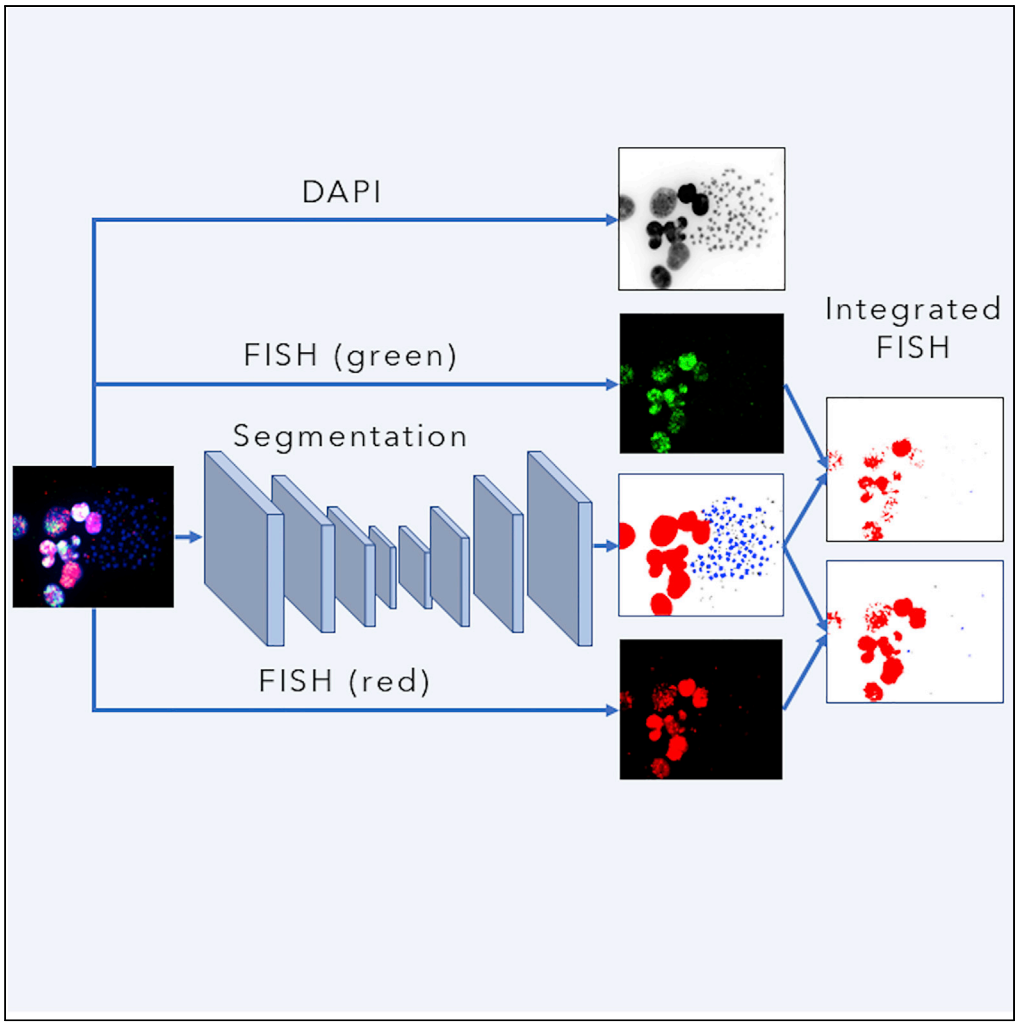
iScience

CellPress

Article

EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA



Utkrisht Rajkumar, Kristen Turner, Jens Luebeck, Viraj Deshpande, Manmohan Chandraker, Paul Mischel, Vineet Bafna

vbafna@ucsd.edu

HIGHLIGHTS

We identify extrachromosomal DNA (ecDNA) in metaphase spreads using deep learning

ecSeg integrates DAPI with FISH probes to provide oncogene amplification location

High intra-tumoral heterogeneity of ecDNA drives cancer pathogenesis

Rajkumar et al., iScience *21*, 428–435
November 22, 2019
https://doi.org/10.1016/j.isci.2019.10.035

iScience

CellPress

Article

EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA

Utkrisht Rajkumar,[1] Kristen Turner,[2] Jens Luebeck,[1] Viraj Deshpande,[5] Manmohan Chandraker,[1] Paul Mischel,[2,3,4] and Vineet Bafna[1,4,6,*]

SUMMARY

Oncogene amplification is one of the most common drivers of genetic events in cancer, potently promoting tumor development, growth, and progression. The recent discovery that oncogene amplification commonly occurs on extrachromosomal DNA, driving intratumoral genetic heterogeneity and high copy number owing to its non-chromosomal mechanism of inheritance, raises important questions about how the subnuclear location of amplified oncogenes mediates tumor pathogenesis. Next-generation sequencing is powerful but does not provide spatial resolution for amplified oncogenes, and new approaches are needed for accurately quantifying oncogenes located on ecDNA. Here, we introduce ecSeg, an image analysis tool that integrates conventional microscopy with deep neural networks to accurately resolve ecDNA and oncogene amplification at the single cell level.

