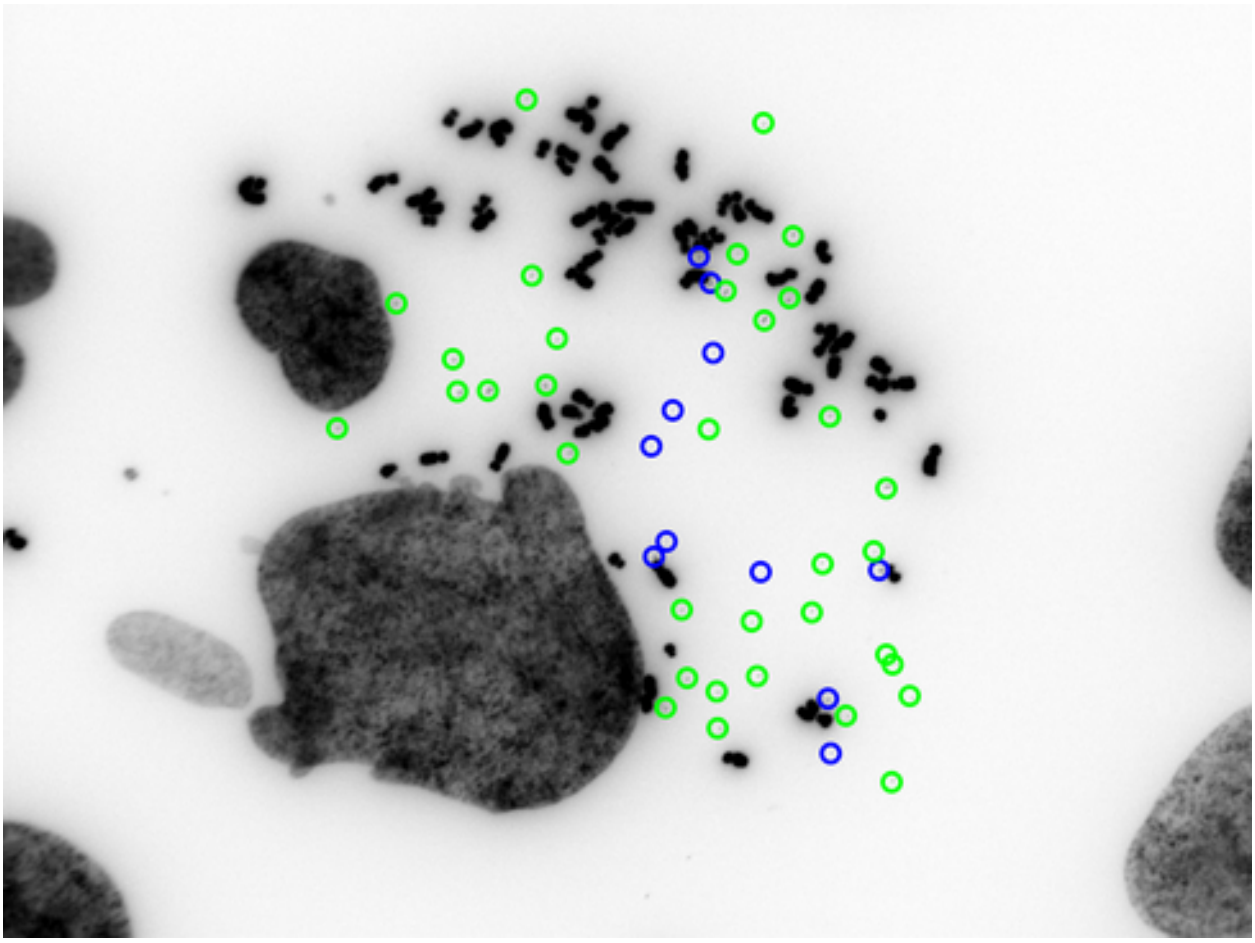


Figure S2.13: CAKI1 - 005

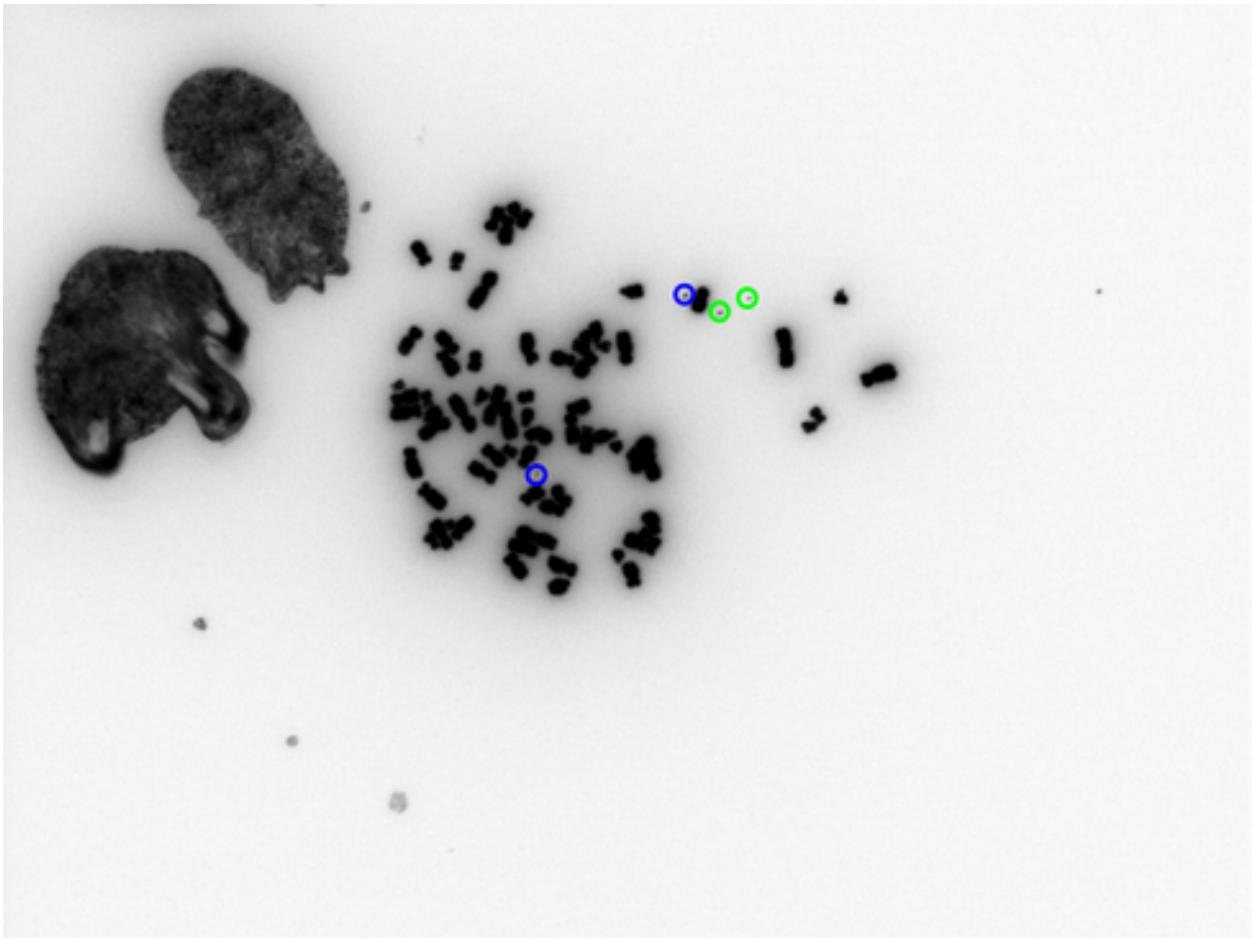


Figure S2.14: CAKI1 - 004

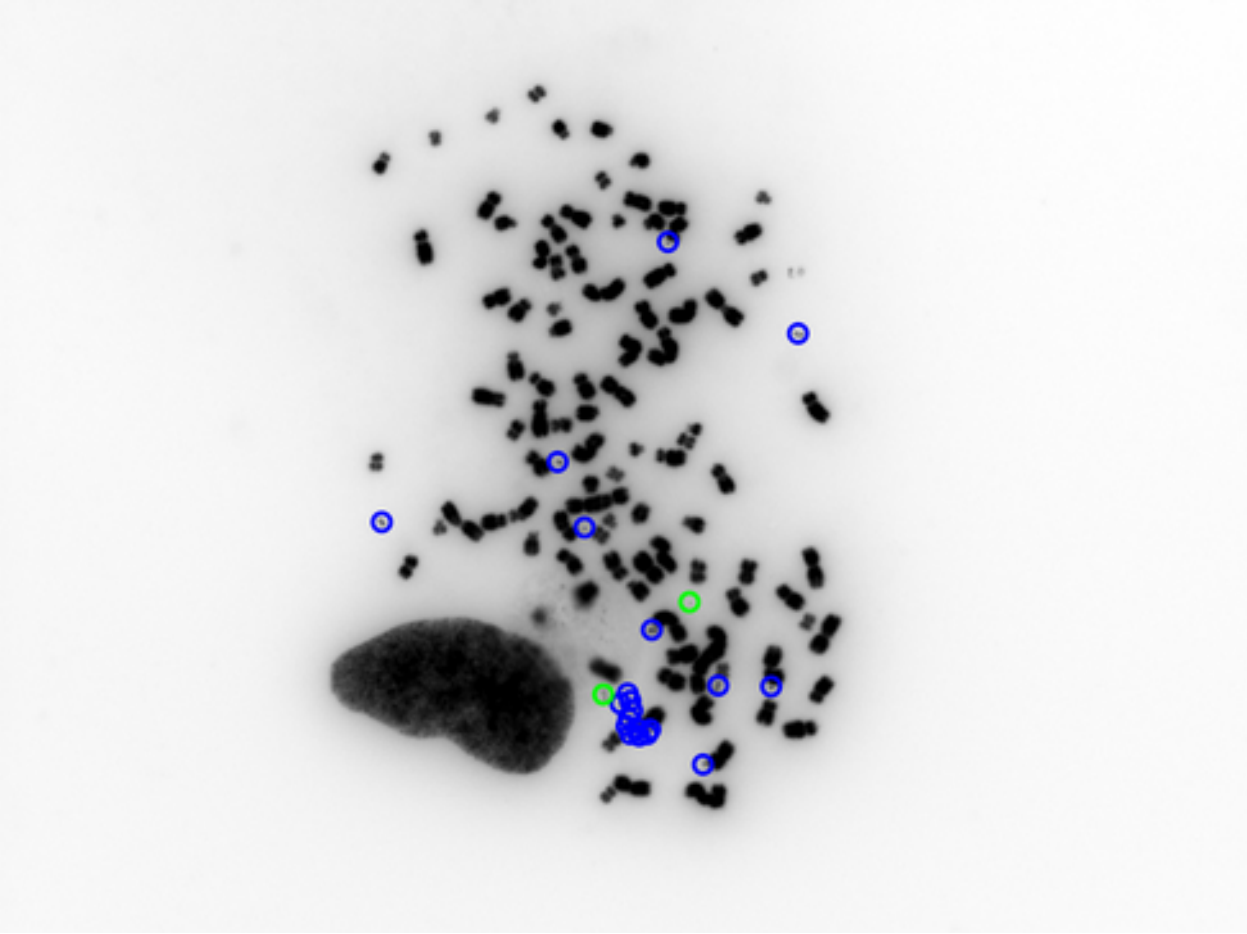


Figure S2.15: Hs578T - 009

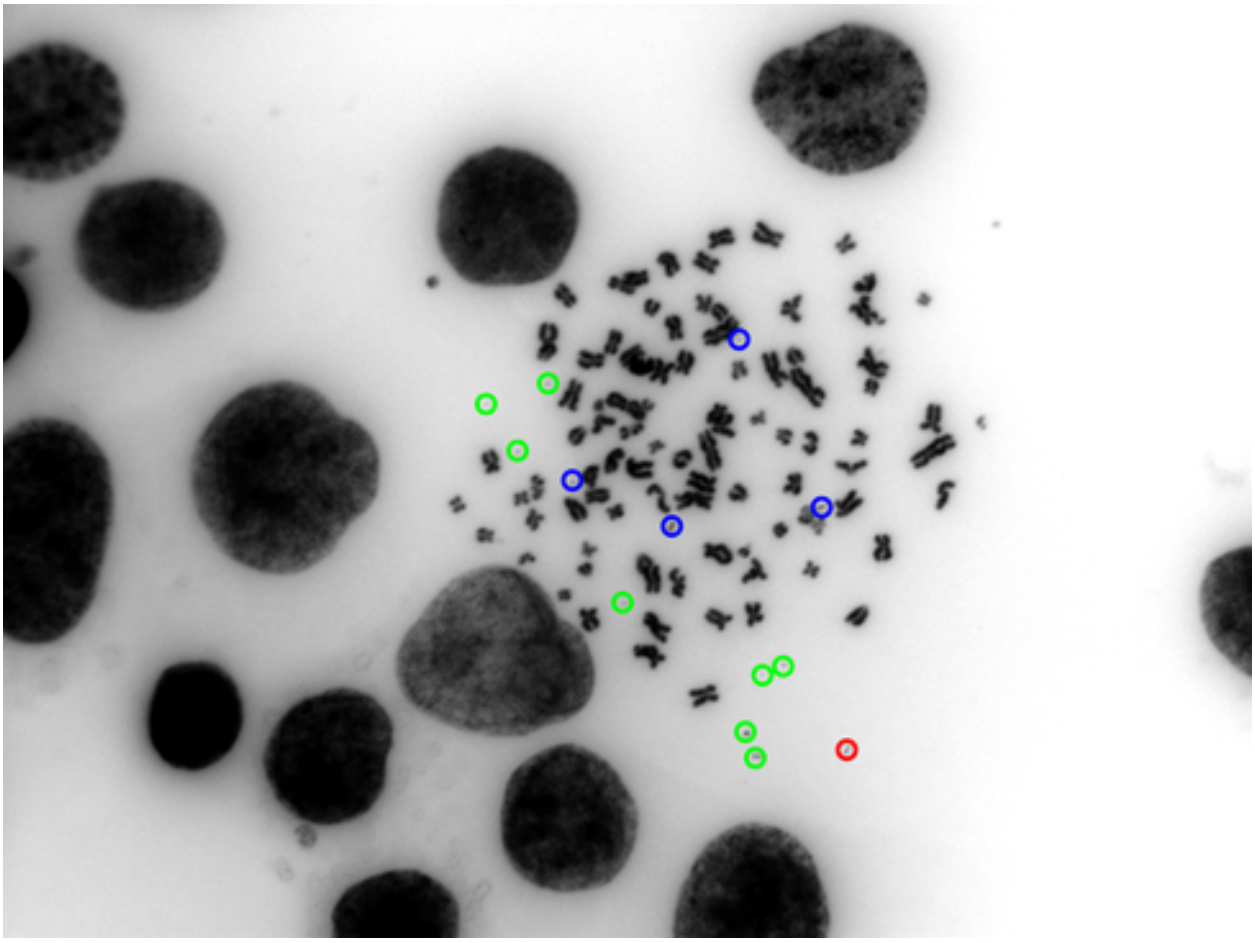


Figure S2.16: IGROV1 - 036

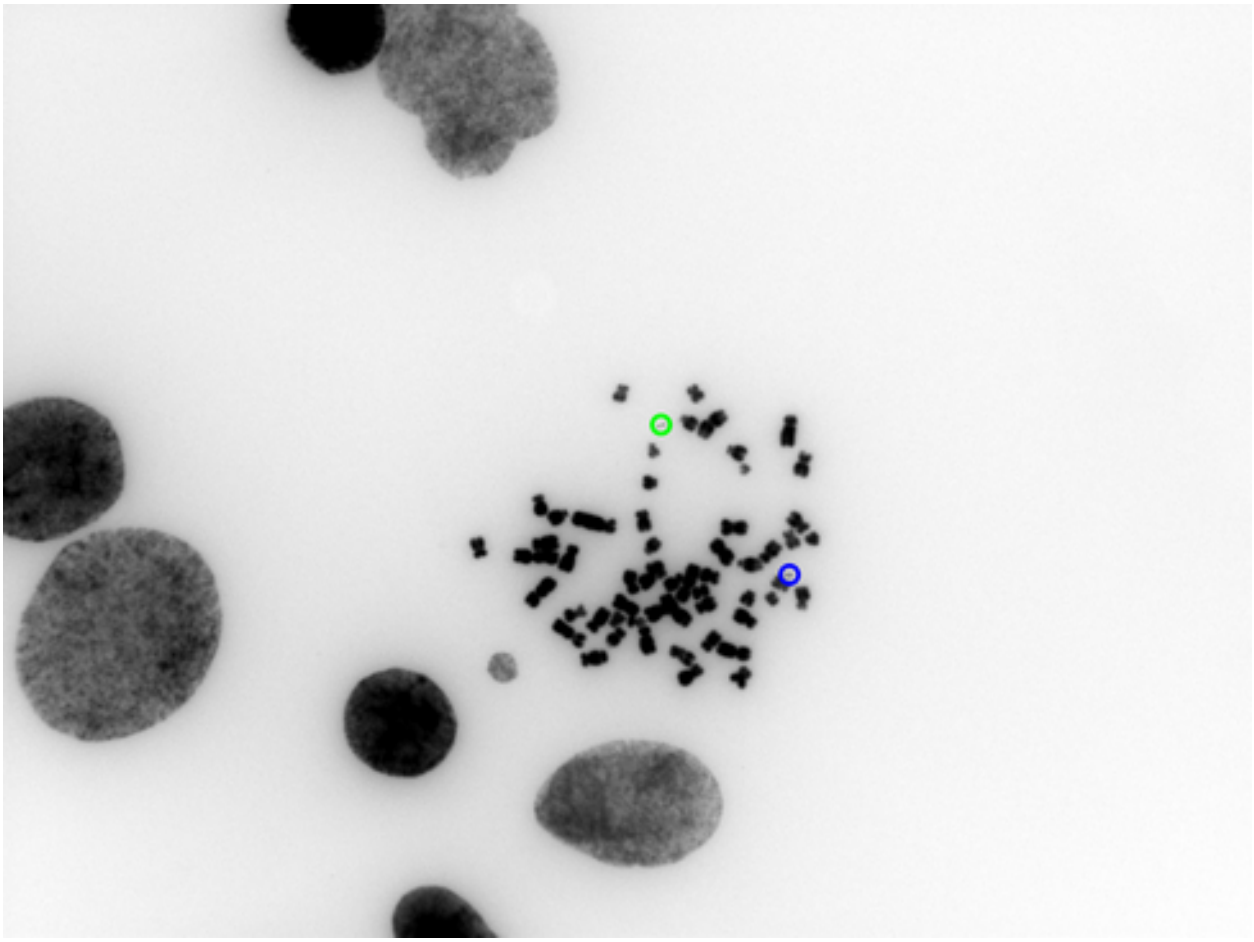


Figure S2.17: H23 - 037

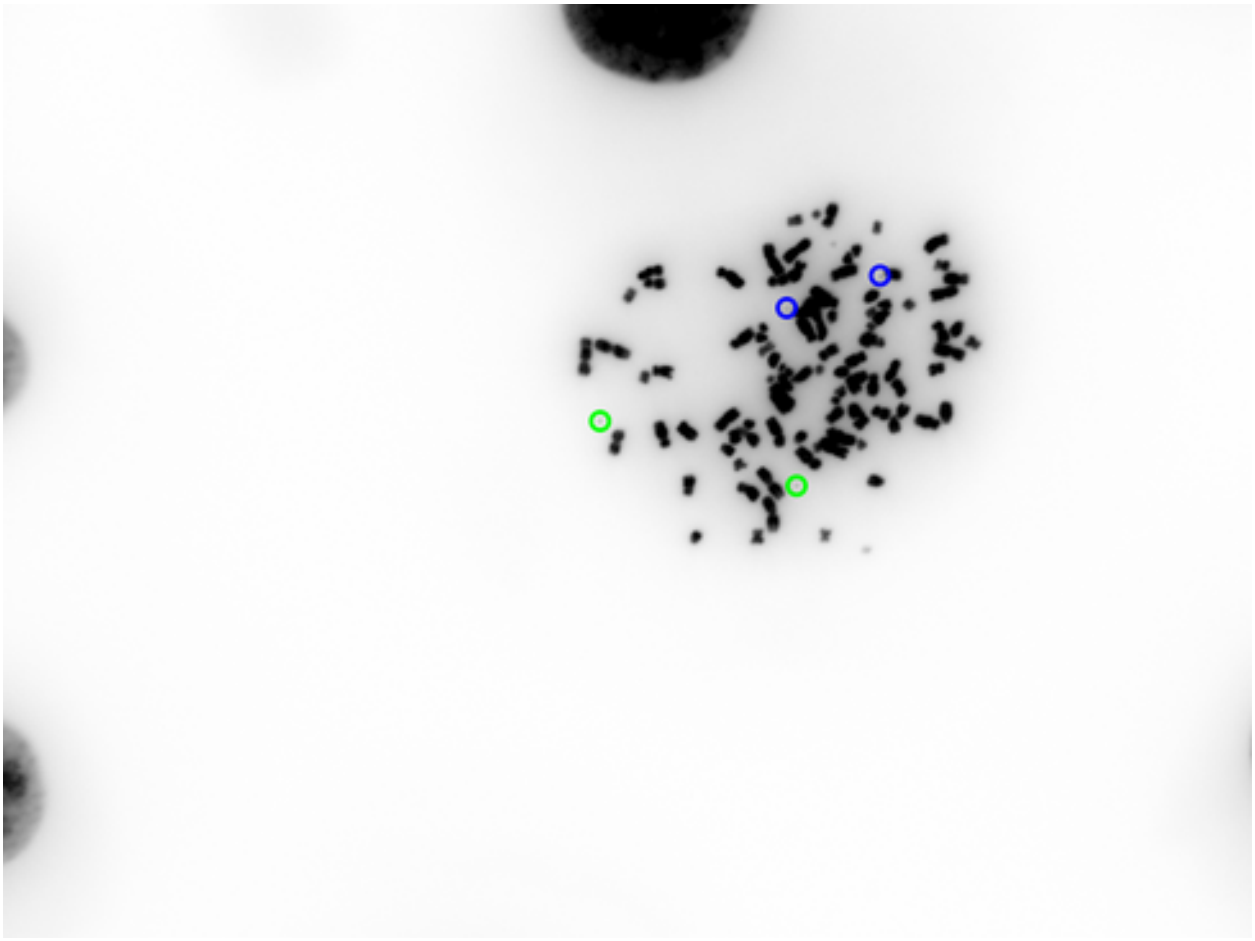


Figure S2.18: U251 - 041

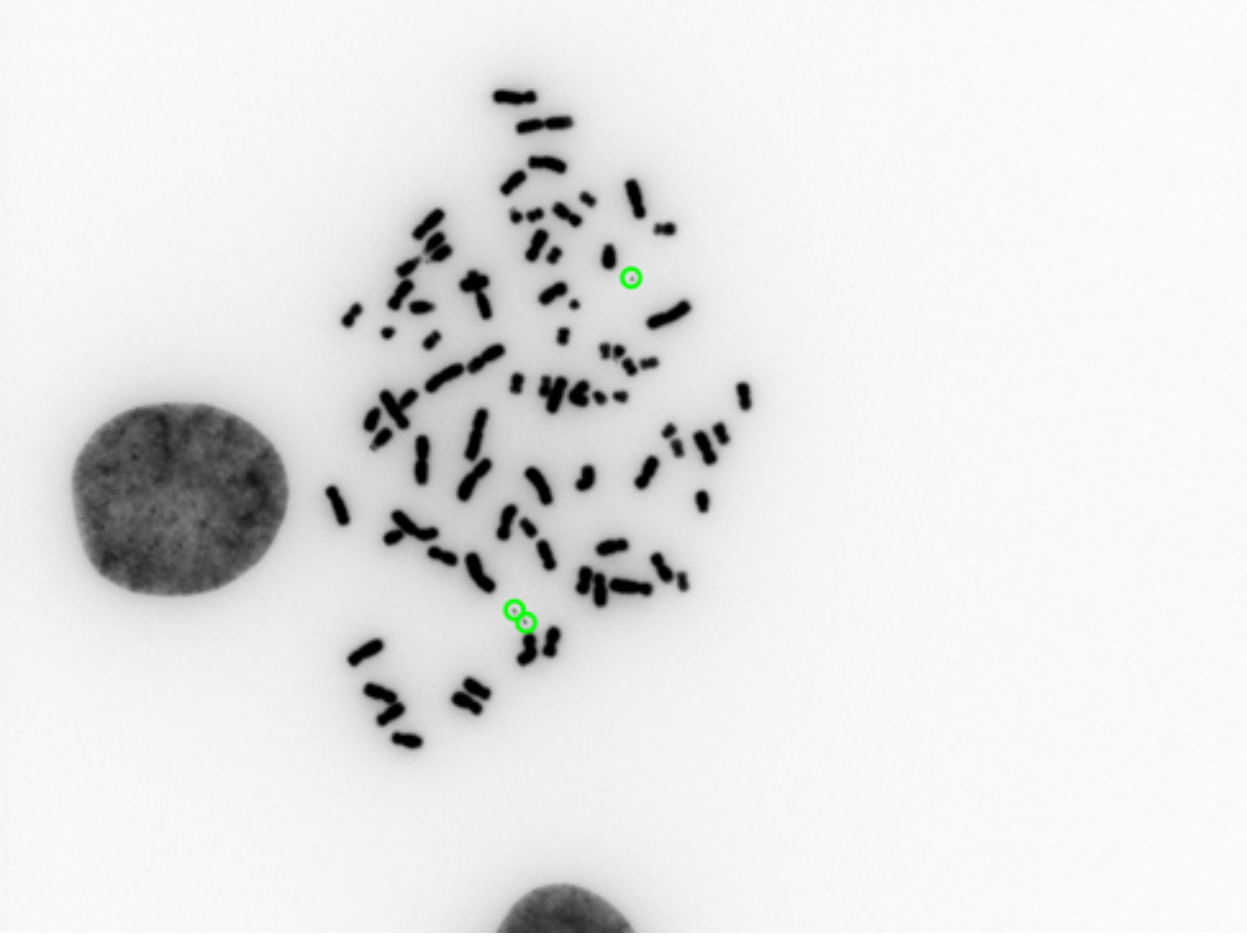