INTRODUCTION

Despite the well-recognized importance of oncogene amplification in cancer pathogenesis (Davoli et al., 2017), the underlying mechanisms remain incompletely understood. How do amplified oncogenes reach such a high copy number in many tumors while still showing considerable cell-to-cell variability? Numerous mechanisms, including tandem duplications (Menghi et al., 2018), breakage fusion breakage cycles (Kitada and Yamasaki, 2008), aneuploidies (Davoli et al., 2017) chromothripsis (Ly and Cleveland, 2017), and neochromosome formation (Garsed et al., 2014) events have been implicated in oncogene amplification, but the recent discovery that extrachromosomal (ecDNA) oncogene amplification is common across a wide variety of tumor types (Turner et al., 2017; Verhaak et al., 2019) has raised new interest in understanding where amplified oncogenes actually reside within the genome of tumor cells.

In fact, ecDNA have long been found to occur in cancer cells studied in metaphase (Cox et al., 1965), referred to as double minutes, but the difficulty in linking these observations with modern cancer genomics led to a massive underestimation of their prevalence (Verhaak et al., 2019). In part, the challenge has been made more difficult by the fact that the 3D structure of DNA in an intact nucleus does not permit unambiguous localization of a particular gene, especially when there are many copies of that gene. Recently, sequence-based methods (Deshpande et al., 2019) have been developed to reconstruct the fine structure of focal amplifications, including ecDNA. However, ecDNA are known to reintegrate into and egress out of chromosomes based on cellular environment (Nathanson et al., 2014) while maintaining their structural features. For example, focal amplifications containing epidermal growth factor receptor (EGFR) have identical structures but can be extrachromosomal or integrated into non-native locations within chromosomes (Figures 1A–1D). Therefore, sequence-based reconstructions have limited power in revealing the spatial location of focal amplifications.

The use of fluorescence in situ hybridization (FISH) probes to study amplified oncogenes in interphase nuclei often reveals a pattern of many FISH probe-positive spots, but with limited ability to discriminate between their chromosomal and extrachromosomal location. During metaphase, the compact alignment of chromosomes enables unambiguous localization of specific genes within the genome and ecDNA can be detected using FISH probes. Moreover, the cell-to-cell variability in terms of ecDNA content and number poses additional challenges.

To accurately quantify ecDNA in cells, we investigated DAPI stained images of cells in metaphase, when chromosomal structures are condensed and separated from ecDNA. However, the large class imbalance in cellular features (Figure 2A), inherently high noise ratio in metaphase images, small size, and paucity of morphological features in ecDNA (particularly in comparison with chromosomes), present challenges

[1]Department of Computer Science & Engineering, UC San Diego, La Jolla, CA, USA

[2]Ludwig Institute for Cancer Research San Diego, UC San Diego, La Jolla, CA, USA

[3]Department of Pathology, UC San Diego, La Jolla, CA, USA

[4]Moores Cancer Center, UC San Diego, La Jolla, CA, USA

[5]Illumina Inc., 5200 Illumina Way, San Diego, CA, USA

[6]Lead Contact

*Correspondence: vbafna@ucsd.edu

https://doi.org/10.1016/j.isci.2019.10.035

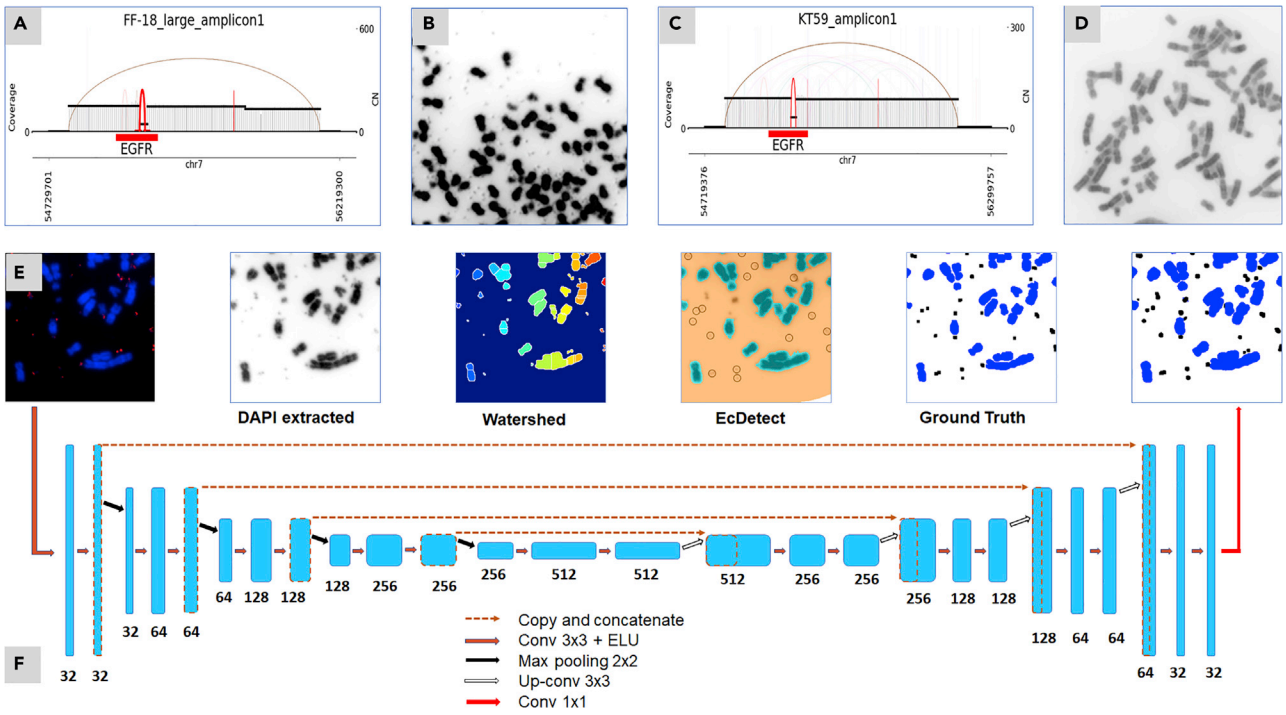


Figure 1. Detecting ecDNA in DAPI-Stained Images

(A and B) Copy number amplification of EGFR in a glioblastoma cell line due to extrachromosomal DNA (ecDNA) formation.

(C and D) Copy number amplification of EGFR in a glioblastoma cell line with no ecDNA. Note that the sequence-based reconstruction does not distinguish between ecDNA (A) and homogeneously stained regions (C).

(E) Identification of ecDNA in DAPI-stained images of cells in metaphase. Although a FISH signal for EGFR is also shown, only the DAPI signals are used for calling ecDNA using the Watershed method, ecDetect, manually annotated ground truth, and ecSeg.

(F) A neural network architecture for semantic segmentation of the pixel into ecDNA, chromosomes, nuclei, and background, as described in Transparent Methods.

to identifying and segmenting ecDNA. Moreover, we observed large variance in prevalence, relative pixel intensity, size, and presence of other partially burst nuclei across samples. It is difficult for traditional image processing algorithms with hand-crafted features to automatically account for the high level of contextual information and different sources of variation between cell types and images within a cell type.

For these reasons, computational tools for identifying ecDNA in DAPI-stained images are limited. Figure 1E shows a patch from a DAPI-stained image of a cell in metaphase. An off-the-shelf method, Watershed (Beucher and Lantuejoul, 1979) detected only 14 of 25 ecDNA. An earlier tool, ecDetect (Turner et al., 2017), used image processing methods, including thresholding and morphological operations, to detect and quantify ecDNA. The tool was optimized for precision and had low recall. For example, when applied to a cropped image from a metaphase PC3 cell line (Figure 1E), ecDetect detected only 16 of 25 ecDNA. Moreover, although the tool performs accurate segmentation of intact nuclei and chromosomes, it does not automatically differentiate between the two classes and requires further post-processing. Therefore, ecDetect cannot identify FISH-stained homogeneously staining regions. Additionally, ecDetect often requires a manual step to demarcate the regions containing metaphase chromosomes. We address these issues in our method.

Deep neural networks, specifically convolutional neural networks, have consistently outperformed traditional image processing algorithms on biological datasets (Hamilton, 2018). By using a large number of learnable parameters, they can recognize complex underlying patterns in large datasets (Krizhevsky et al., 2012). In contrast to image classification tasks, fully convolutional networks were constructed to perform pixel-wise classification (Long et al., 2015). Pixel-wise classifications allows for images to be semantically segmented, i.e., acquire class labels while retaining their spatial organization. However, ecDNA are small and irregularly shaped and can be confused with other proximal objects. Thus, resolving local details