Figure S2.19: UACC62 - 001

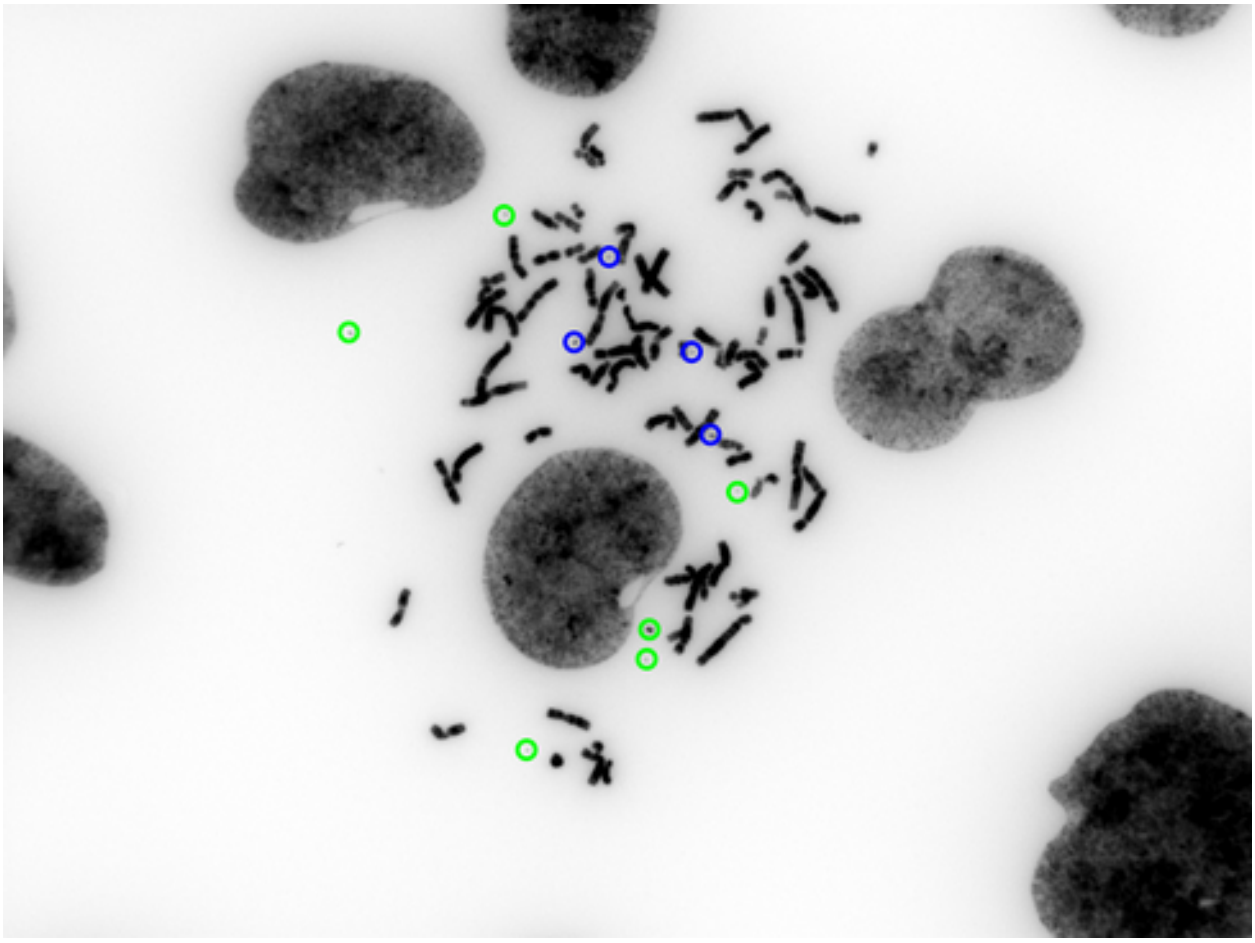


Figure S2.20: 786-0 - 037

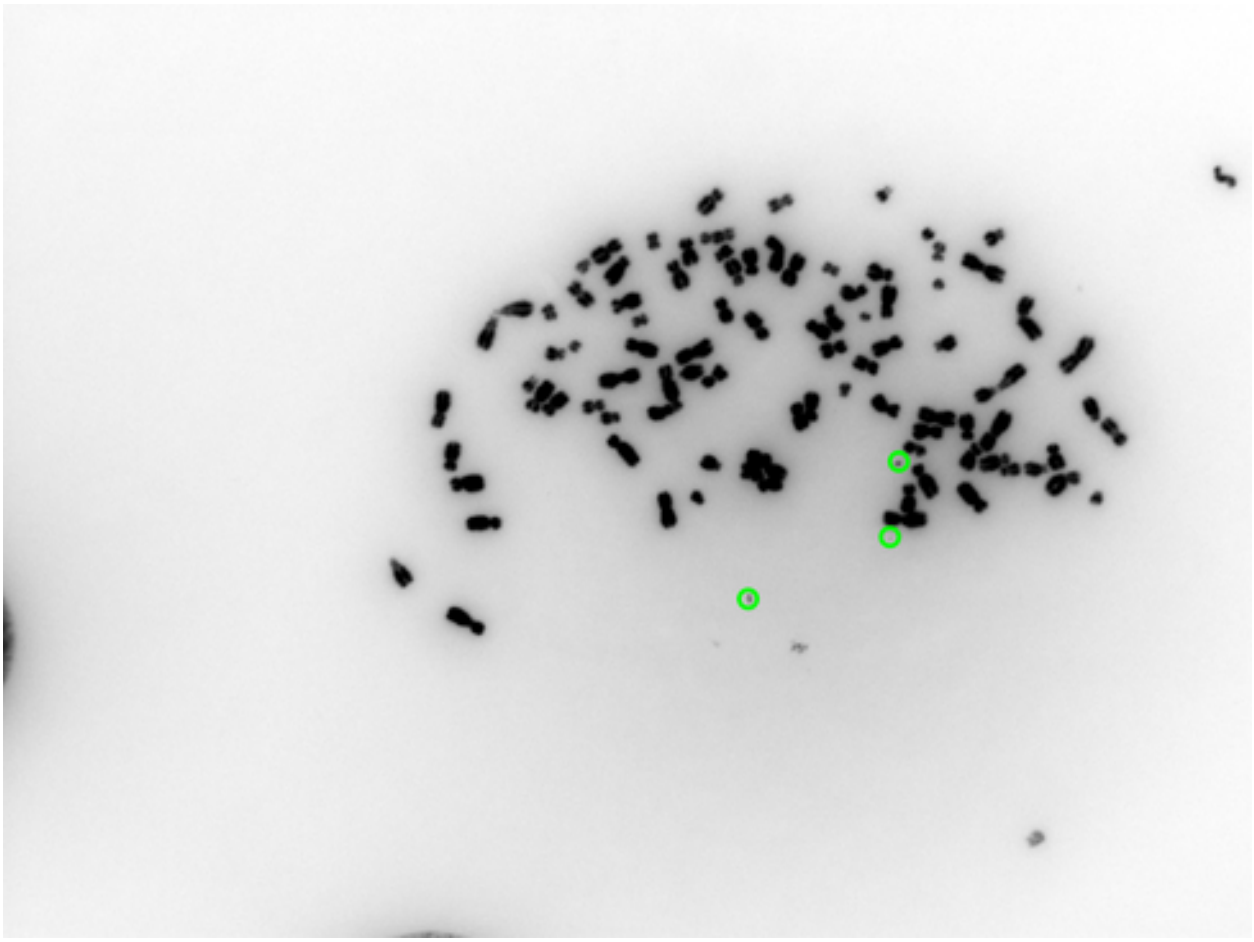


Figure S2.21: SkMel2 - 24

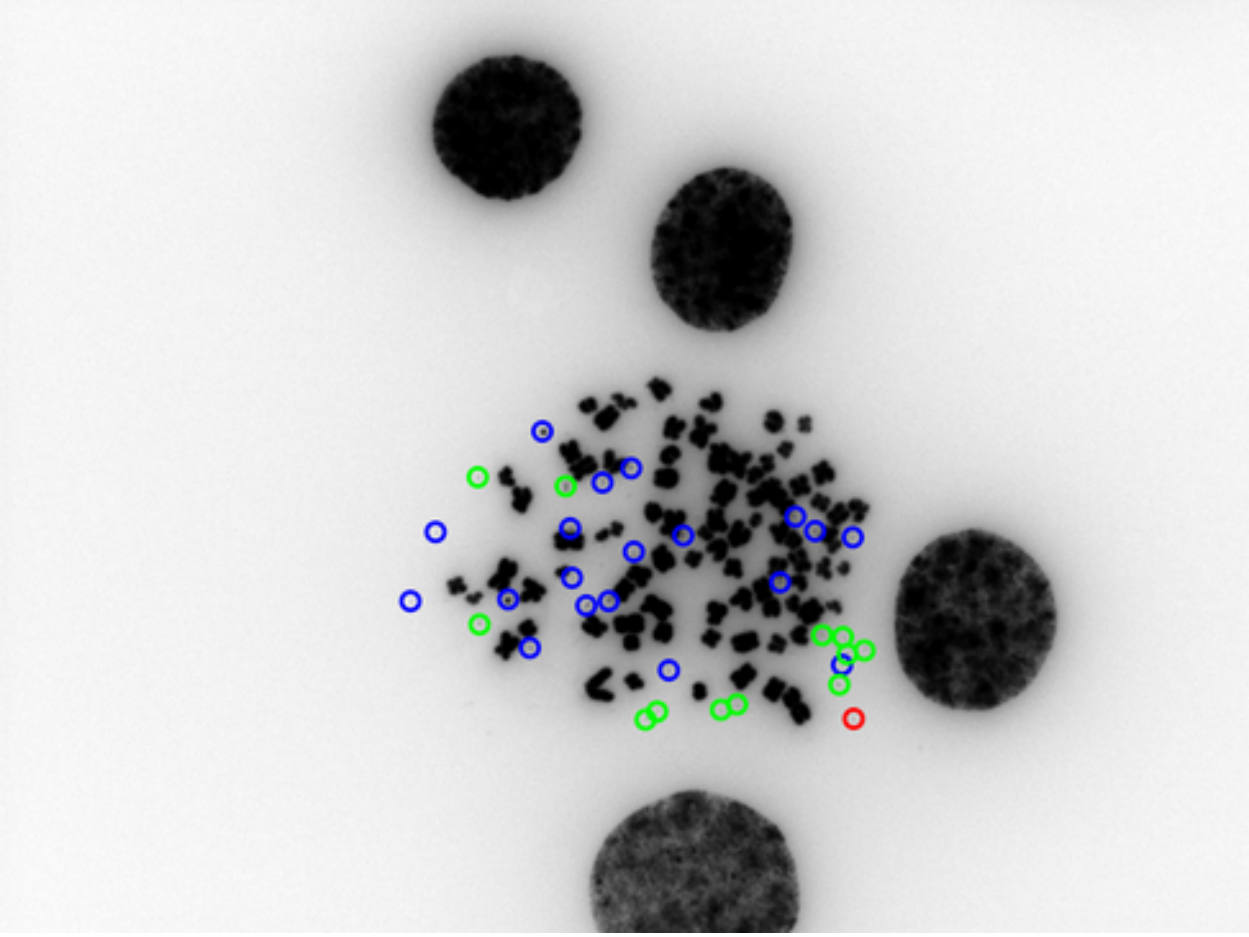


Figure S2.22: SKOV3 - 019

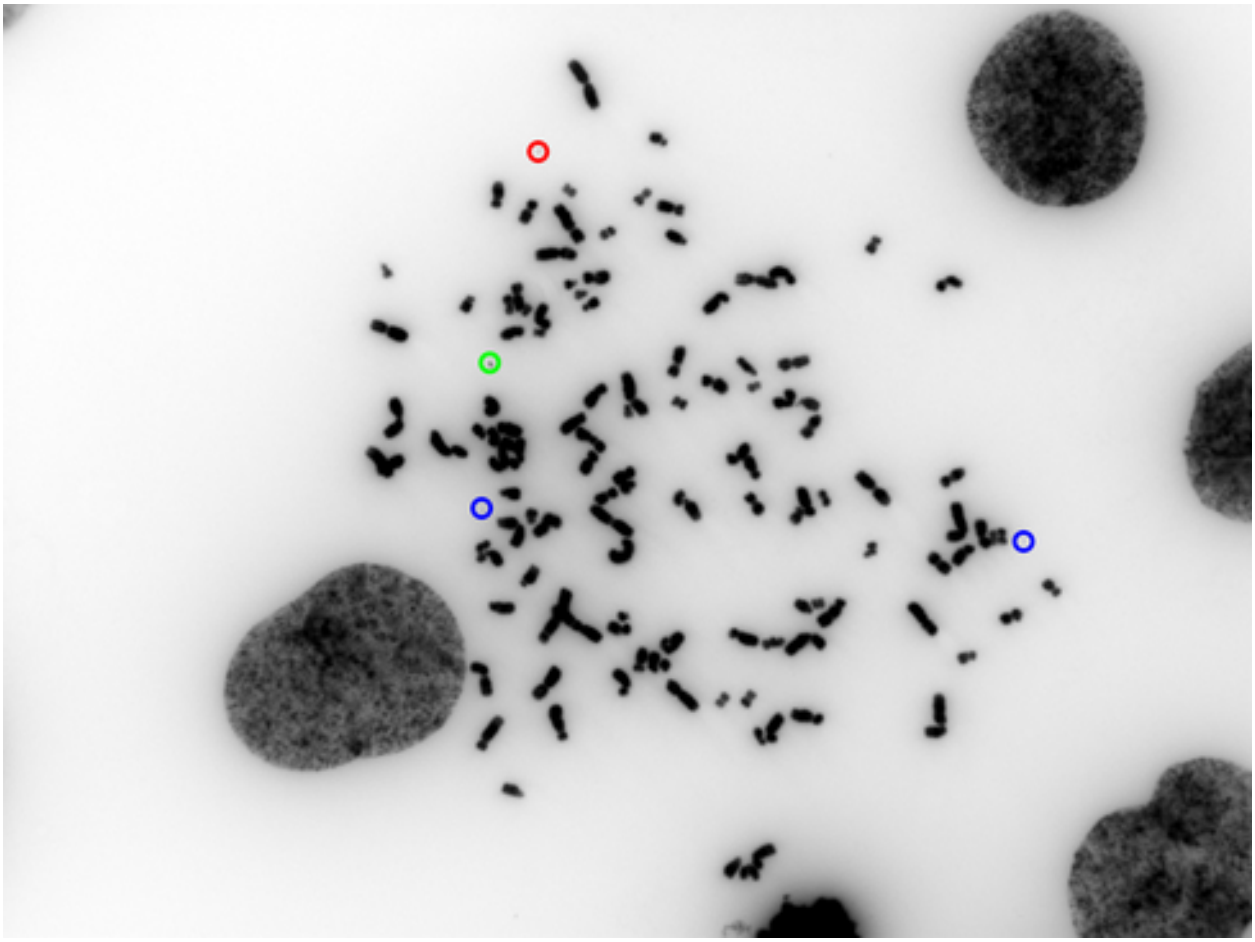


Figure S2.23: RXF623 - 001

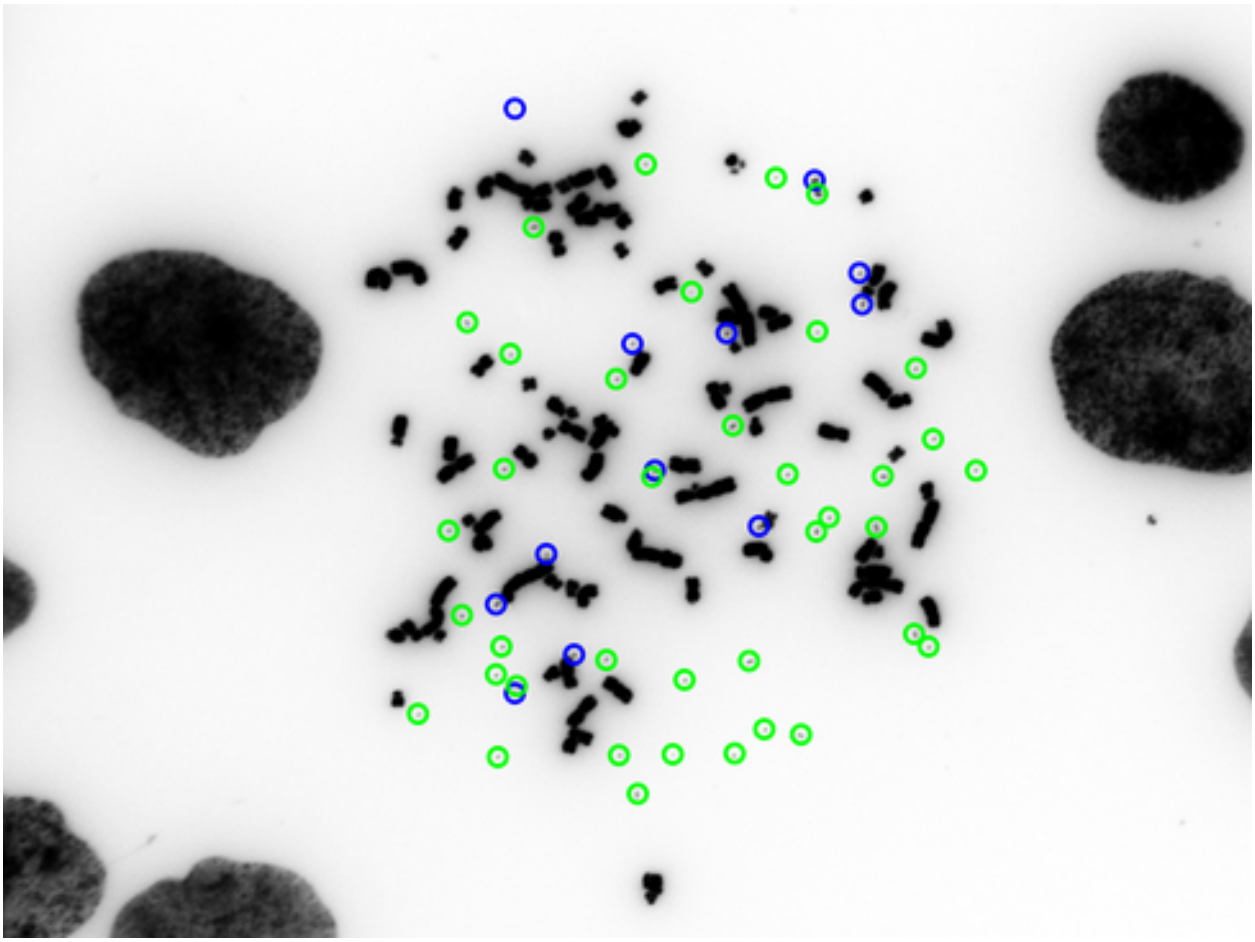


Figure S2.24: BT549 - 031

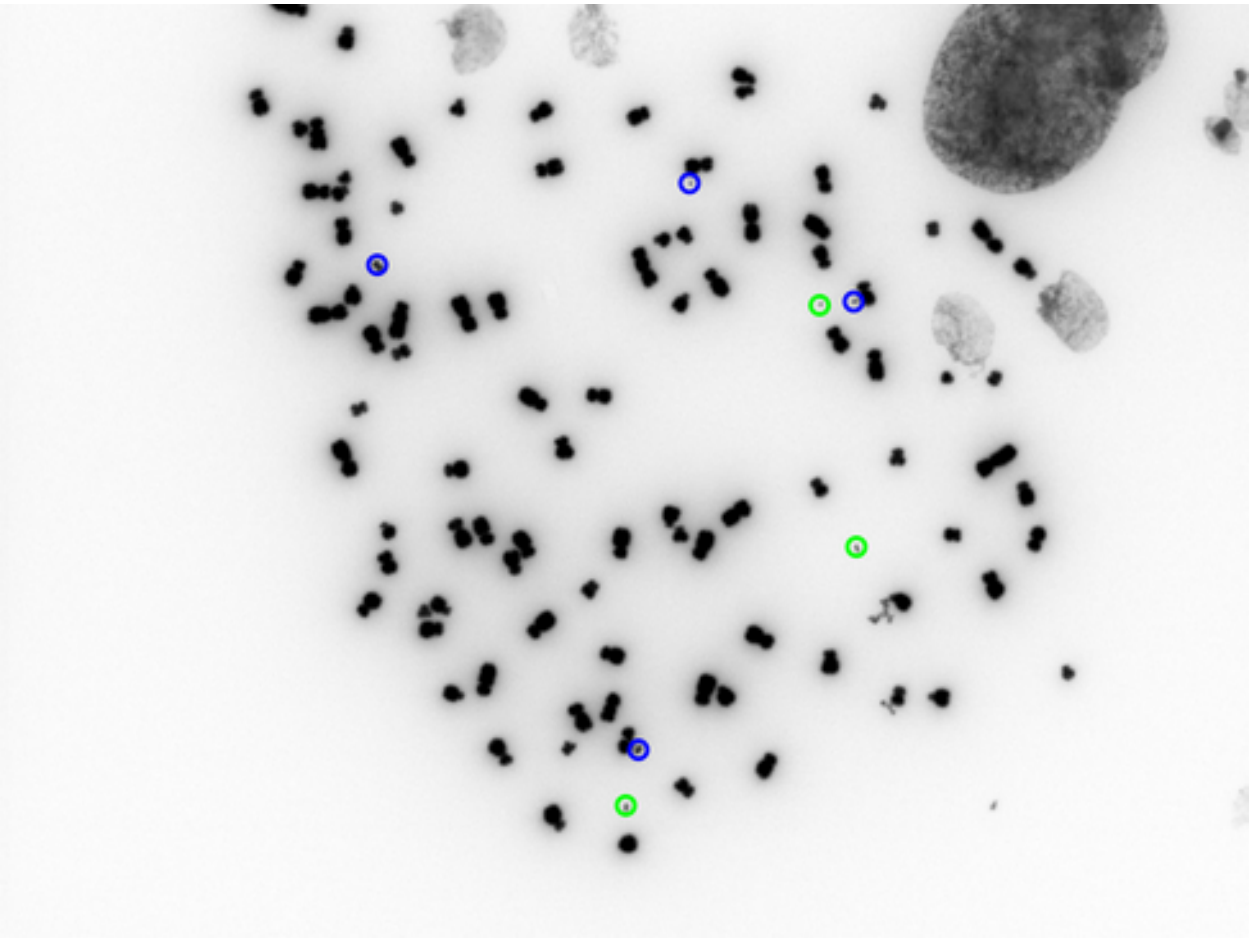


Figure S2.25: CAKI1 - 014