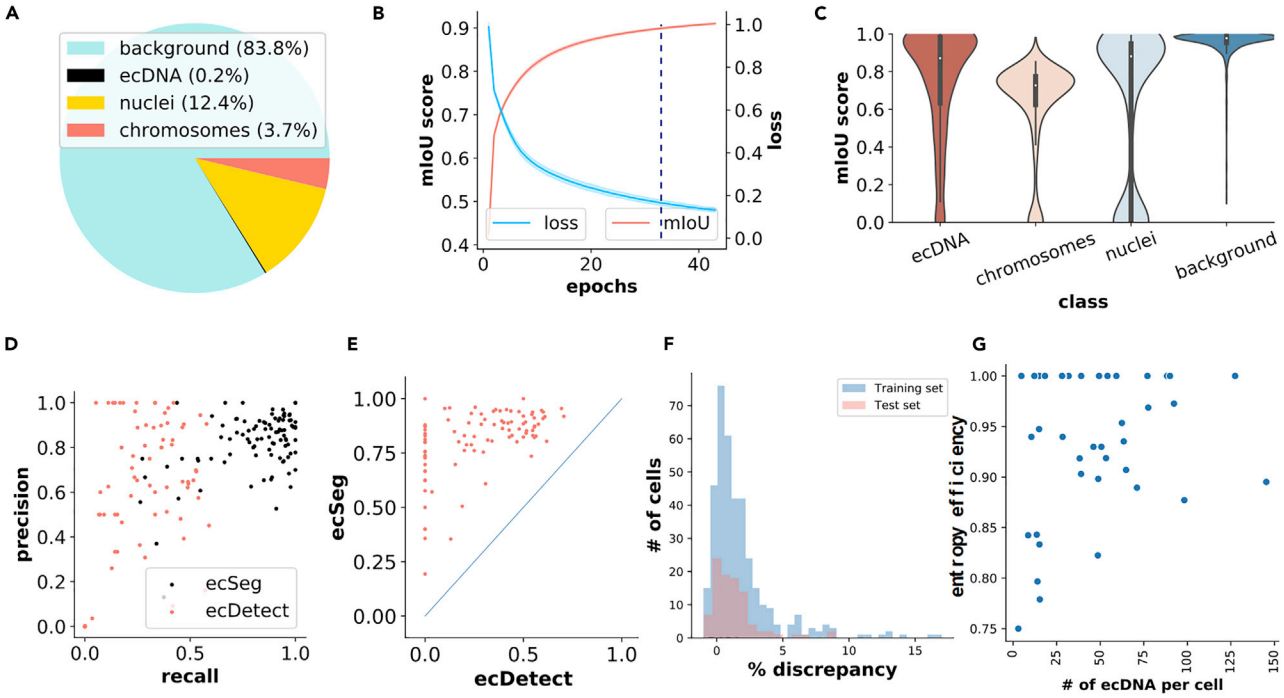


Figure 2. ecSeg Performance and Applications

(A) Pie chart showing class imbalance.

(B) Loss and mIoU on validation data as a function of training epochs. Only the loss function is used for training.

(C) mIoU score distributions for ecDNA, chromosomes, nuclei, and background on test data.

(D) Precision versus Recall for ecDetect and ecSeg on test data. Each point represents a complete image.

(E) F1 score comparison between ecSeg and ecDetect on test data. Notably, the ecDetect F1 scores rarely exceed 0.5 because of low recall, whereas ecSeg F1 scores are generally above 0.75.

(F) Distribution of Discrepancy in ecDNA counts ([ecSeg count - ground truth counts]/ground truth counts) shows a slight over-estimate for ecSeg, with 90% of the calls being within 5%.

(G) Entropy efficiency for 40 cell lines.

at a fine spatial resolution, as well as reasoning about categorical information based on global context, is necessary to successfully segment ecDNA. These dual goals can be achieved using U-nets (Ronneberger et al., 2015), a variant of FCNs, which gradually up-sample features and use skip connections to recover spatial resolution. U-nets have become widely recognized as the common choice of architecture in the medical image community for their superior performance on a number of imaging challenges (Litjens et al., 2017).

In this research, we developed ecSeg, a U-net-based platform (Figure 1F) for automatically classifying DAPI signal, identifying and quantifying ecDNA, and incorporating FISH data to clarify the location of oncogene amplification on ecDNA and chromosomes. It accepts DAPI and FISH-stained metaphase images and classifies each image pixel into one of the following classes: Cytoplasm, Nucleus, Chromosome, and ecDNA (Figure 1E, right panel). Subsequently, it computes connected components of ecDNA pixels (Transparent Methods) to demarcate and count ecDNA. When FISH probes are present, it quantifies their spatial location in a separate post-processing step and correlates those locations with ecDNA and chromosomes.

RESULTS

Network Training Procedure

To train ecSeg, we developed and made available a unique dataset containing ground truth labeling of nuclei, chromosomes, and ecDNA, starting from 483 unlabeled images of dimensions 1040 × 1392 from Turner et al. (2017). The ground truth labeling was created by multiple scientists involved in independent annotation (Transparent Methods). Owing to the difficulty of annotating 18.9K ecDNA across 483 images, a decision was made early on to use a coarsely annotated dataset that allowed for the possibility of a few

missed and/or false ecDNA calls. For the training and testing of learning frameworks, we generated 5,949 image patches (256 × 256 each) that were cropped from the larger images. We randomly split these patches into training (4,760 patches) and test (1,189 patches) datasets. Importantly, to have higher fidelity while testing the network, we further annotated the test data to reduce false ecDNA calls. The test data was a "holdout" set that was used only for final quantification of the model and had no direct effect on the training itself.

In training the network, we used a weighted loss function comprising the binary cross entropy and Dice coefficient to correct for the severe class imbalance. We also modified the architecture and adjusted hyper-parameters to account for the small size and lack of discriminating features on ecDNA (Transparent Methods). To optimize the model, various iterations of the architecture and hyper-parameters were trained using the training data on 8 GeForce GTX 1080 Ti GPUs (Transparent Methods). We also tested with different network architectures such as U-net with multi-scale context aggregation using dilated convolution (Yu and Koltun, 2015), pre-trained weights from VGG16 (Simonyan and Zisserman, 2014) trained on ImageNet, and the base U-net. The performance was optimized on a network with 32 filters in the first layer and doubling the number of filters in each layer, input image sizes of 256 × 256, and an L2 regularization parameter of 0.0001. For the optimal model, we found that the loss converged after 33 epochs (Figure 2B). As the loss function did not provide an intuitive explanation of performance, we additionally used a "mean Intersection over Union" (mIoU) score (Transparent Methods) to measure the fraction of true calls. The mIoU score showed similar convergence behavior on the training data (Figure 2B).

Test Set Segmentation Accuracy

On the test data (1,189 patches) ecSeg displayed good performance for each of the classes with mIoU scores of 0.75 for ecDNA, 0.68 for chromosomes, 0.78 for nuclei, and 0.97 for background (Figure 2C). Notably, 50% of the patches had an ecDNA mIoU score of at least 0.871 and 25% had a score of 0.938. The relatively worse performance for chromosomes was partially due to images in which the chromosomes are tightly clustered, making it difficult to differentiate them from intact nuclei (Figure S1).

Although we used a pixel-based image segmentation approach, the primary goal of ecSeg is to detect and count ecDNA in entire images. For example, an incorrect pixel classification adjacent to a correctly annotated ecDNA pixel does not change the fact that the ecDNA was detected. Therefore, ecSeg also post-processes the output by computing connected components of adjacent pixels with the same class label (Transparent Methods). We defined true-positive or TP (respectively, false-positive or FP) predictions as an ecDNA connected component whose centroid was within (respectively, outside) a pixel-distance threshold α (=5) of a manual annotation (see Transparent Methods). Similarly, we defined a false-negative (FN) call as a manual annotation with no ecSeg prediction within 5 pixels. On the test patches, the mean precision (TP/TP + FP) and recall (TP/TP + FN) were measured as 85% and 86%, respectively.

Comparison of Segmentation Methods

To compare against ecDetect predictions, we combined the predictions of all patches for an image. We plotted the precision versus recall performance of ecSeg for each image, along with the ecDetect predictions (Figure 2D, Table S2). At the image level, the mean precision and recall values were 82% each, in contrast with 59% and 23% achieved by ecDetect, which rarely achieved recall above 50%, and had a worse F1 (combined) score than ecSeg for each image (Figure 2E). The ecSeg performance varied across cell lines (Table S3). Thus COLO205, where the ecDNA are notably larger in the nine images (Figure S2) had worse performance (75% precision, 64% recall) compared with CA718 (84%, 90%). Moreover, in at least some cases, ecSeg predictions that did not match the manual annotation were in fact true calls as verified by external annotators who were not involved in the original annotation process. Similarly, a small number of manual annotations not called by ecSeg were truly not ecDNA (Figures S3 and S4). Including the totality of 483 training and test images, the number of ecDNA called by ecSeg were within 5% of the manual annotation calls in 88% of the images, validating the applicability of ecSeg in providing an accurate estimate of ecDNA abundance (Figure 2F).

ecDNA Heterogeneity

The ecDNA model of focal amplification (Deshpande et al., 2019; Verhaak et al., 2019) suggests that ecDNA segregate randomly into daughter cells, driving and maintaining intra-tumoral genetic heterogeneity of

ecDNA counts. For a sample with n metaphase images, let n_i denote the number of samples with exactly i ecDNA counts. The Shannon Entropy, measured using

$$\mathcal{H}_n = \sum_{i:n_i>0} -\frac{n_i}{n}\log_2\frac{n_i}{n},$$

showed large variation across different cell lines (Table S4). Noting that the entropy value depends on the number n of sampled cells (images), we also plotted the normalized *entropy-efficiency* value ($\mathcal{H}_n/\log_2 n$) for 40 cell lines. Interestingly, most (21 of 29) cell lines whose ecDNA copy numbers exceeded 10 per cell had an entropy efficiency above 90% (Figure 2G, Table S4), suggesting an important role for ecDNA in maintaining copy number heterogeneity.

Modeling the Effect of Environmental Changes (Drug Treatment) on ecDNA

Activated oncogenes on ecDNA can provide a selective advantage to cells with higher ecDNA counts, leading to rapid proliferation of those cells and focal amplification (Turner et al., 2017). However, environmental changes that restrict metabolite availability may impose a selective disadvantage on ecDNA-containing cells. Indeed, a previous report had shown a dramatic decrease of ecDNA in a glioblastoma cell line when targeted with the anti-EGFR drug Erlotinib (Eb), followed by a rapid increase in ecDNA upon withdrawal of drug treatment (Nathanson et al., 2014). To test the effect of drugs and other environmental factors in modulating ecDNA counts, we used ecSeg to quantify ecDNA counts in cells before Eb treatment and followed up 2 and 4 weeks after treatment.