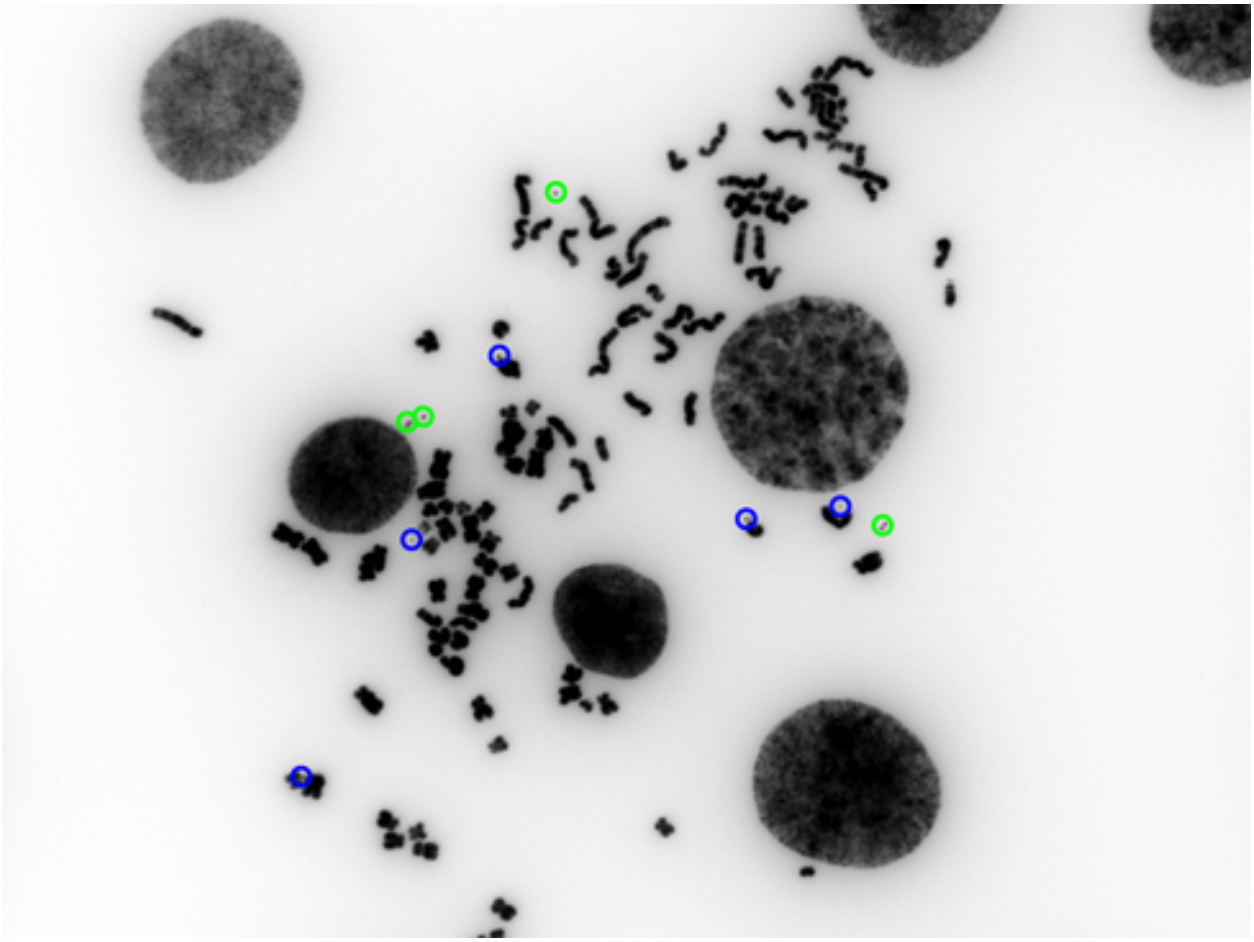


Figure S2.26: H322M - 023

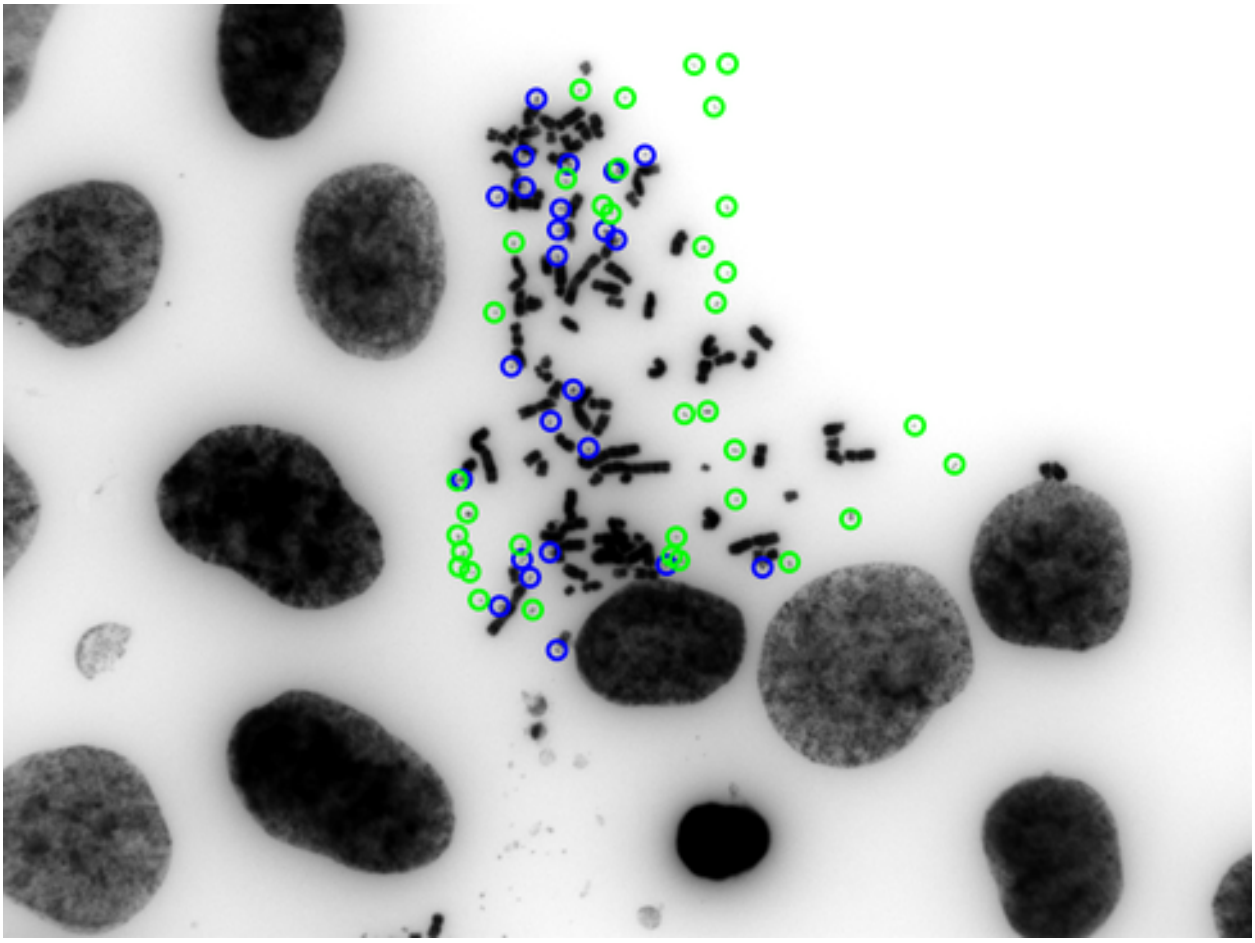


Figure S2.27: PC3 - 006

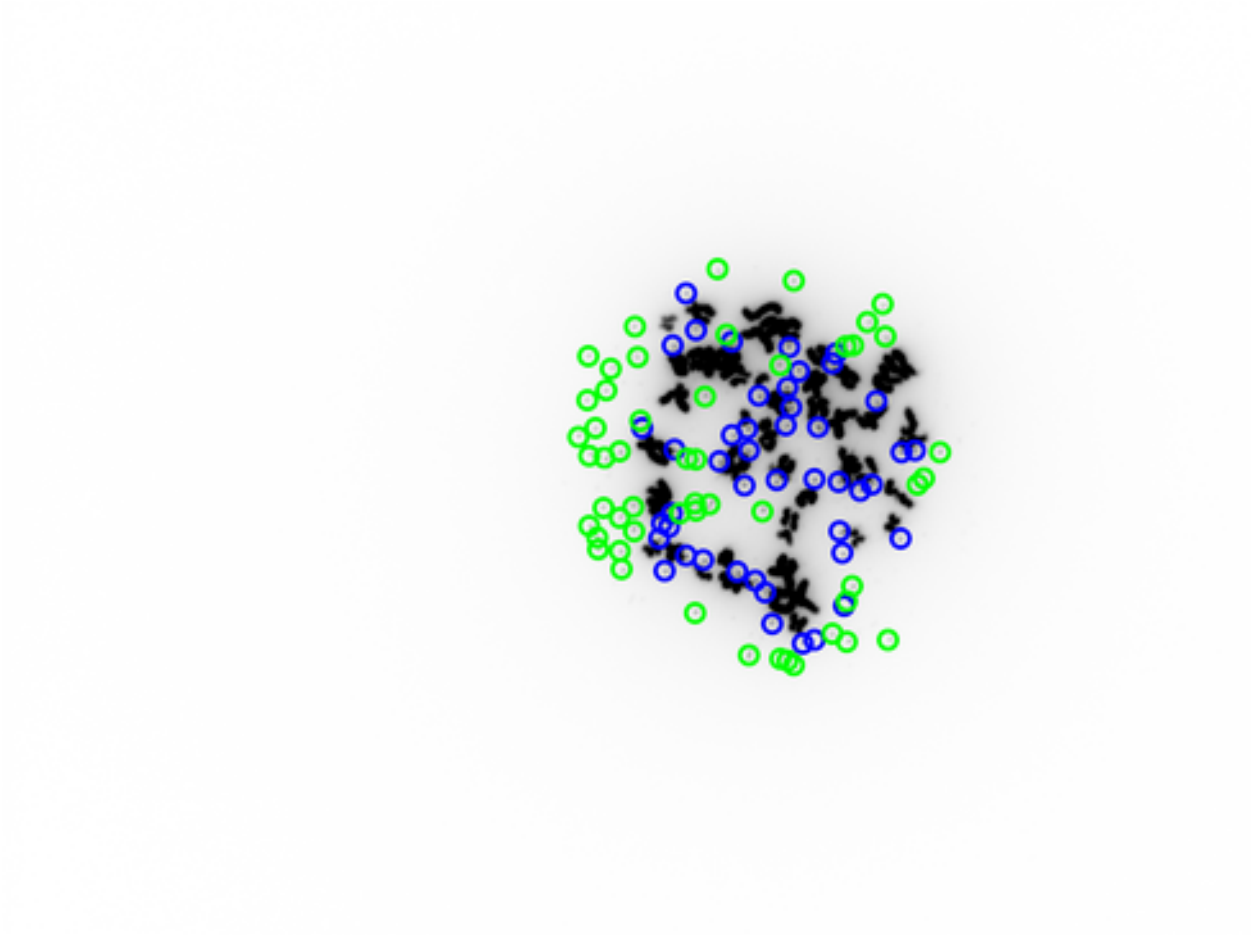


Figure S2.28: HK301 - 016

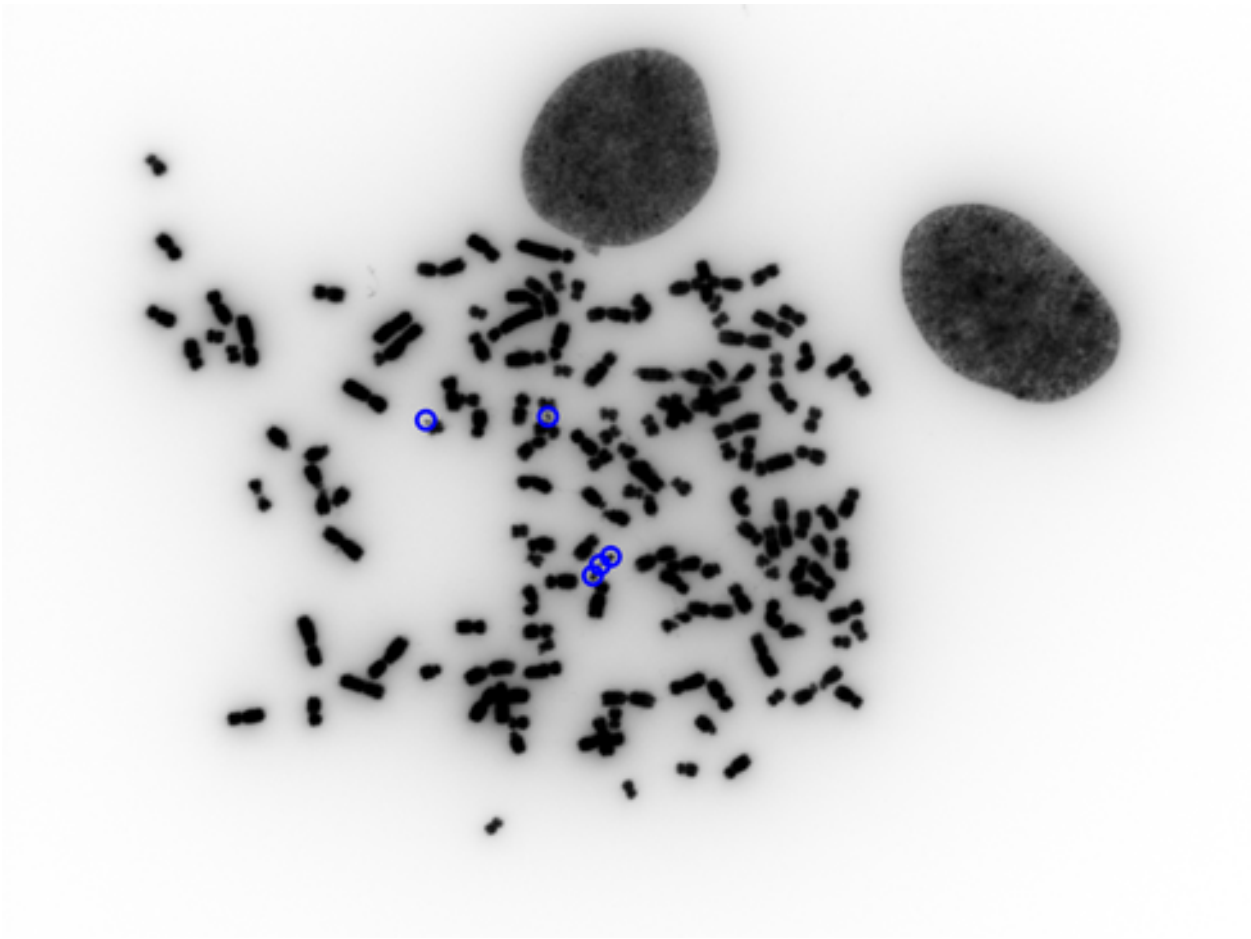


Figure S2.29: UACC62 - 022

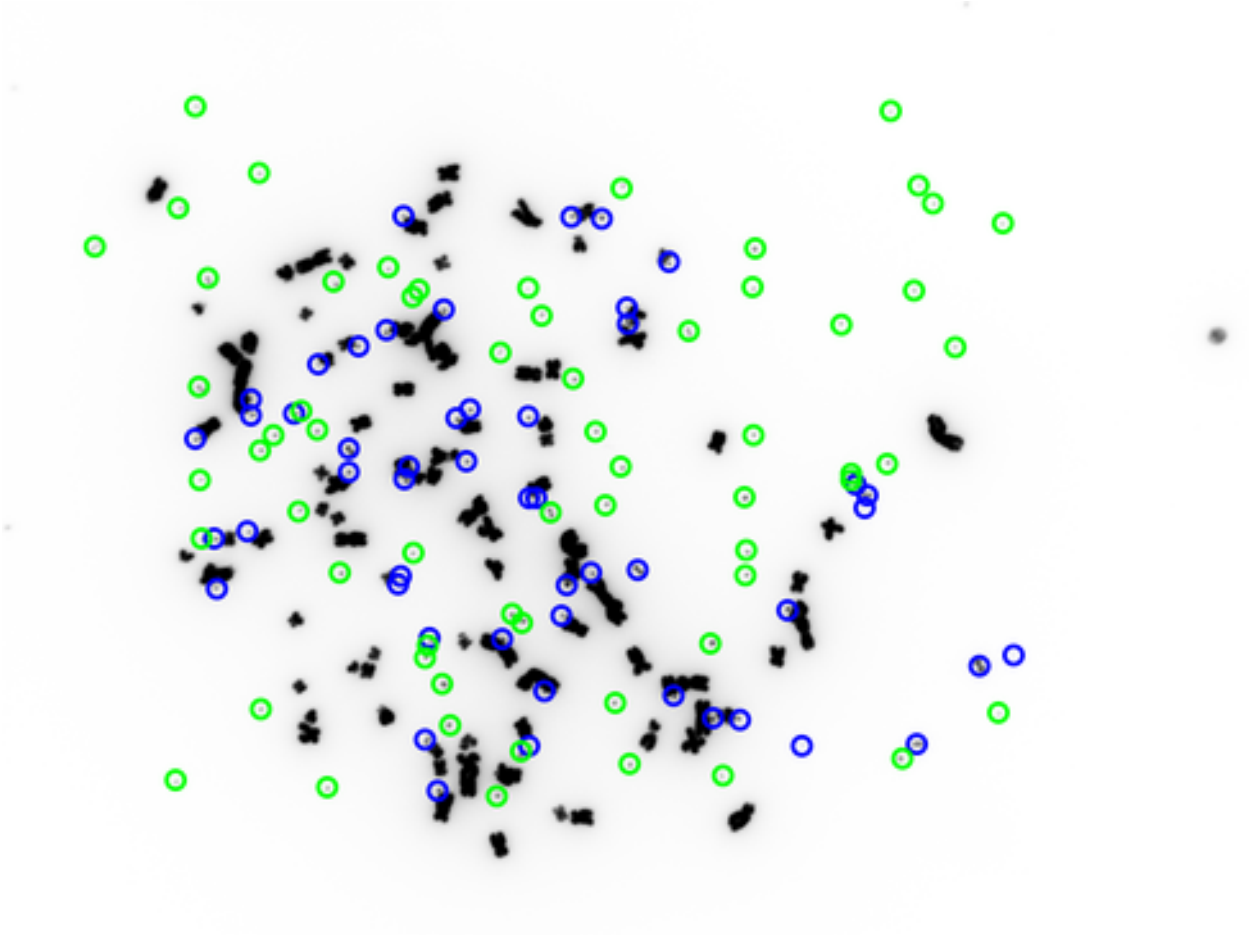


Figure S2.30: BT549 - 053

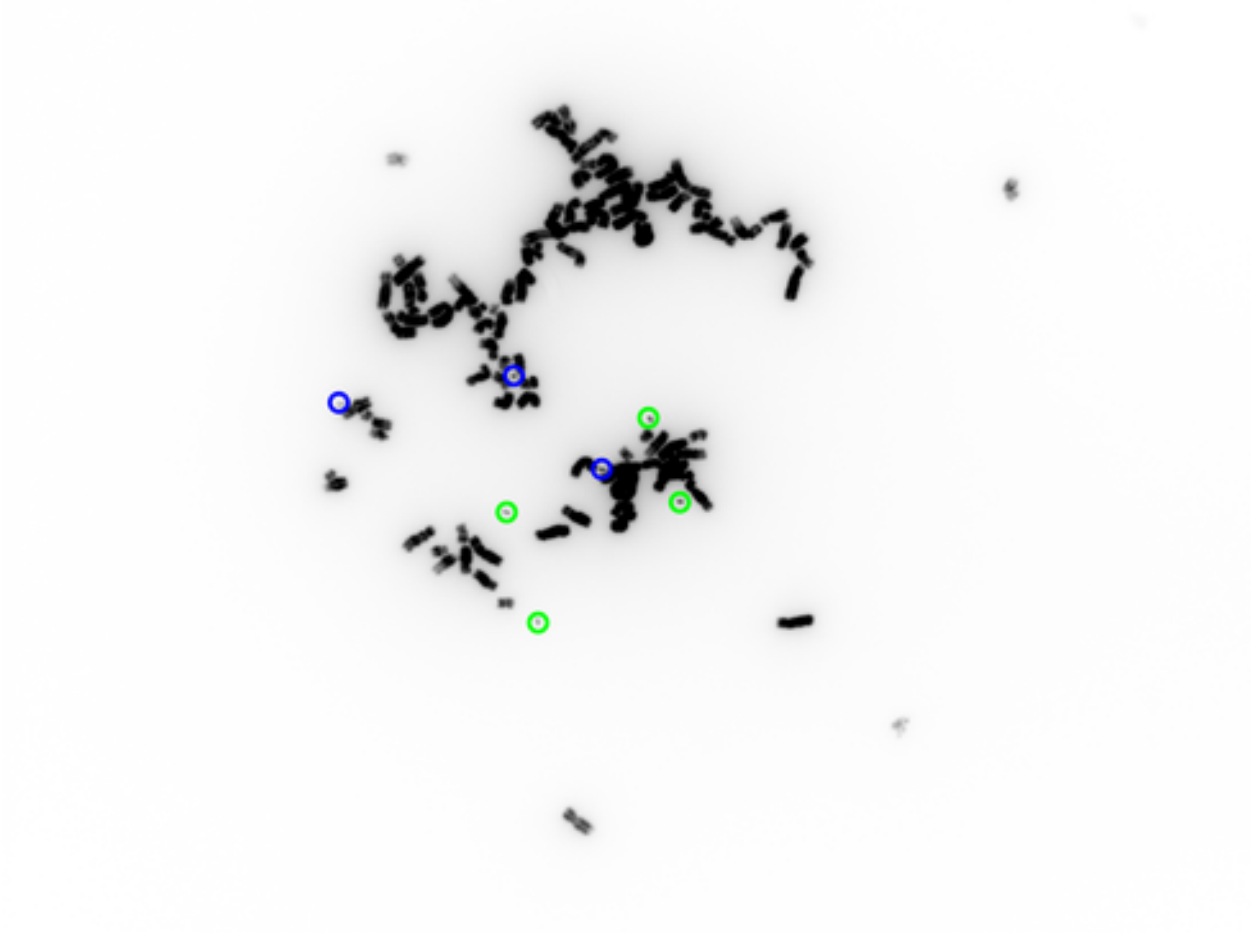


Figure S2.31: HOP62 - 038

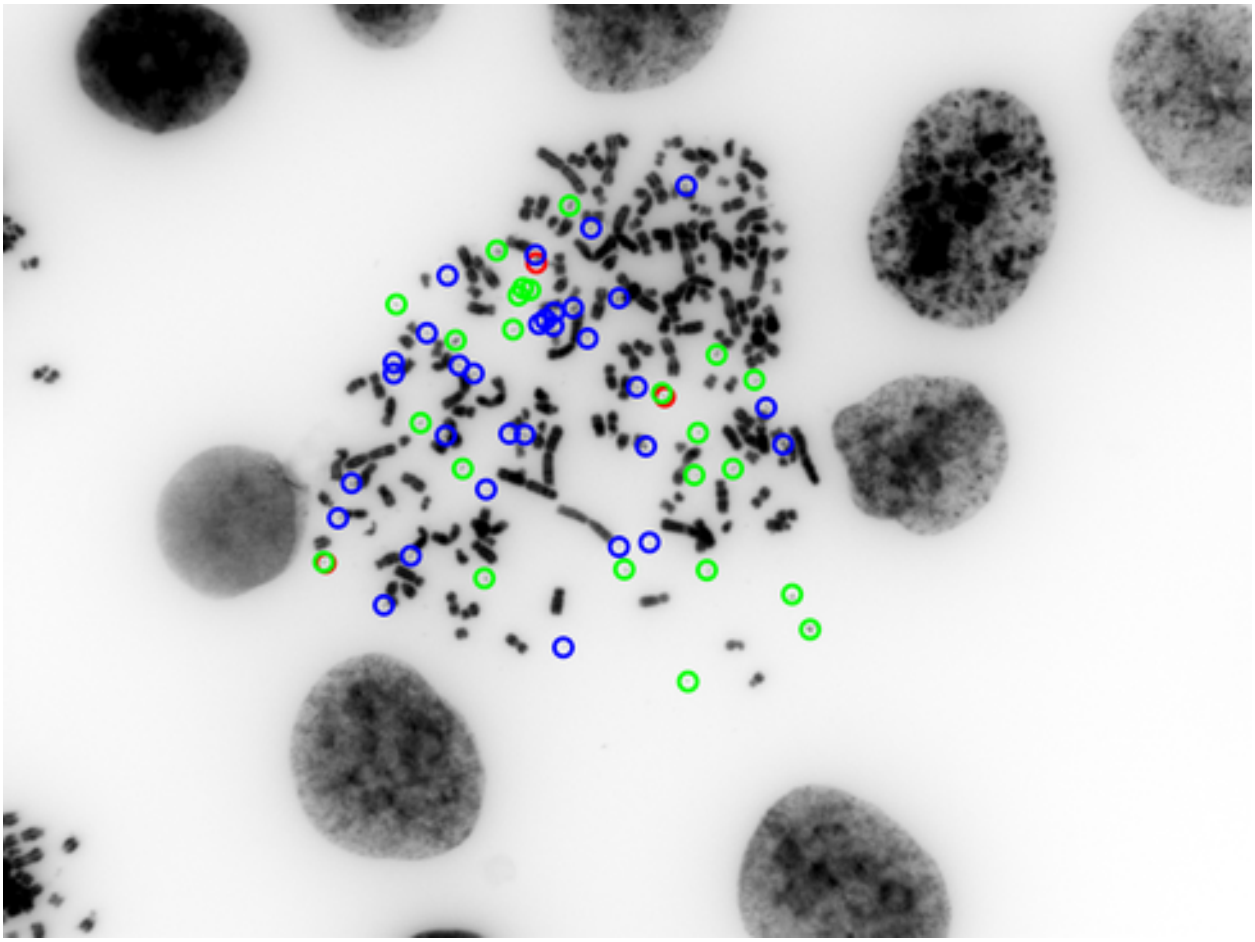


Figure S2.32: PC3 - 003

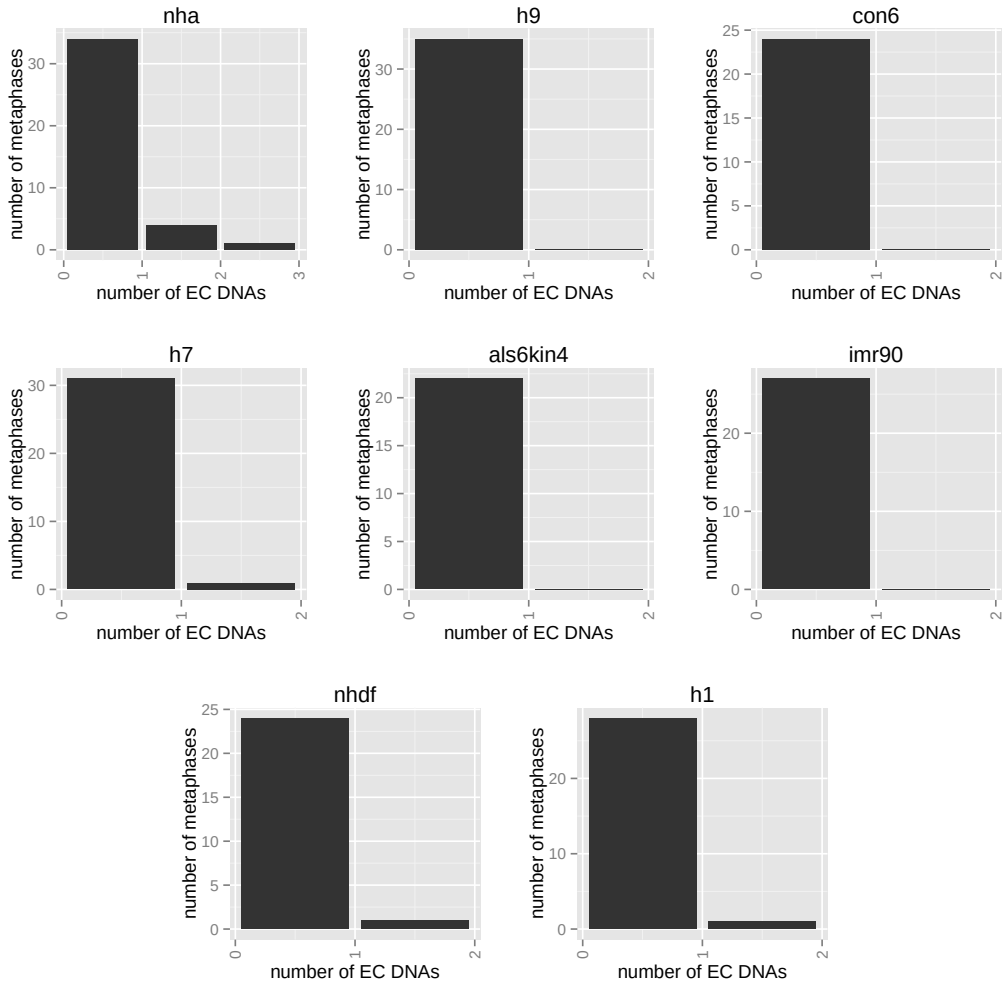


Figure S2.33: ECDNA count histograms of normal samples.

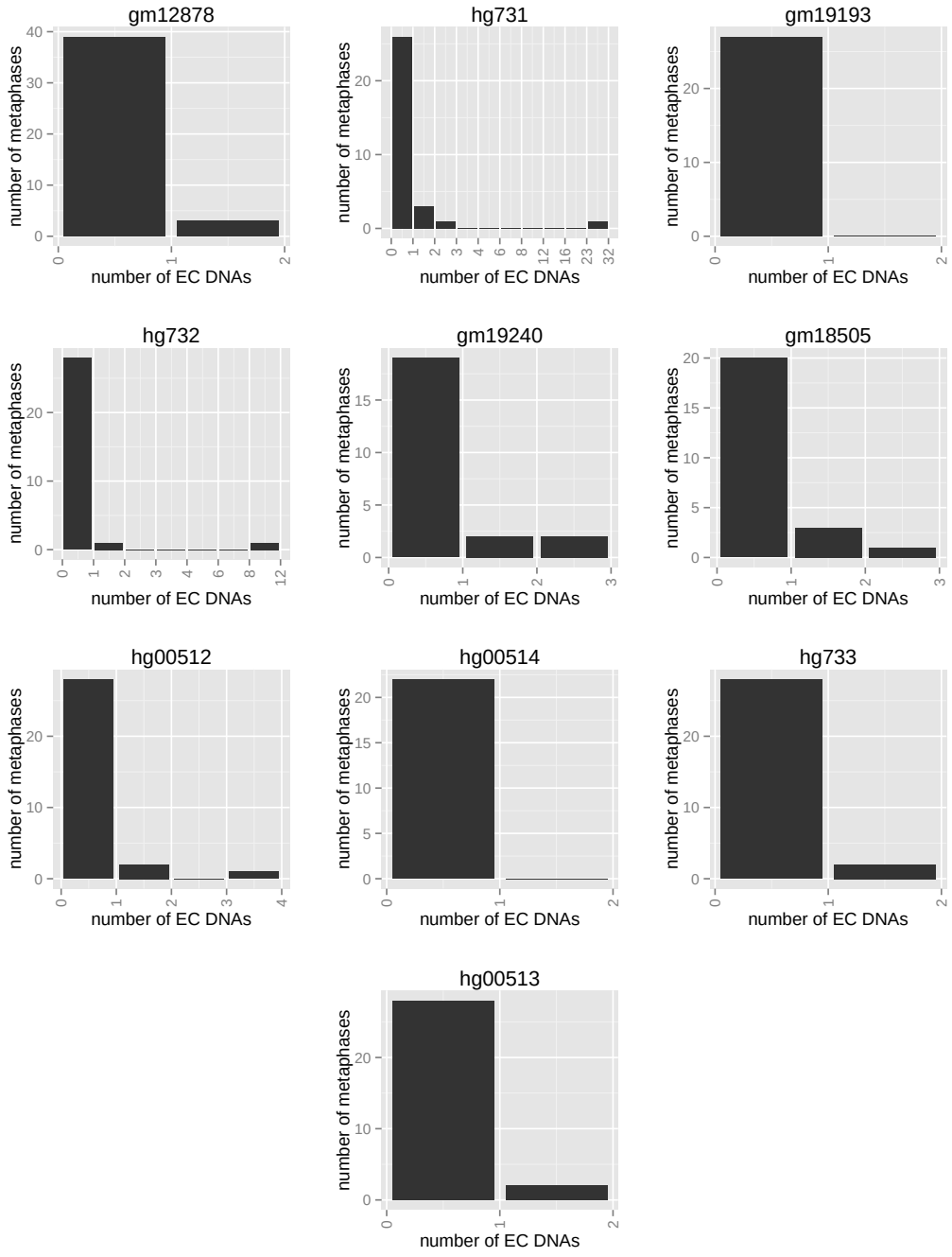


Figure S2.34: ECDNA count histograms of immortalized samples.

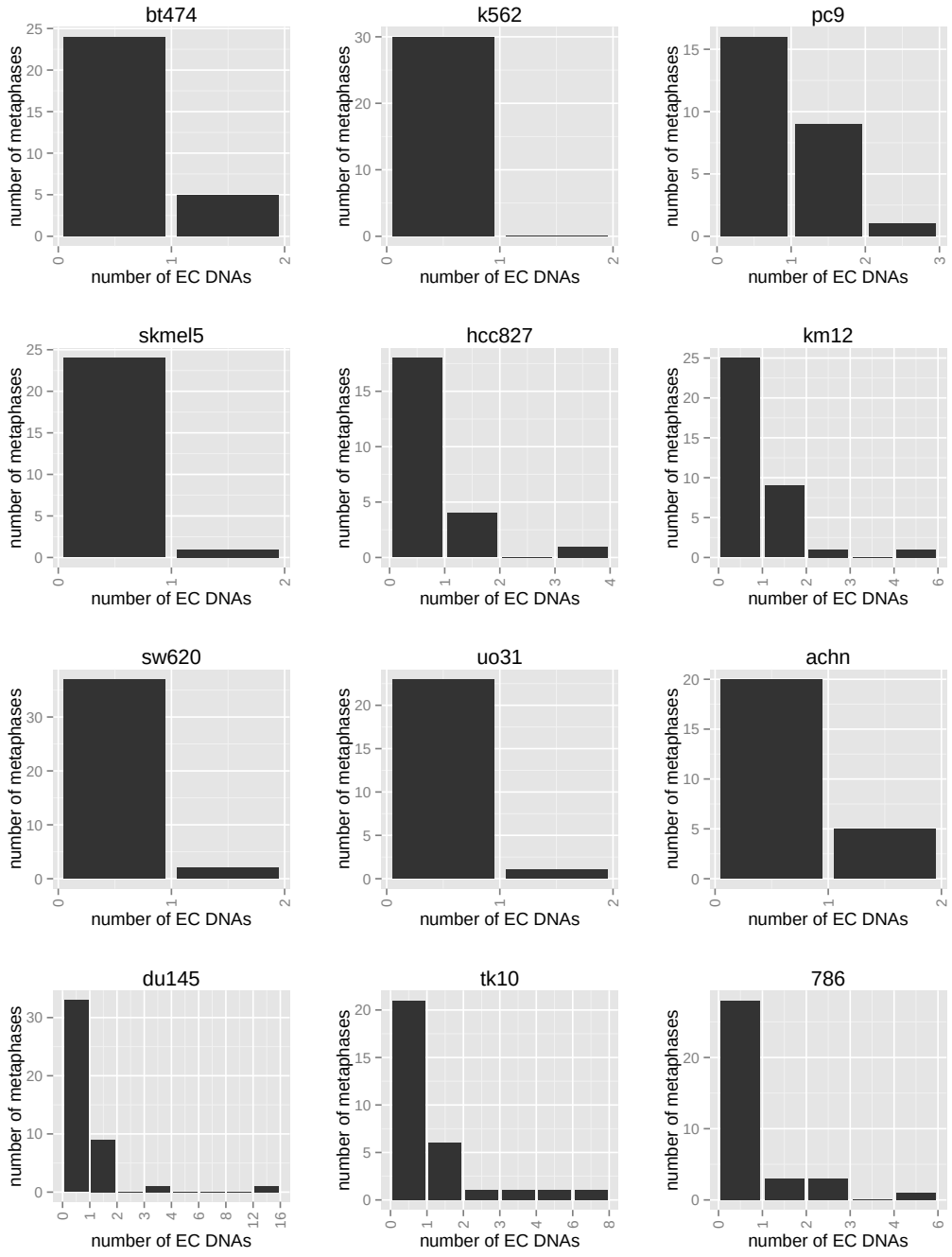


Figure S2.35: ECDNA count histograms of tumor cell line samples.

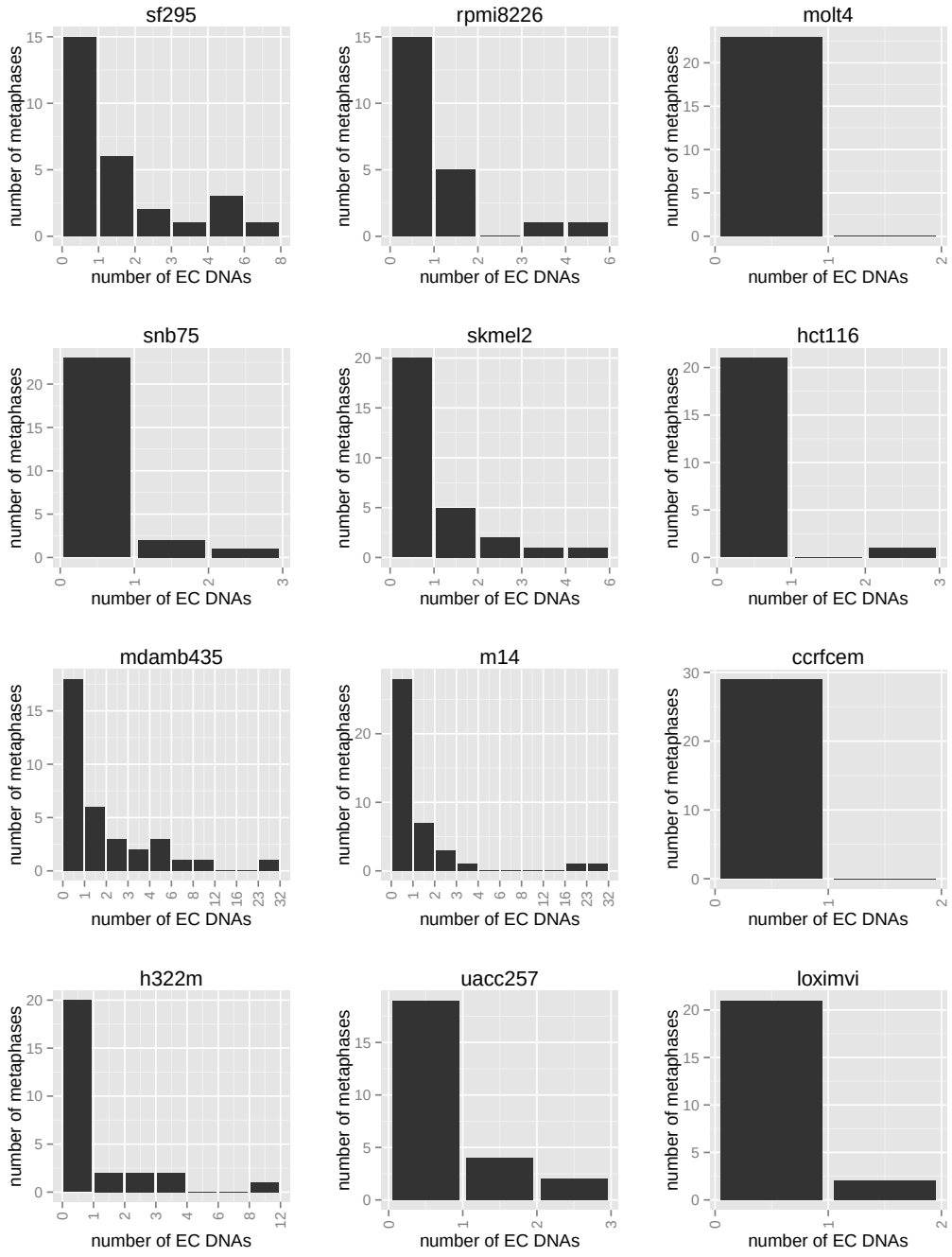


Figure S2.36: ECDNA count histograms of tumor cell line samples.