To quantify the effect of drug treatment, we extended earlier work that modeled these selective forces using a Galton-Watson branching process (Bozic et al., 2010; Turner et al., 2017) (Transparent Methods), where each cell containing k ecDNA either replicates with probability b_k, or dies (probability $d_k = 1 - b_k$), to create the next generation. Positive selection was modeled by setting $b_k - d_k \propto f_{m,\alpha}(k)$, where

$$f_{m,\alpha}(k) = \begin{cases} \dfrac{k}{M_s} & (0 \le k \le M_s)\,, \\ \dfrac{1}{1+e^{\alpha(k-m)}} & (M_s < k < M_a). \end{cases} \qquad \text{(Equation 1)}$$

is positive, increasing for small values of k and decreasing logistically to 0 for larger values of k (Figure 3A, black line). To this model, we added the effect of a drug targeting the protein product of the oncogene by using f_k that logistically decreases to a negative value for increasing k (Figure 3A, blue line and Transparent Methods).

$$f_{r,\alpha}(k) = -\frac{e^{\alpha(k-r)}}{1+e^{\alpha(k-r)}}. \qquad \text{(Equation 2)}$$

Different choices of the decay parameters r, α all predicted a sharp decrease in ecDNA per cell, and a decrease in heterogeneity (Figure S5), but show very different rates of decrease in ecDNA.

On the experimental data, ecDNA counts, estimated by ecSeg, reduced dramatically from a mean of 50 per cell (median 26) at week 0 to 38 (median 14) at 2 weeks and 10 (median 1) at 4 weeks (Figures 3B and 3C, Table S5). The entropy efficiency of the cells changed from 0.98 at week 0 to 0.73 at week 4. The results closely matched simulations for $r = 20$, $\alpha = 0.04$. Although the theoretical models are admittedly simplistic, they showcase the power of ecSeg in inferring model parameters and providing quantitative comparisons of drugs used to target ecDNA.

Oncogene Amplification on Homologously Stained Regions and ecDNA

The tumor cell can respond rapidly to a changing environment by dynamically modulating RNA expression through ecDNA formation as well as reintegration of ecDNA as homologously stained regions (HSRs) (Nathanson et al., 2014). This is shown in the example of two glioblastoma cell lines where EGFR amplifications occur either as ecDNA ("ec" cell line) or as HSR ("hsr" cell line, Figures 3E and 3F). To quantify this phenomenon, we used an EGFR FISH probe and ecSeg analysis to locate EGFR (Transparent Methods) in the two cell lines. The median fraction of FISH signal explained by ecDNA was 0% in the hsr cell line but rose to 14% in the ec cell line (Figure 3G). In contrast, 71% (respectively, 15%) of the FISH signal was found on chromosomes in the hsr (respectively, ec) cell line. The results document the ability of ecSeg to provide

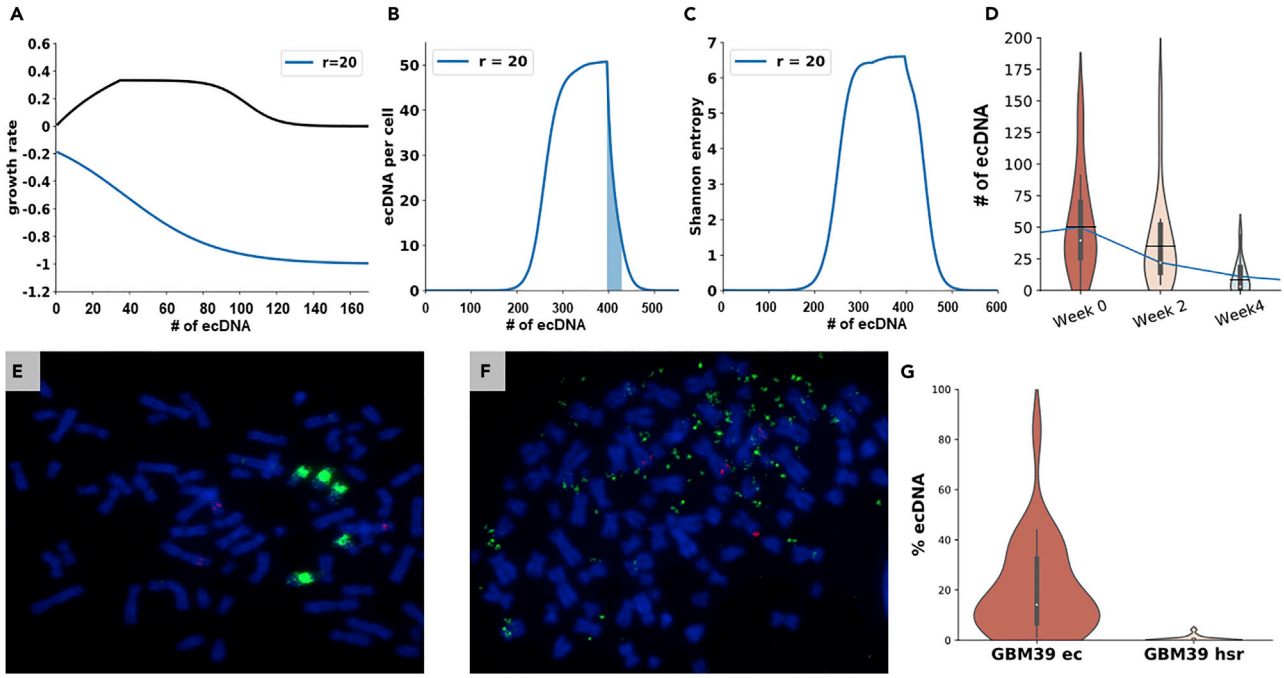


Figure 3. ecSeg Applications

(A) The black line shows the growth rate $b_k - d_k$ for ecDNA-driven amplification (parameters $\alpha = 0.1$, $m = 100$), which rises initially and slowly decreases to 0. The effect of a drug on growth rate (blue line)is modeled using a negative selection function $f_{r,\alpha}(k)$ for parameters $\alpha = 0.04$, $r = 20$.