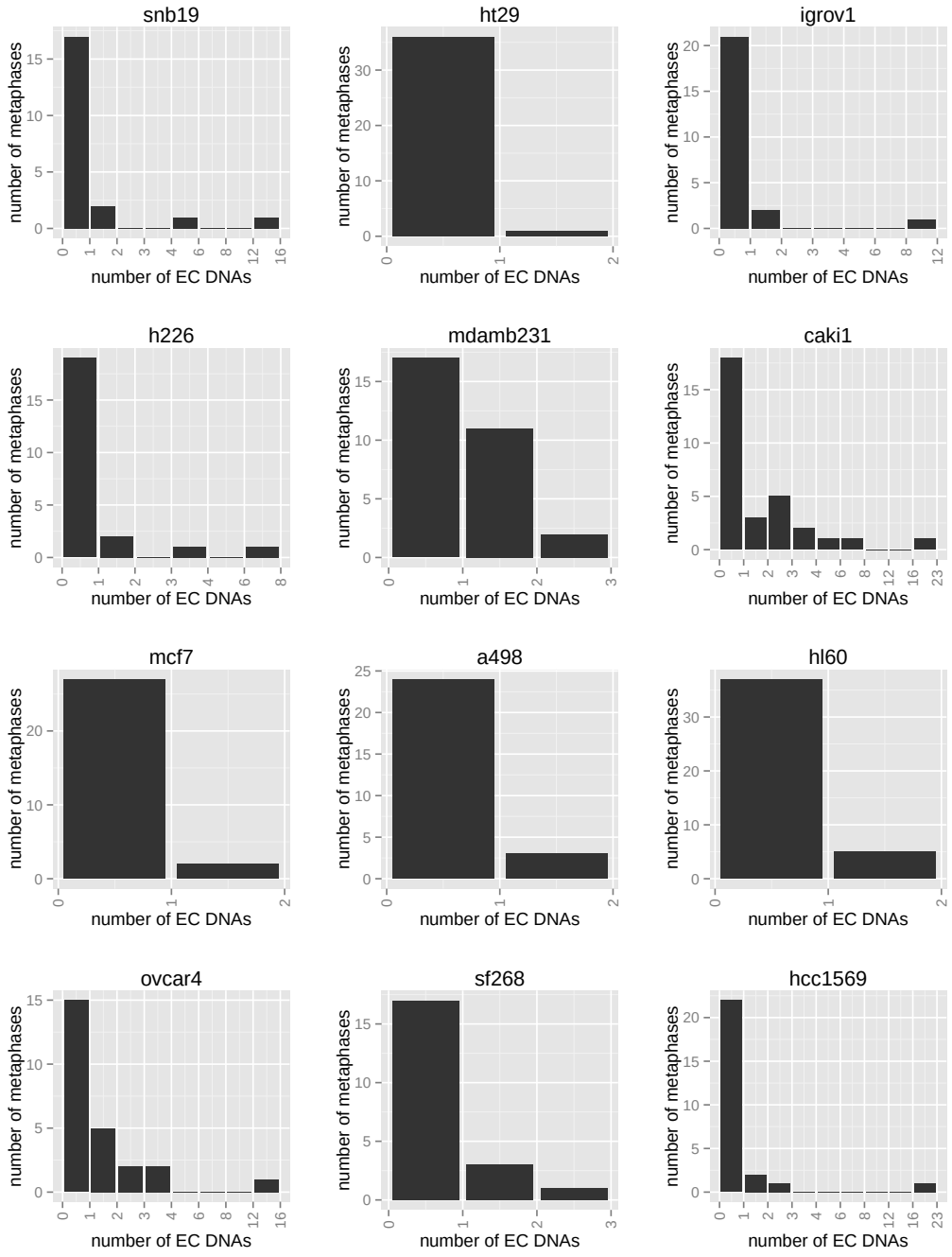


Figure S2.37: ECDNA count histograms of tumor cell line samples.

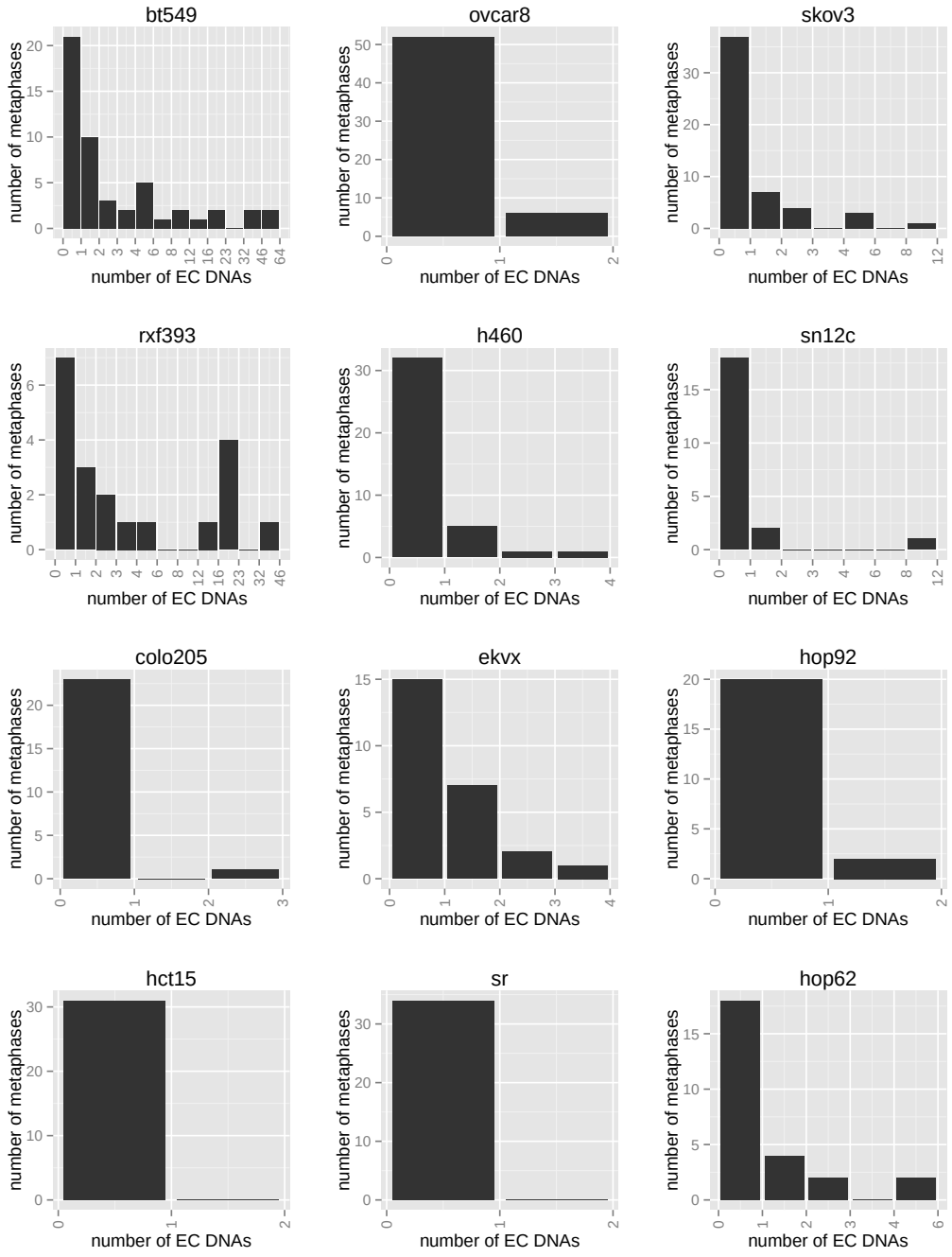


Figure S2.38: ECDNA count histograms of tumor cell line samples.

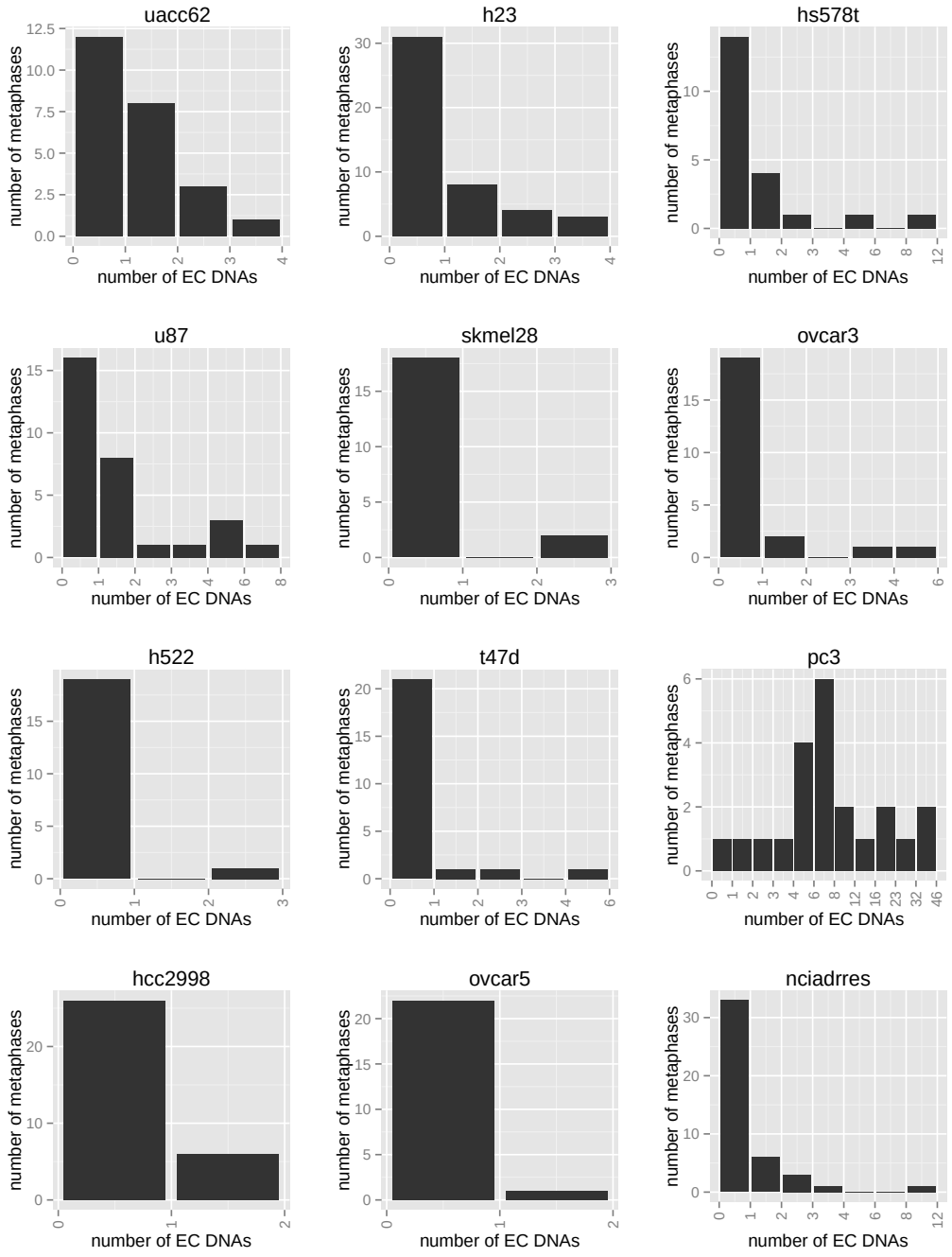


Figure S2.39: ECDNA count histograms of tumor cell line samples.

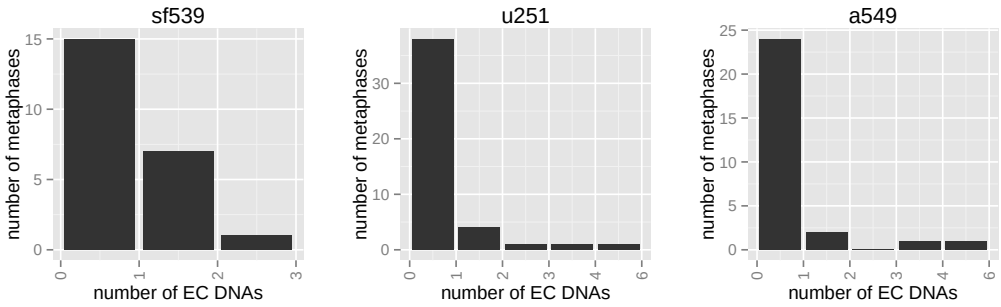


Figure S2.40: ECDNA count histograms of tumor cell line samples.

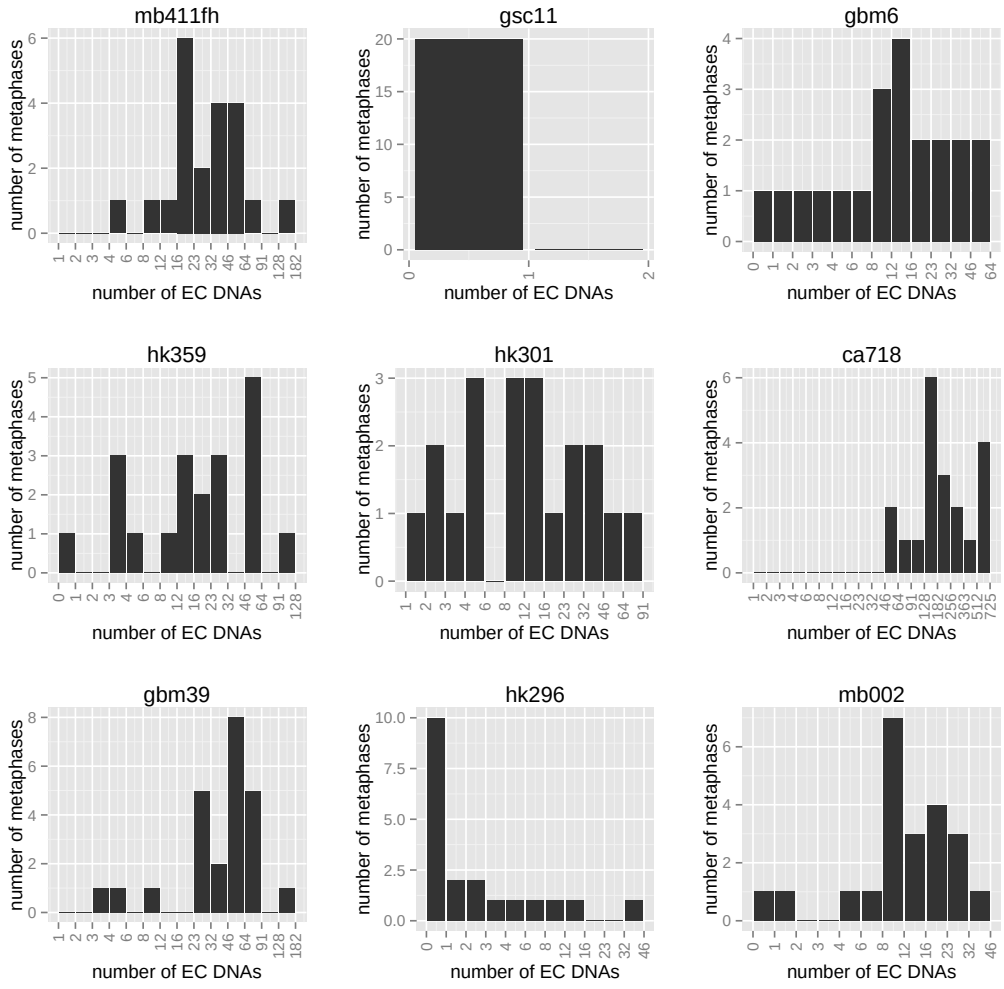


Figure S2.41: ECDNA count histograms of tumor PDX samples.

3. AmpliconArchitect: Sequence analysis for identification and reconstruction of focal amplifications

For the purpose of the AmpliconArchitect software, we focused on a set of genomic intervals that are simultaneously amplified to a high copy number. We define a *focal amplification* or an *amplicon* as a set of genomic intervals that are amplified to a high copy number, such that the intervals may be either contiguous or discontiguous on the reference genome, but are connected in the tumor cells in circular or linear structures. Different cells may contain different combinations of these genomic elements, and as long as they share common segments, we consider them as one amplicon in a sample. While we do not distinguish between the terms *focal amplifications* and *amplicons*, we do separate these events from *aneuploidies* where large chromosomal scale segments are amplified.

Using cytogenetic (mainly FISH) analysis, we can observe the existence of focal amplifications of the probed regions. By using multiple metaphase spreads, we can determine if those probes are amplified extra-chromosomally, intra-chromosomally, or both, and may be able to observe some heterogeneity in terms of size differences. However, cytogenetic analysis is limited to a few cells, does not reveal the fine structure of the amplicons. In contrast, genome sequencing techniques enable us to zoom into the fine-scale structure of genomic variants [4, 5], but provide additional complexities due to sampling from a heterogenous mix of amplicons from many cells. For this reason, existing computational tools (mainly tools that allow structural variation, or SV detection) are limited to identification of one or more rearrangement events and do not provide information of the connectivity and architecture of the larger genomic architecture (layout of genomic segments in one or more structures in a heterogenous mixture). We designed and developed AmpliconArchitect to enable the reconstruction of complex rearrangements in cancer amplicons from WGS data. AmpliconArchitect uses pre-processed data from mapped WGS reads, as described below.

3.1 Pre-processing

Identification of amplified regions.

We mapped whole genome paired-end Illumina reads from each tumor and normal sample to the hg19 (GRCh37) human reference sequence [6] downloaded from the UCSC genome browser site [7]. The BWA software version 0.7.9a was used with default parameters for mapping [8]. We inferred copy number variants from these mapped reads using the ReadDepth CNV software [9] version 0.9.8.4 with parameters FDR= 0.05 and overDispersion parameter= 1.

Filtering amplicons.

We used stringent filtering criteria to select amplified regions from both sequencing and TCGA datasets. In our starting set, we considered only CNV gain segments with copy count > 5 for samples from each dataset. We merged segments within 300kbp of each other into a single region and considered regions $>$ 100kbp in size. We applied 3 criteria to filter amplicons in repetitive/low-copy genomic regions as well as amplified regions reported in normal tissue samples to avoid sequencing and mapping artefacts:

1. **Regions amplified in normal samples**: Regions which had copy number of > 5 in 2 or more normal samples were labelled as uninteresting and extended by 1Mbp. A high copy region from a tumor sample which overlapped an uninteresting region was required to be at least 2Mbp in size after the part which overlapped the uninteresting region was trimmed.

2. **Repetitive regions**: We eliminated segments with average repeat count of > 2.5 (5 accounting for diploid genome) in the reference genome. The average reference repeat count of the region was calculated by defining a duke35 score [10, 11] of a genomic region based on Duke35 mappability. The duke35 score for an interval I was defined as
$$\text{duke35}(I) = \frac{\sum_{s \in I}(\text{length}(s)/\text{d35}(s))}{\text{length}(I)} \tag{3.1}$$
where s refers to each genomic segment defined in the Duke35 file which overlaps our region of interest, length(s) refers to length in base-pairs of the part of segment which overlaps the region and d35(s) refers to the value assigned to the segment in the Duke35 files. 1/d35(s) corresponds to the repeat count of the segment (extended by 34 base-pairs) in the reference genome. Thus regions with duke35(I) > 2.5 were eliminated.

3. **Segmental duplication regions**: We eliminated the regions of segmental duplications from the human paralog project [11–13] depending on the observed copy counts in our samples. If an interval I overlapped one or more segmental duplications, then the copy count of this interval was revised as the
$$\text{NewCount}(I) = \frac{\text{OriginalCount}(I) \cdot \text{length}(I)}{\text{length}(I) + \sum \text{length}(\text{overlapping segmental duplications})} \tag{3.2}$$
Only regions which had a revised copy count > 5 were retained.

3.2 Reconstructing amplicon architecture using AmpliconArchitect.

For each sample, AmpliconArchitect(AA) takes as input, an initial list of amplified intervals and whole genome sequencing (WGS) paired-end reads aligned to the human reference. The high level steps in AA are as follows:

1. Identify boundaries of segments in the reference genome that are part of the amplicon.

2. Build a breakpoint graph with nodes corresponding to segment-endpoints, and edges connecting pairs of nodes. The pairs may be from the same or different segments.

3. Use an optimization to estimate copy numbers of edges.

4. Extract paths and cycles in the graph that explain most of the copy number. These paths and cycles correspond to putative amplicon structures.

These steps are expanded upon below.

Sequencing statistics.

AmpliconArchitect samples a random subset of paired-end WGS reads to estimate sequencing parameters like read length, insert size, depth of coverage, and variability in coverage. We also estimate percentage of read pairs mapping concordantly (in the expected size and orientation). and expected number of read pairs that map across a genomic location. This expected number of read pairs within 3 standard deviations is used to identify clusters of discordant read pairs that indicate a genomic rearrangement.