(B and C) (B) Simulated changes in the mean copy number and Shannon entropy (C) as a function of time, when the drug is applied at day 400 with $\alpha = 0.04$, $r = 20$.

(D) Reduction of ecDNA counts in glioblastoma cell line GBM39 upon Erlotinib treatment. Black lines inside the violin plots show sample means, whereas white circles and box plots show the median and the middle 50th percentile. The blue line shows mean values of the simulation from (B) (shaded region). The mean, median counts per cell were (50,26) at week 0, (38,14) at week 2, and (10,1) at week 4, consistent with the theoretical model.

(E and F) (E) A glioblastoma cell line with EGFR proto-oncogene (stained using green FISH signal) found in homologously stained regions (HSR) (F) A glioblastoma cell line with EGFR found on ecDNA.

(G) Percentage of FISH signal on ecDNA in the ec and hsr cell lines.

insight into potentially important biological processes. Specifically, they suggest that ecDNA-driven amplifications, which are inherently capable of rapidly changing tumor copy number, can be "stabilized" by reintegrating into chromosomes, validating the prescient concept that ecDNA-based amplification (aka double minutes) is "unstable," whereas chromosomal amplification on HSRs is stable (Haber and Schimke, 1981).

DISCUSSION

The finding that ecDNA-based oncogene amplification is common in cancer raises some challenges for our current topological maps of cancer genes, including the fact that oncogene location within the nucleus could greatly impact tumor aggressiveness, as well as through non-chromosomal mechanisms of ecDNA inheritance. Nevertheless, it is difficult with existing genomic tools to quantify the extrachromosomal origin of copy number amplification. ecSeg provides a new tool for the research community to quantify ecDNA-based amplification at the single cell level.

FISH-based methods have been used to probe for oncogenes involved in tumor development, to identify cellular location of other proteins, including those involved in DNA repair, and for foci scoring (Verhaak et al., 2019; Nathanson et al., 2014). ecSeg can be used to determine the sub-cellular location of these proteins, helping to differentiate between intra-chromosomal and extrachromosomal repair mechanisms.

Genomic tools have been invaluable for precise measurements of copy number amplification, but bulk sequencing does not reveal the cell-to-cell variability in the copy number counts. Tools for quantifying

copy number heterogeneity are very limited as single-cell genomic analyses of copy number variation are often confounded by PCR-mediated artifacts. Automated cytogenetic analysis allows for an automated measurement of heterogeneity and understanding of its consequence. The ecDNA model of oncogene amplification suggests that ecDNA segregate independently into daughter cells and selection helps modulate a rapid change in copy number. An identical mechanism allows cells to rapidly reduce copy numbers under negative selection from a drug. ecSeg allows for the measurement of the rate of change and helps quantify the positive or negative selection strength. In summary, ecSeg can provide new insight into how cell-to-cell variability with respect to specific oncogenes contributes to tumor growth, progression, and drug resistance.

Limitations of the Study

We note that the network was trained on data from established cell lines, and the abundance and physical characteristics of ecDNA could be different in primary cancers, thus requiring additional fine-tuning. Most of the metaphase spreads were generated in a single laboratory. Differences in metaphase spread preparation could change the resolution of the input data.

Currently, the ecSeg method is trained only to analyze up to two FISH probes and will need to be updated to handle multiple probes.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.10.035.

ACKNOWLEDGMENTS

This research was supported in part by grants from the NSF (DBI-1458557) and from the NIH (R01GM114362).

AUTHOR CONTRIBUTIONS

U.R., V.B., and P.M. conceived the method and designed the experimental strategy to test it and wrote the manuscript. U.R., M.C., and V.B. designed the neural architecture and the computational methodology. K.T. and P.M. designed the experimental strategy. J.L. and V.D. performed the genomic reconstructions used in data analysis. All authors analyzed the results and reviewed the manuscript.

DECLARATION OF INTERESTS

V.B. is a co-founder of, serves on the scientific advisory board of, and has an equity interest in Boundless Bio, Inc. (BB) and Digital Proteomics, LLC (DP) and receives income from DP. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. BB and DP were not involved in the research presented here.

Received: July 23, 2019
Revised: October 3, 2019
Accepted: October 16, 2019
Published: November 22, 2019

REFERENCES

Beucher S. and Lantuejoul C. (1979). Use of watersheds in contour detection. In: International Workshop on Image Processing Real-Time Edge and Motion Detection Estimation, pp. , 17–21.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K.W., Vogelstein, B., and Nowak, M.A. (2010). Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. U S A *107*, 18545–18550.

Cox, D., Yuncken, C., and Spriggs, A.I. (1965). Minute chromatin bodies in malignant tumors of childhood. Lancet *286*, 55–58, https://doi.org/10.1126/science.aaf8399.