Detecting segment boundaries.

We used two genomic signatures that suggest segment boundaries, as well as connections.

- Discordant read pair clusters: Recall that a genomic rearrangement can be indicated by a set of discordantly mapping read pair [4, 5]. The coordinates where the two reads map also provide the boundary of the segment, and indicate that the two segments are connected in the tumor genome. We used clusters of reads supporting the same rearrangement to identify segment boundaries as well as interconnections. We used filtering strategies based on the Duke35 mappability score described above to minimize false signals for rearrangements.

- Meanshift in coverage: Segment boundaries were also detected by a steep copy number change between adjacent or nearby locations. We used a mean-shift technique used in image processing for edge detection [14]. Specifically, we used a smoothed Gaussian kernel density function for coverage to find a span of genomic coordinates with similar values followed by a second span with different kernel density values (See also [15]). The locations determined to have shift in coverage were further investigated for rearrangements using discordant read clusters with less stringent criteria e.g., fewer number (~ 3) of read pairs.

Breakpoint graph construction.

Segment boundaries represent vertices in the breakpoint graph. Consecutive vertices that represented the beginning and end of a segment along the genome were connected by *sequence-edges*. Vertices linked by discordant read-pair clusters were connected using *breakpoint-edges*. We also used breakpoint edges to connect the end of one segment to the beginning of an adjacent segment. We introduced a special *source* vertex to represent ends of linear contigs or unidentified connections. A breakpoint edge was used to connect an existing vertex and the source vertex if we observed one-end mapping reads on the vertex, under the assumption that it represented an undiscovered rearrangement because one of the end-points was located in repetitive or novel/mutated sequence.

Copy count determination.

We assigned edge weights proportional to the number of reads mapping to each sequence-edge and breakpoint-edge. Assuming that shotgun reads follow a Poisson process, we formulated and optimized an objective function to normalize raw read counts into estimated copy counts for all edges of the breakpoint graph.

Paths and cycles in the graph that have a uniform copy number on all edges correspond to an amplified genomic sequence in the tumor genome. Given that the breakpoint graph represents the union of all of these amplifications, we obtain linear constraints on the copy numbers. The linear constraint (balanced-flow constraint) enforces that copy counts for breakpoint-edges incident at a breakpoint vertex should sum up to the copy count of the sequence-edge connected to the vertex. The optimized counts represent edge-weights in the breakpoint graph.

Amplicon Architecture determination.

We processed the edge-weighted breakpoint graph and extracted cycles. Cycles containing the source vertex represent paths beginning and ending at the two vertices adjacent to the source. The balanced-flow constraint ensures that we can always decompose the breakpoint graph into cycles and linear contigs such that the copy counts of edges in the subgraphs add up to the copy counts in the original graph. We used a polynomial-time heuristic which iteratively identifies the most dominant cycle or path, i.e. the cycle or path with the highest copy count until 80% of the genomic content in the breakpoint graph was accounted for in the extracted cycles. We note that the short insert lengths do not always allow an unambiguous and complete reconstruction of the amplified segment. However, the cycles provide a 'basis' decomposition, and cycles with common sequence-edges may be combined in multiple ways to form larger cycles to explore the full architecture and heterogeneity in the amplicon. An example of such a basis decomposition is presented in Figure S3.1 and the corresponding fine structure interpretation and visualization is presented in Figure S3.2.

3.3 Results

We sequenced 117 tumor samples including 63 cell lines, 19 neurospheres (PDX) and 35 cancer tissues with coverage ranging from 0.6× to 3.89×, excluding one sample with 0.06× coverage. See Extended Data Figure E4 for the coverage distribution across samples. We also sequenced additional 8 normal tissues as controls.

While the sequencing depth is low, it is sufficient to capture large regions with increased copy number. Consider the lowest mean coverage in our samples $c = 0.6$. For a region of size w ($w = 10^5$ in our tests), and copy count d, the expected number of 100bp reads with diploid genome

$$\lambda = \frac{wcd}{100 \cdot 2} = \frac{10^5 \cdot 0.6d}{200} = 150d$$

We assume the Null hypothesis that the number of reads in the region is Poisson distributed with parameter λ. Our goal is to exclude all regions with normal copy count, while including all regions with high copy numbers (e.g. $d \geq 6$). Consider an experiment where we select all regions of size w, containing at least 750 mapped reads. Then, the probability of a Type I error (including a region with copy count 2) is given by

$$1.0 - \text{Poisson-cdf}(750, \lambda = 300) \simeq 0.0$$

The probability of a Type II error (missing a region with $d \geq 6$) is at most

$$\text{Poisson-cdf}(750, \lambda = 900) = 1.5 \cdot 10^{-7}$$

The numbers are better for samples with higher sequence coverage, and larger amplified regions.

We identified 265 high-copy amplifications in 61 samples (see methods 3.1). We analyzed putative genomic connections between amplified regions to identify amplicon structures consisting of 1 or more amplified regions. The amplifications were assembled in 183 independent amplicons with copy count ranging from 2.64 to 132.11 and size ranging from 111Kbp to 67Mbp.

In order to estimate the significance of our observations, we downloaded copy number variation calls for 11079 tumor-normal samples covering 33 different tumor types from TCGA [16]. After merging and filtering the variant calls according to our criterion in Section 3.1, we identified 16408 amplicons in 3919 samples.

For each dataset, genome sequencing and TCGA, we computed a histogram for percentage of samples displaying an amplification at each genomic position. The weight in the histogram for samples in the genome sequencing dataset was adjusted to reflect the frequency of corresponding tumor types in TCGA samples. We found 20 peak regions amplified in more than 1% of TCGA samples. We compared these regions against 522 oncogenes from the COSMIC database (Aug 2014) [17] 13 out 20 regions contained an oncogene. We observed that 17 out of 20 regions were also captured by amplifications reported from our sequencing dataset, including all 13 oncogene regions most of each were amplified in multiple samples.

The genome sequencing samples displayed a wide variety of amplicon structures ranging from a simple circularization of a single genomic segment to mixtures of multiple structures(Sw620-MYC Supplementary Figure S3.2), amplicons containing complex rearrangements (MB002-MYC Supplementary Figure S3.3), similar structure simultaneously in EC and HSR (H460-MYC Supplementary Figure S3.4), multiple connected genomic regions. We identified one instance of a Breakage Fusion Bridge (HCC827-EGFR Supplementary Figure S3.5). FISH analysis revealed that some of these amplicons occurred as ECDNAs, HSRs or sometimes both, in the same sample. Many amplicons could be represented as cycles or closed walks on the breakpoint graph indicative of either ciruclar ECDNAs or tandemly duplicated HSRs. For many amplicons, most of the copy count could be explained by one or only a few cycles/walks indicating that the copies of amplificons consisted of a single or mixture of only a few distinct structures arising from a common origin.

```
Sw-620:  Amplicon 1
List of cycle segments:
 Segment 1 chr8 128603201 128773339
 Segment 2 chr8 128604396 128773339
 Segment 3 chr8 128604396 129210095
 Segment 4 chr8 129014958 129210095
 Segment 5 chr8 129014958 129307341
 Segment 6 chr8 129014958 129465168
 Segment 7 chr8 129216719 129307341
 Segment 8 chr8 129307359 129337939
 Segment 9 chr8 129307359 129338888
 Segment 10 chr8 129307359 129787552
 Segment 11 chr8 129371801 129465168
 Segment 12 chr8 129415256 129465168
 Segment 13 chr8 129451283 129465168
 Segment 14 chr8 129465169 129555740
 Segment 15 chr8 129471266 129555740
 Segment 16 chr8 129471266 129787552
 Segment 17 chr8 129480064 129555740
 Segment 18 chr8 129485384 129555740
 Segment 19 chr8 129485384 129789000
List of cycles:
 Cycle=1;Copy_count=9.8734302058;Segments=6+,15+,16-,12-,2-
 Cycle=2;Copy_count=7.92364061752;Segments=7+
 Cycle=3;Copy_count=3.06666021945;Segments=16+,15-,11-,9-,4-,2+,12+
 Cycle=4;Copy_count=1.9861498683;Segments=0+,2-,5+,0-
 Cycle=5;Copy_count=1.68895699871;Segments=0+,11-,0-
 Cycle=6;Copy_count=1.68623749159;Segments=0+,1-,0-
 Cycle=7;Copy_count=1.4485094014;Segments=0+,14-,0-
 Cycle=8;Copy_count=1.43019802705;Segments=0+,19-,0-
 Cycle=9;Copy_count=0.797613789551;Segments=16+,18-,13+
 Cycle=10;Copy_count=0.682924332011;Segments=0+,3+,8+,17+,10-,3-,0-
```

Figure S3.1: Sample output from AmpliconArchitect for amplicon reconstruction for Sw-620-cMYC amplicon

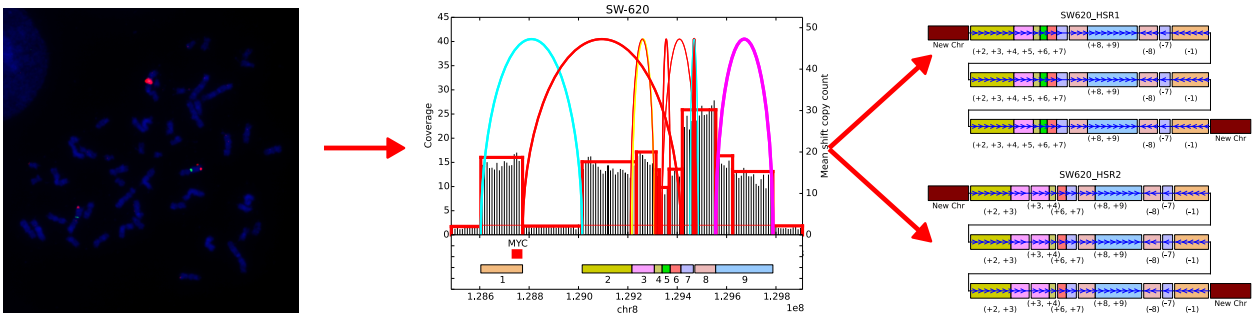


Figure S3.2: Fine structure analysis of c-MYC Amplification in Chromosomal DNA in Sw620 Colon Cancer Cells

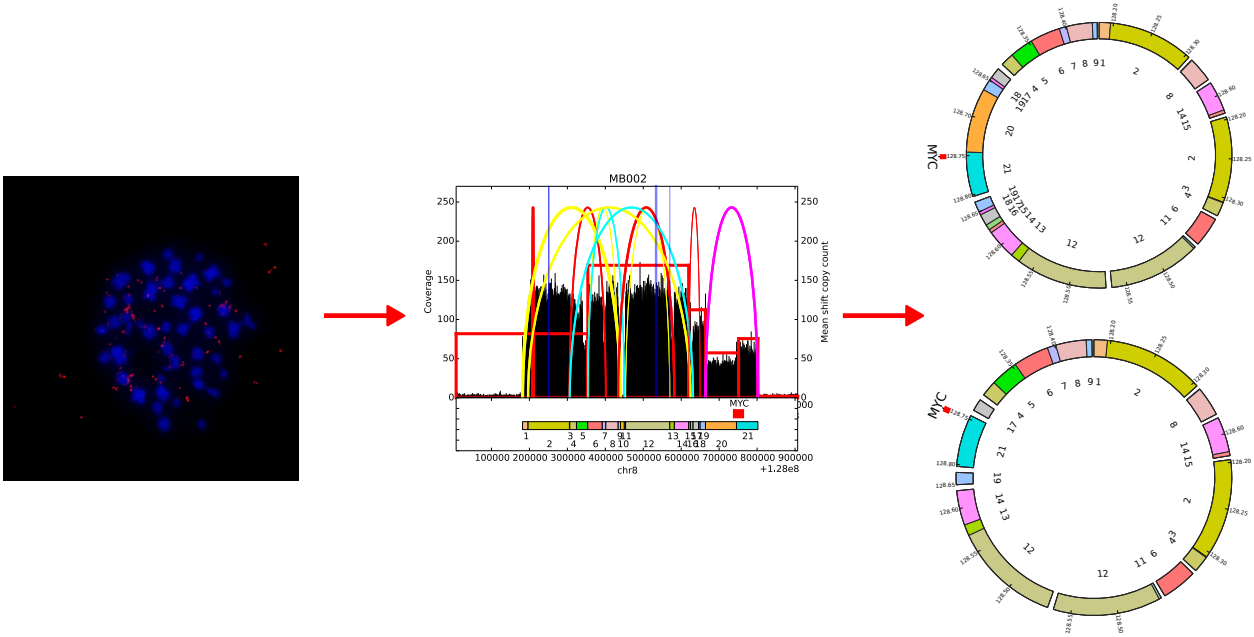


Figure S3.3: Fine structure analysis of c-MYC Amplification in Extrachromosomal DNA in Medulloblastoma MB002 Cells

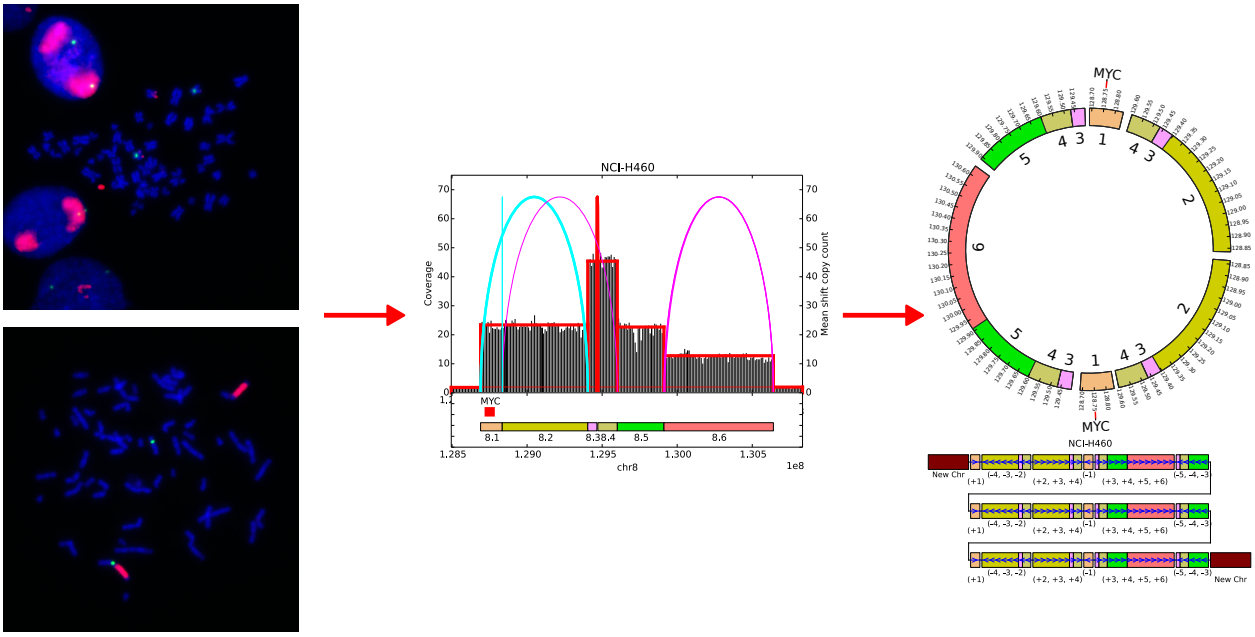


Figure S3.4: Fine structure analysis of c-MYC Amplification in Extrachromosomal and Chromosomal DNA in NCI H460 Non-Small Cell Lung Cancer Cells

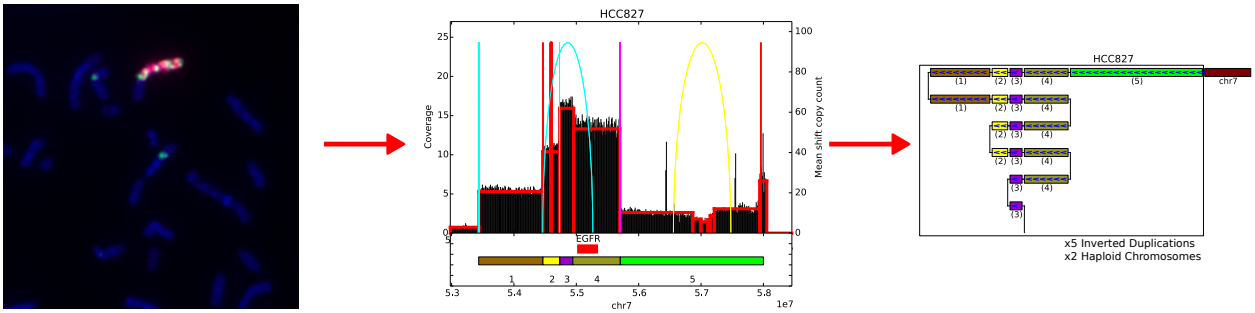


Figure S3.5: Fine structure analysis of EGFR Amplification in Chromosomal DNA via Breakage-Fusion-Bridge (BFB) mechanism in HCC827 Lung Adenocarcinoma Cells displays inverted duplications.

4. A theoretical model of extrachromosomal and intrachromosomal duplication

4.1 Model

Consider an initial population of N_0 cells, of which N_a cells contain a single extra copy of an oncogene. We model the population using a discrete generation Galton-Watson branching process [18]. In this simplified model, each cell in the current generation containing k amplicons (amplifying an oncogene) either dies with probability d_k, or replicates with probability b_k to create the next generation. We set the selective advantage

$$\frac{b_k}{d_k} = \begin{cases} 1 + sf_m(k), & 0 \leq k < M_a \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

$$d_k = 1 - b_k \tag{4.2}$$

In other words, cells with k copies of the amplicon stop dividing after reaching a limit of M_a amplicons. Otherwise, they have a selective advantage for $0 < k \leq M_a$, where the strength of selection is described by $f_m(k)$, as follows:

$$f_m(k) = \begin{cases} \frac{k}{M_s} & (0 \leq k \leq M_s)\ , \\ \frac{1}{1+e^{-\alpha(k-m)}} & (M_s < k < M_a). \end{cases} \tag{4.3}$$

Here, s denotes the selection-coefficient, and parameters m and α are the 'mid-point', and 'steepness' parameters of the logistic function, respectively. Initially, $f_m(k)$ grows linearly, reaching a peak value of $f_m(k) = 1$ for $k = M_s$. As the viability of cells with large number of amplicons is limited by available nutrition [19], $f_m(k)$ decreases logistically in value for $k > M_s$ reaching $f_m(k) \to 0$ for $k \geq M_a$. We model the decrease by a sigmoid function with a single mid-point parameter m s.t. $f_m(m) = \frac{1}{2}$. The 'steepness' parameter α is automatically adjusted to ensure that $\min\{1 - f_m(M_s), f_m(M_a)\} \to 0$.

The copy number change is effected by different mechanisms for extrachromosomal (EC) and intrachromosomal (HSR) models. In the EC model, the available k amplicons are on EC elements which replicate and segregate independently. We assume complete replication of EC elements so that there are $2k$ copies which are partitioned into the two daughter cells via independent segregation. Formally, the daughter cells end up with k_1 and k_2 amplicons respectively, where

$$k_1 \sim \mathcal{B}(2k, \frac{1}{2}) \tag{4.4}$$

$$k_2 = 2k - k_1 \tag{4.5}$$

In contrast, in the intrachromosomal model, the change in copy number happens via mitotic recombination, and the daughter cell of a cell with k amplicons will acquire either

$k+1$ amplicons or $k-1$ amplicons, each with probability p_d. With probability $1-2p_d$, the daughter cell retains k amplicons.

4.2 Model parameters

We started with an initial population $N_0 = 10^5$ and a small number of cells ($N_a = 100$) with one extra copy of an amplicon. We set $M_s = 15, M_a = 10^3$ for both, based on the observation of cells with $\sim 10^3$ EC elements (e.g. Extended Data Figure E10). While the number is excessive for intrachromosomal amplifications, we kept M_s, M_a identical for both EC and intrachromosomal events to allow for direct comparisons. It is well known that tumor cells have a selective advantage and proliferate; the rates are however different for different tumors and also within a sample, as cells acquiring multiple oncogenic mutations quickly grow more aggressively [18]. We chose different values of s $\{0.5, 1.0\}$ to explore different growth rates. For $s = 0.5$, $\frac{b_k}{d_k} \leq 1.5$, implying a tumor growth rate of $b_k - d_k = 2b_k - 1 = 0.2$ per generation. For $s = 1$, $\frac{b_k}{d_k} \leq 2$ implying a growth rate of 0.33 per generation. The results are not substantially different across different choices of s, with impact only on the rate of amplification and heterogeneity. While these choices provide maximum growth rate, the choice of the selection function $f_m(k)$ reduces the growth rate with increasing number of amplicons to model the effect of excessive metabolic demands on the cell. Once a cell reaches $M_a = 1000$, it stops replicating. The decay in selection function is modeled by a single parameter m, denoting the number of amplicon copies at which the selection strength is half of the peak strength.

Exponential growth of amplicon containing cells is seen in both extrachromosomal and intrachromosomal duplications. However, the tumor mass cannot grow indefinitely. We model the tumor as a sphere, and assume that 10^9 cells account for a tumor of 1cm diameter [20] although more recent accounts put the number for tumor cells as 10^8 cm^{-3} [21]. A physical limit of 20cm for the tumor diameter [22] implies a limit of 10^{13} tumor cells. We stop the simulation once the number of tumor cells reach 10^{14}. Note that more realistic models have been proposed where growth rate depends upon spatial constraints (e.g., see [23]). Tumors are modeled as spheres, but can only replicate on the surface of the sphere, or when there is dispersion of the tumor cells. Here, we work with the simpler model to focus on the differences between extrachromosomal and intrachromosomal methods of amplification.

In summary, the main difference in the two models is in the differing mechanisms for amplification. For intrachromosomal model, we experimented with different duplication probabilities ($0.01 \leq \text{HSR} \leq 0.1$). We chose a generation time of 3 days to measure time in days.

4.3 Results

Figures S4.1-S4.5 give the results for $s = 0.5$, while Figures S4.6-S4.9 show the results for $s = 1.0$. For each choice of s, the different figures vary only in the mid-point of the logistic

decay of the selection function (parameter m), which models the metabolic constraints.

The results are consistent in all cases. We see an exponential growth in the overall cell population, as well as in cells containing amplicons (Figures S4.1-S4.10). The amplicon containing cells take some time to establish, and then grow exponentially (Panel A in Figures). The rate of growth depends upon selection coefficient (s), and metabolic constraints (m). Our model is somewhat simplified as in most real situations, the growth does not continue indefinitely, but stabilizes due to spatial and metabolic constraints. We model metabolic constraints, but not spatial, in order to keep the model simple and to focus on the differences between extrachromosomal and intrachromosomal amplification.

The copy number of the amplicon (average number of copies per cell) grows for all cases, but the growth is slower for intrachromosomal compared to extrachromosomal (Panel B in all Figures). Similar behavior is observed for the number of amplicons per cell (Panel C in all Figures), and heterogeneity of copy number, measured as the Shannon entropy of the copy number distribution of amplicons (Panel D in all Figures). We note that when the metabolic constraints are weak (high values of m), heterogeneity and average number of amplicons per cell continue to grow. However, for stringent metabolic constraints, both heterogeneity and number of amplicons per cell stabilize, and even decrease, consistent with some long term studies [24].

Finally, heterogeneity grows along with copy number, but stabilizes (Panel E in all Figures). These model predictions are robust to choice of model parameters, and are borne out by experimental observations (Figure 4F of main paper).

Figure S4.10 shows the variance in trajectories in 10 simulation runs. We note that much of the variance comes from the fact that the amplicon containing cells take some time to establish, or reach their maximum growth rate. This time to establishment varies due from experiment to experiment due to the stochastic nature of the experiment. Otherwise, the results are consistent from run to run. As there can be a significant time gap between the establishment of cells, we did not compute the variance in number of cells between runs, but showed each trajectory separately.

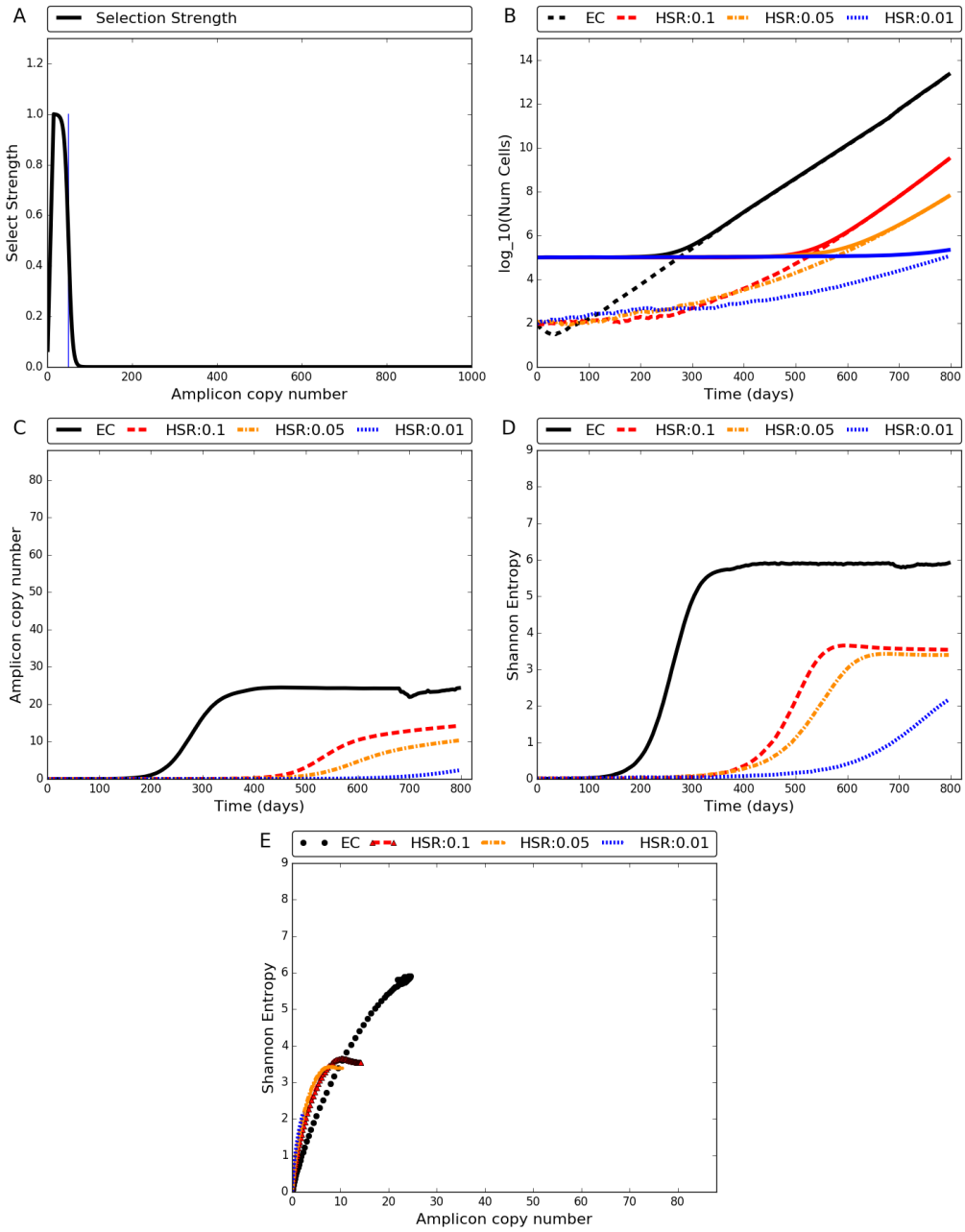


Figure S4.1: Evolution of tumor amplicons, with Initial Population $N_0 = 10^5$, selection-coefficient $s = 0.5$, decay parameter $m = 50$. (A) The selection function $f_m(k)$ with $m = 50$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in the amplicon copy number per cell over time. (D) Change in Shannon entropy of the number of amplicons per cell with time. (E) Change in entropy compared to change in copy number.

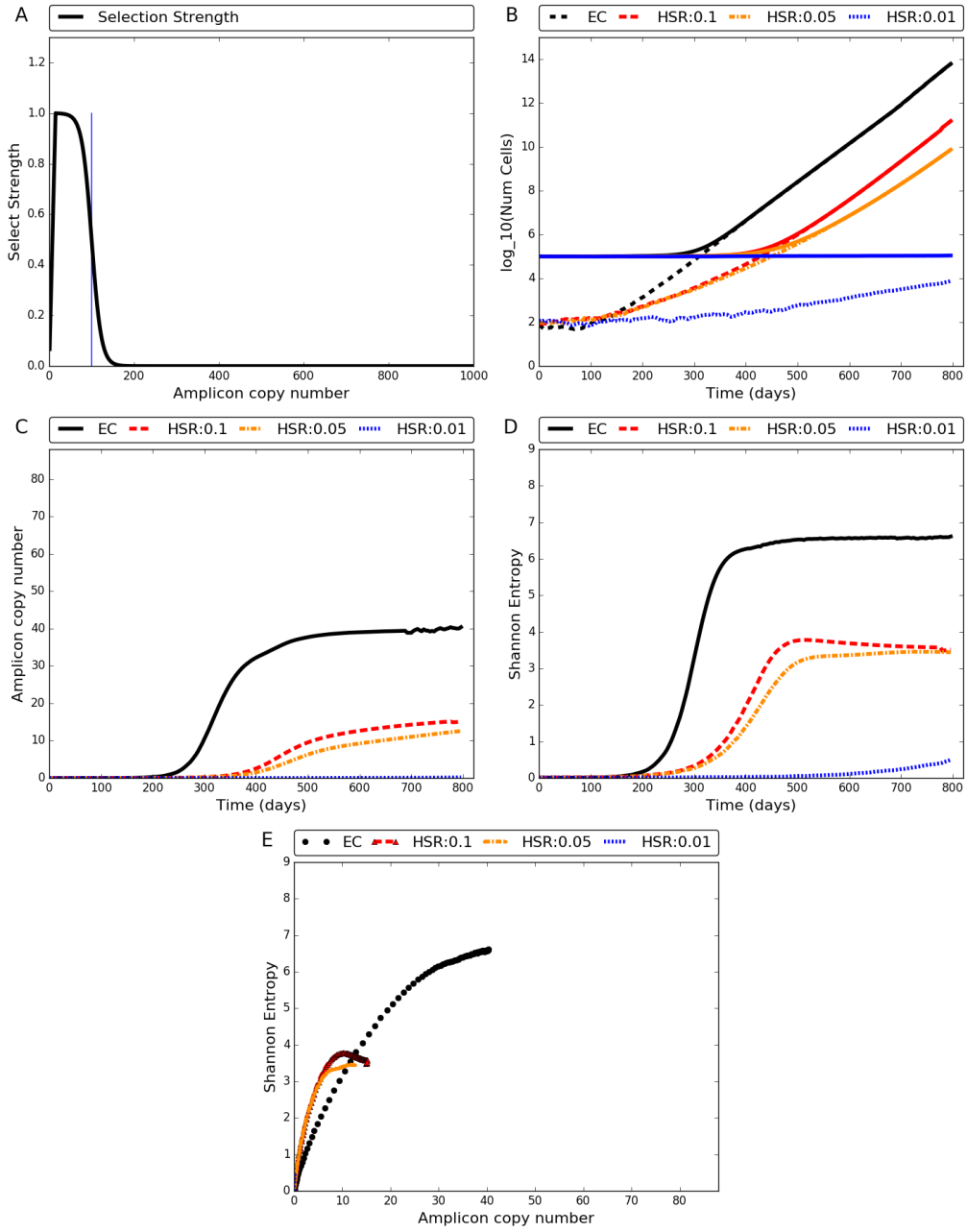


Figure S4.2: Tumor evolution with $N_0 = 10^5$, $s = 0.5$, $m = 100$. (A) The selection function $f_{100}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in the amplicon copy number per cell over time. (D) Change in Shannon entropy of the number of amplicons per cell with time. (E) Change in entropy compared to change in copy number.

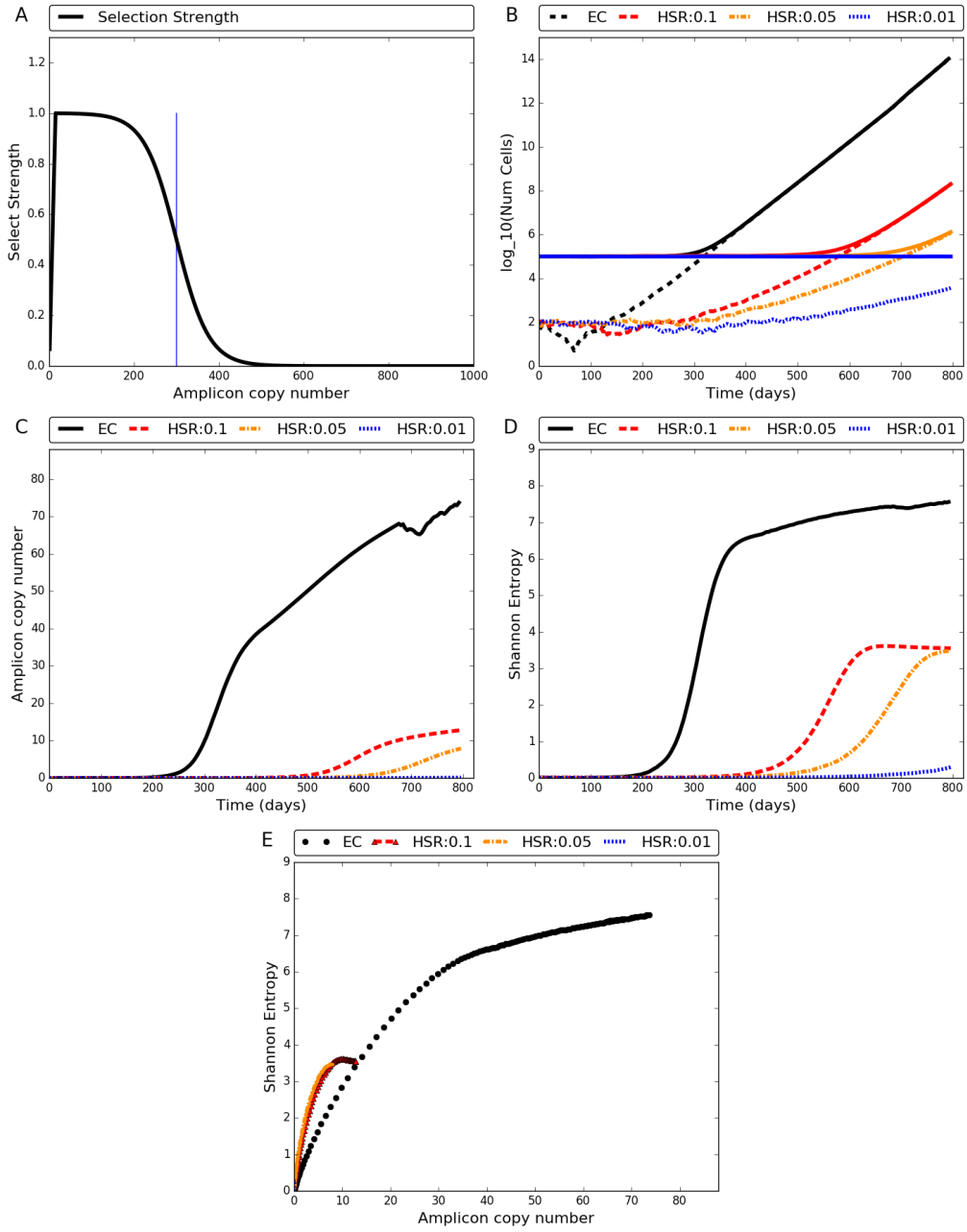


Figure S4.3: Tumor evolution with $N_0 = 10^5$, $s = 0.5$, $m = 300$. (A) The selection function $f_{300}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in the amplicon copy number per cell over time. (D) Change in Shannon entropy of the number of amplicons per cell with time. (E) Change in entropy compared to change in copy number,

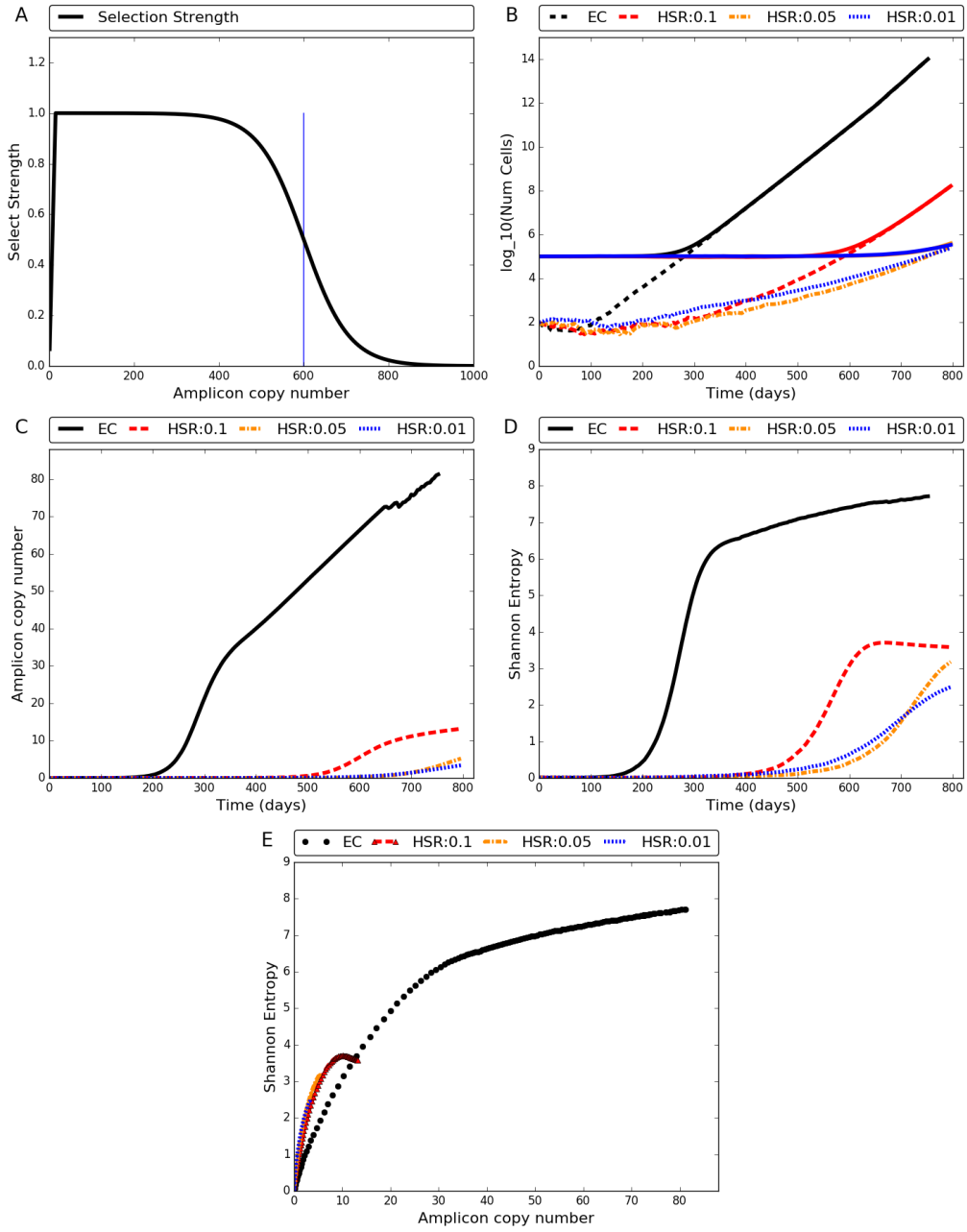


Figure S4.4: Tumor evolution with $N_0 = 10^5$, $s = 0.5$, $m = 600$. (A) The selection function $f_{600}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in the amplicon copy number per cell over time. (D) Change in Shannon entropy of the number of amplicons per cell with time. (E) Change in entropy compared to change in copy number.

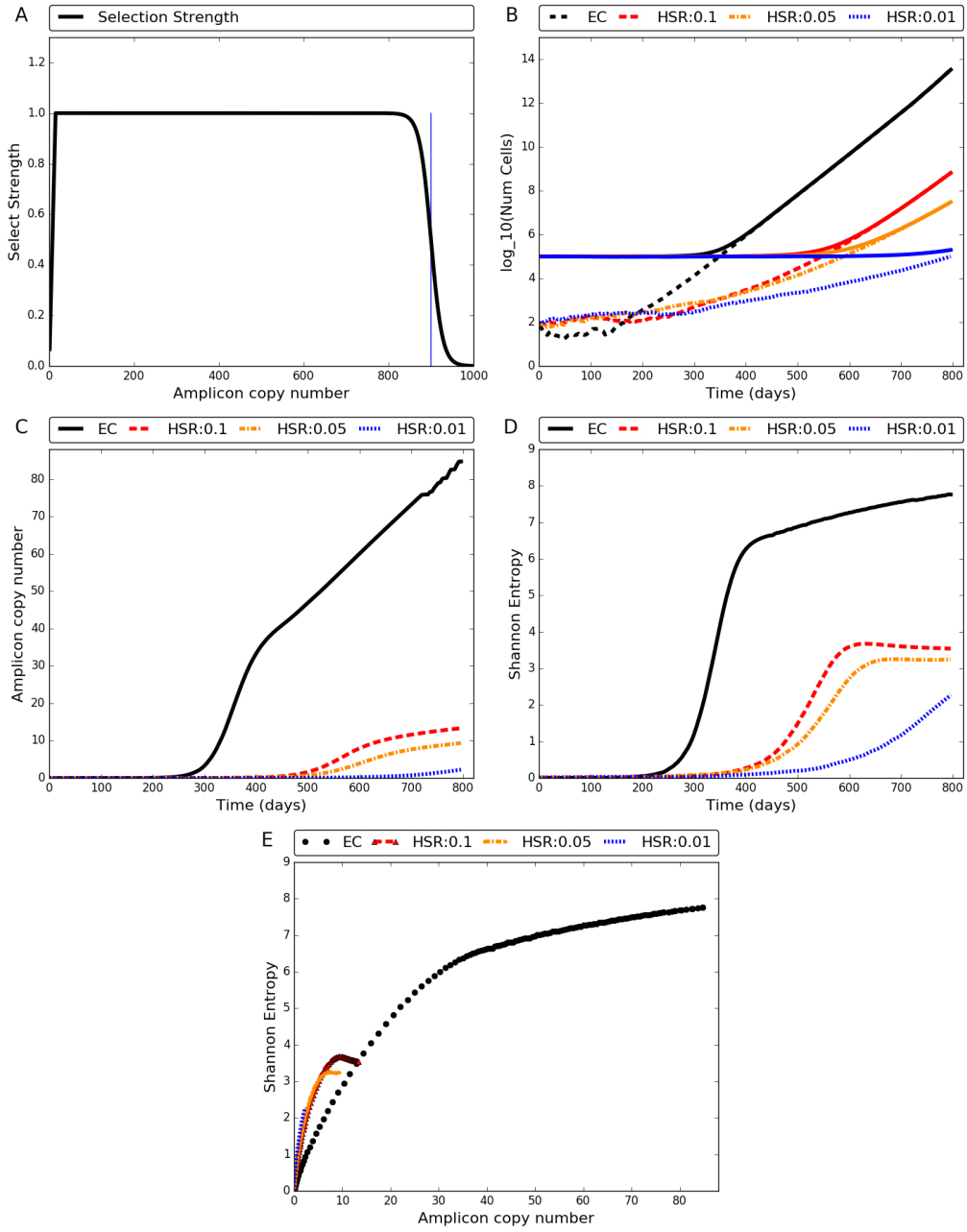


Figure S4.5: Tumor evolution with $N_0 = 10^5$, $s = 0.5$, $m = 900$. (A) The selection function $f_{900}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in average amplicon copy number over time. (D) Change in Shannon entropy with time. (E) Change in entropy compared to change in copy number.

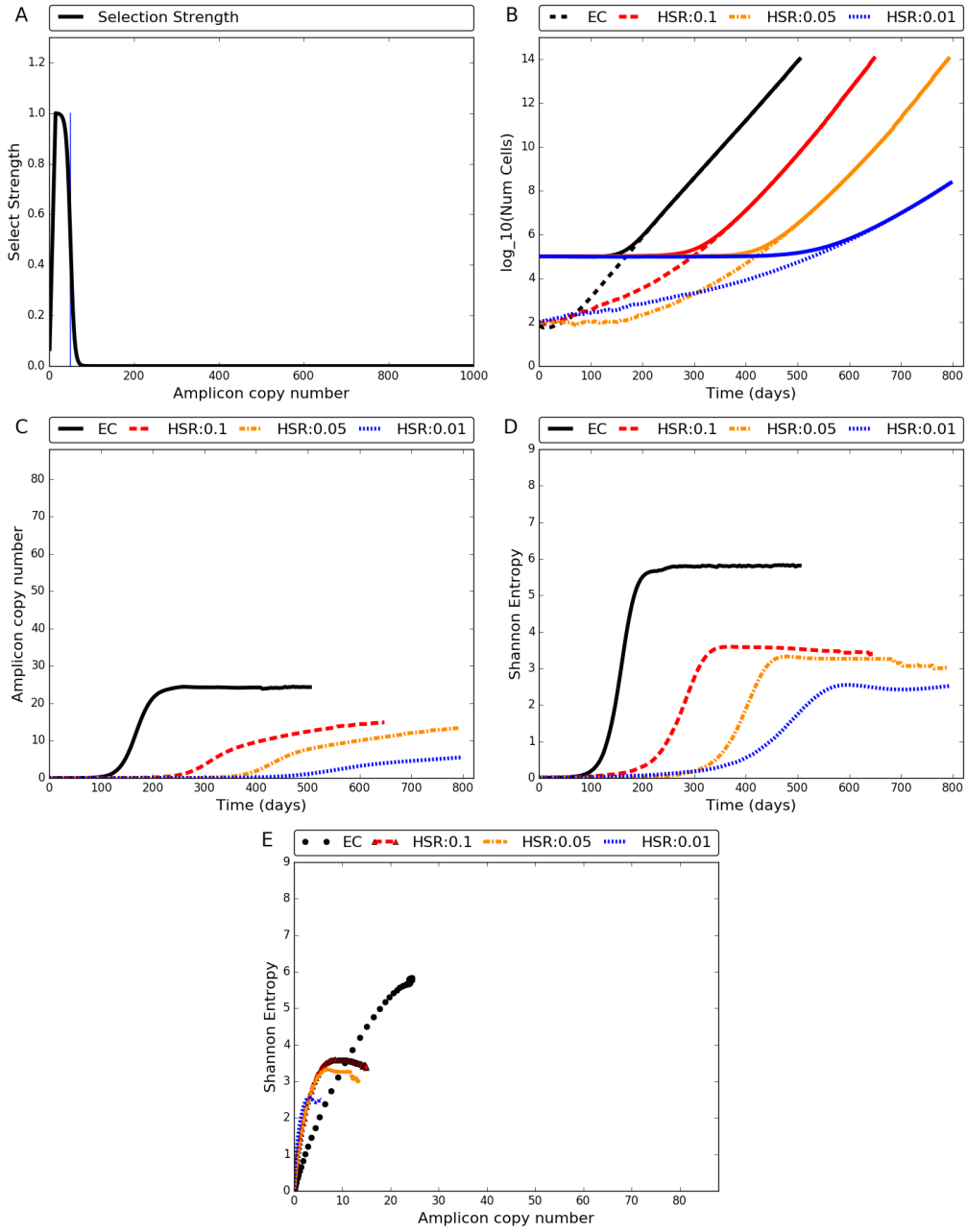


Figure S4.6: Tumor evolution with $N_0 = 10^5$, $s = 1.0$, $m = 50$. (A) The selection function $f_{50}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in average amplicon copy number over time. (D) Change in Shannon entropy with time. (E) Change in entropy compared to change in copy number.

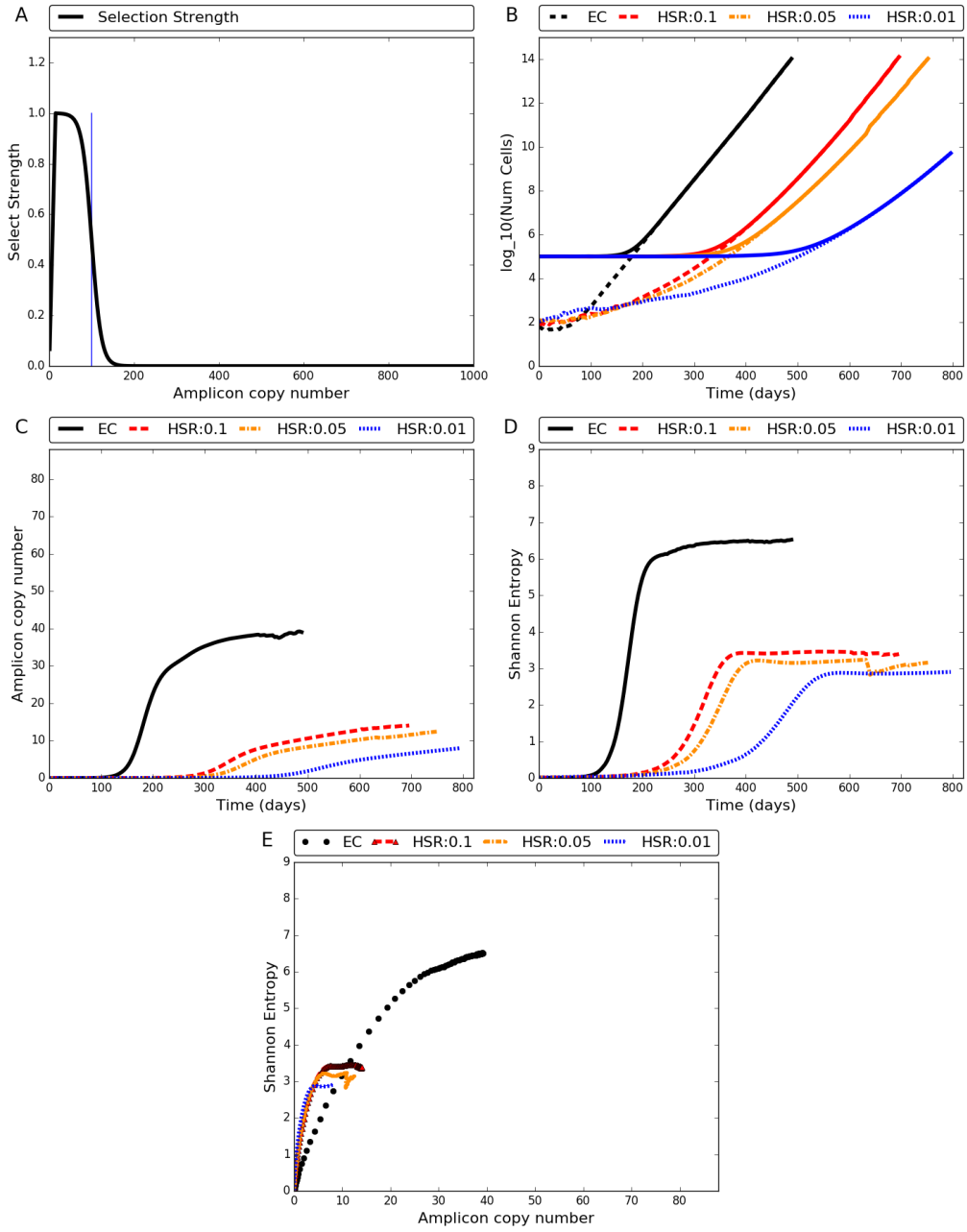


Figure S4.7: Tumor evolution with $N_0 = 10^5$, $s = 1.0$, $m = 100$. (A) The selection function $f_{100}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in average amplicon copy number over time. (D) Change in Shannon entropy with time. (E) Change in entropy compared to change in copy number.

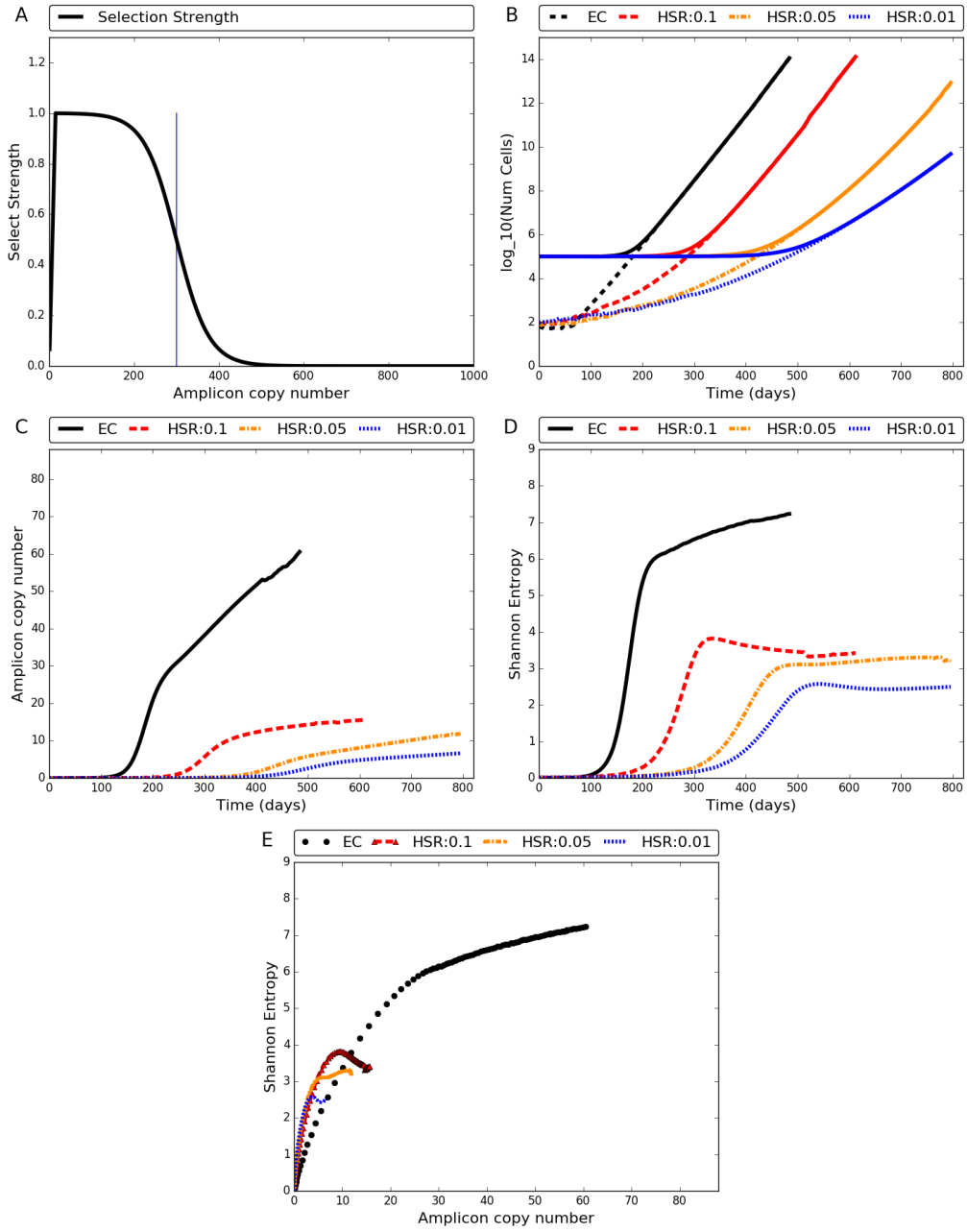


Figure S4.8: Tumor evolution with $N_0 = 10^5$, $s = 1.0$, $m = 300$. (A) The selection function $f_{300}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in average amplicon copy number over time. (D) Change in Shannon entropy with time. (E) Change in entropy compared to change in copy number.

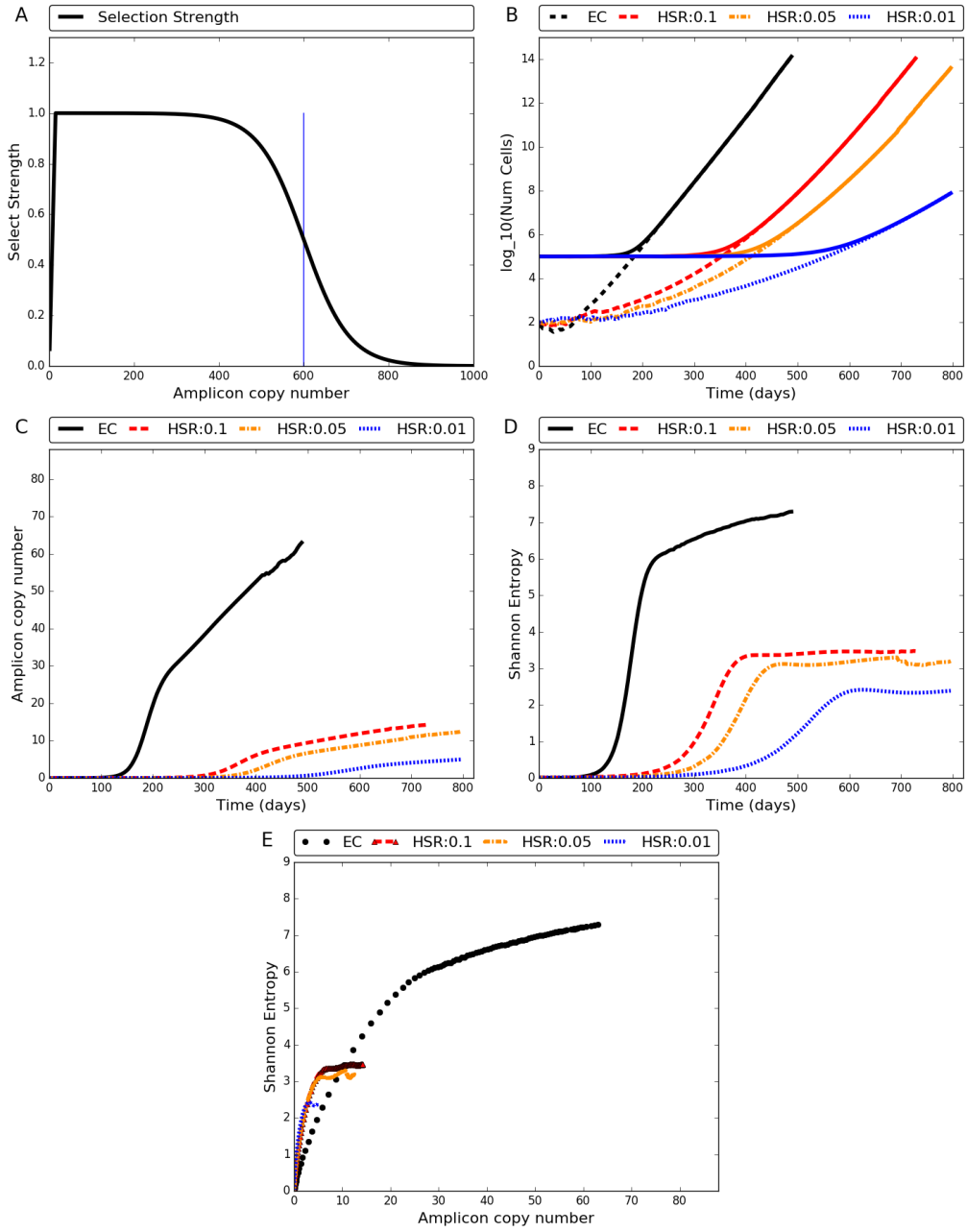


Figure S4.9: Tumor evolution with $N_0 = 10^5$, $s = 1.0$, $m = 600$. (A) The selection function $f_{600}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1+sf_m(k)$. (B) Growth of cells over time with EC amplicon (black) compared to growth with intrachromosomal amplification (HSR) with duplication probabilities 0.1 (red-line); 0.05 (dark-orange); 0.01 (blue). The dotted lines represent the number of cells containing amplicons, starting with 100 amplicon containing cells, while solid lines depict the total number of cells in the population. (C) Increase in average amplicon copy number over time. (D) Change in Shannon entropy with time. (E) Change in entropy compared to change in copy number.

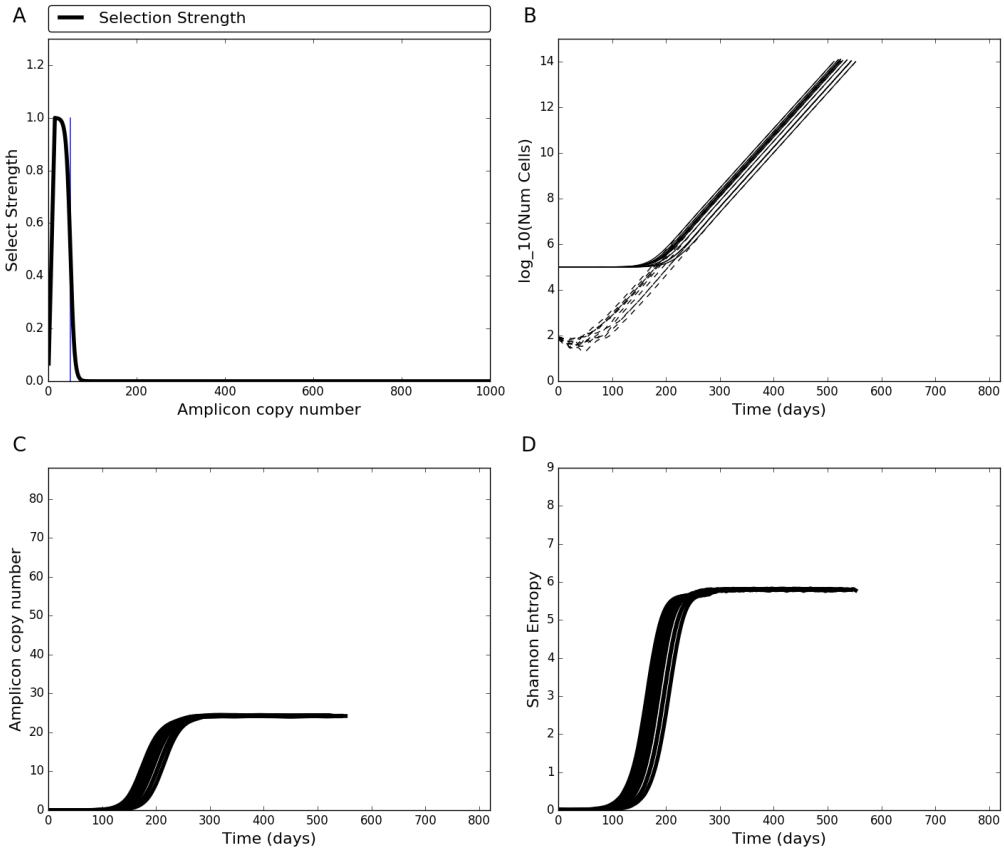


Figure S4.10: Tumor evolution trajectories with $N_0 = 10^5$, $s = 1.0$, $m = 50$. (A) The selection function $f_{50}(k)$. The ratio of birth to death rate for a cell with k amplicon copies is given by $1 + sf_m(k)$. (B-D) 10 simulation trajectories showing growth of cells over time (B); Increase in average amplicon copy number over time (C); and, Change in Shannon entropy with time (D). The trajectories are consistent, with variation due to difference in 'establishment time' of amplicon containing cells.

References

[1] P. M. Lee, *Bayesian statistics: an introduction* (John Wiley & Sons, 2012).

[2] D. Bradley, G. Roth, *Journal of graphics, gpu, and game tools* **12**, 13 (2007).

[3] Jan Motl, Available at: `https://www.mathworks.com/matlabcentral/fileexchange/40854` (2015). Last Accessed: 11 July 2016.

[4] E. Tuzun, *et al.*, *Nat. Genet.* **37**, 727 (2005).

[5] E. E. Eichler, *et al.*, *Nature* **447**, 161 (2007).

[6] E. S. Lander, *et al.*, *Nature* **409**, 860 (2001).

[7] W. J. Kent, *et al.*, *Genome Res.* **12**, 996 (2002).

[8] H. Li, R. Durbin, *Bioinformatics* **25**, 1754 (2009).

[9] C. A. Miller, O. Hampton, C. Coarfa, A. Milosavljevic, *PLoS ONE* **6**, e16327 (2011).

[10] T. Derrien, *et al.*, *PLoS ONE* **7**, e30377 (2012).

[11] K. R. Rosenbloom, *et al.*, *Nucleic Acids Res.* **43**, D670 (2015).

[12] J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res.* **11**, 1005 (2001).

[13] J. A. Bailey, *et al.*, *Science* **297**, 1003 (2002).

[14] D. Comaniciu, P. Meer, *IEEE Transactions on pattern analysis and machine intelligence* **24**, 603 (2002).

[15] A. Abyzov, A. E. Urban, M. Snyder, M. Gerstein, *Genome Res.* **21**, 974 (2011).

[16] The Cancer Genome Atlas (TCGA) Research Network, *Nature* **455**, 1061 (2008).

[17] S. A. Forbes, *et al.*, *Nucleic Acids Res.* **43**, D805 (2015).

[18] I. Bozic, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18545 (2010).

[19] N. N. Pavlova, C. B. Thompson, *Cell Metab.* **23**, 27 (2016).

[20] V. T. DeVita, R. C. Young, G. P. Canellos, *Cancer* **35**, 98 (1975).

[21] U. Del Monte, *Cell Cycle* **8**, 505 (2009).

[22] M. F. Dempsey, B. R. Condon, D. M. Hadley, *AJNR Am J Neuroradiol* **26**, 770 (2005).

[23] B. Waclaw, *et al.*, *Nature* **525**, 261 (2015).

[24] X. Li, *et al.*, *Cancer Prev Res (Phila)* **7**, 114 (2014).