Davoli, T., Uno, H., Wooten, E.C., and Elledge, S.J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science *355*.

Deshpande, V., Luebeck, J., Nguyen, N.D., Bakhtiari, M., Turner, K.M., Schwab, R., Carter, H., Mischel, P.S., and Bafna, V. (2019). Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat. Commun. *10*, 392.

Garsed, D.W., Marshall, O.J., Corbin, V.D., Hsu, A., Di Stefano, L., Schröder, J., Li, J., Feng, Z.P., Kim, B.W., Kowarsky, M., et al. (2014). The architecture and evolution of cancer neochromosomes. Cancer Cell *26*, 653–667.

Haber, D.A., and Schimke, R.T. (1981). Unstable amplification of an altered dihydrofolate reductase gene associated with double-minute chromosomes. Cell *26* (3 Pt 1), 355–362.

Hamilton, B.A. (2018). Kaggle 2018 Data Science Bowl. https://www.kaggle.com/c/datascience-bowl-2018/overview.

Kitada, K., and Yamasaki, T. (2008). The complicated copy number alterations in chromosome 7 of a lung cancer cell line is explained by a model based on repeated breakage-fusionbridge cycles. Cancer Genet. Cytogenet. *185*, 11–19.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. *25*, 1097–1105.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., and Sánchez, C.I. (2017). A survey on deep learning in medical image analysis. Med. Image Anal. *42*, 60–88.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Ly, P., and Cleveland, D.W. (2017). Rebuilding chromosomes after catastrophe: emerging mechanisms of chromothripsis. Trends Cell Biol. *27*, 917–930.

Menghi, F., Barthel, F.P., Yadav, V., Tang, M., Ji, B., Tang, Z., Carter, G.W., Ruan, Y., Scully, R., Verhaak, R.G.W., et al. (2018). The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. Cancer Cell *34*, 197–210.e5.

Nathanson, D.A., Gini, B., Mottahedeh, J., Visnyei, K., Koga, T., Gomez, G., Eskin, A., Hwang, K., Wang, J., Masui, K., et al. (2014). Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. Science *343*, 72–76.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for largescale image recognition. In: International Conference on Learning Representations, pp. 1–14.

Turner, K.M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., Li, B., Arden, K., Ren, B., Nathanson, D.A., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature *543*, 122–125.

Verhaak, R.G.W., Bafna, V., and Mischel, P.S. (2019). Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat. Rev. Cancer *19*, 283–288.

Yu, F. and Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. In: International Conference on Learning Representations.

ISCI, Volume *21*

Supplemental Information

EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA

Utkrisht Rajkumar, Kristen Turner, Jens Luebeck, Viraj Deshpande, Manmohan Chandraker, Paul Mischel, and Vineet Bafna

Supplemental Table Titles and Legends

Supplemental Table 1: **Performance on different neural architectures.** The table reports (1) mIOU scores of ecDNA, chromosomes, nuclei, and cytoplasm, and (2) precision and recall scores for ecDNA for each variant of neural architecture tested. Related to Figure 2a, b, c.

Supplemental Table 2: **Precision and recall scores for ecSeg and ecDetect on entire data set.** The table reports (1) the precision and recall scores for ecSeg and ecDetect, (2) ground truth ecDNA counts per image, and (3) the predicted numbed of ecDNA from ecSeg for each of the 483 images in the data set. Related to Figure 2f.

Supplemental Table 3: **Performance on test data set.** Precision and recall scores from ecSeg and ecDetect for the 7 cell lines in the test set. Related to Figure 2d,e.

Supplemental Table 4: **Entropy.** The entropy and entropy efficiency for all cell lines present across the entire data set (training, validation, and test). Related to Figure 2g.

Supplemental Table 5: **Drug treatment ecDNA counts** Sheet 1 has raw ecDNA counts for both control and case for week 0,2, and 4. Sheet 2 has the entropy values for the cases in week 0, 2, and 4. Related to Figure 3d, e, f, g.

Transparent Methods

Data set. We started with a data set from (Turner et al. 2017). To capture relevant spatial information, cells were cultured according to standard protocol, and Karyomax was added to enrich for cells in metaphase. Cells were collected and treated with a 0.075 M KCl hypotonic solution for 10 minutes, followed by fixation in 3 : 1 methanol/glacial acetic acid solution. Interphase and mitotic cells were dropped onto humidified glass slides, and mounting medium with DAPI was applied to the slides. Cells in metaphase were imaged with an Olympus BX43 microscope equipped with a QiClick CCD camera. No 3D imaging was performed. Our dataset contains 483 images of dimensions 1392×1040 sampled from 27 different tumor cell lines. All images were stained with 4,6-diamidino-2-phenylindole (DAPI). DAPI is a blue-fluorescent stain that binds to any DNA structure represent in the sample. Thus, in our data set, it defines ecDNA, chromosomal, and nucleic regions. Some components in the image are also stained with fluorescence in situ hybridization (FISH) for specific probes on the ecDNA. However, we ignored the FISH signals when constructing our ground truth as (a) some ecDNA may not carry the probe target due to heterogeneity, and (b) not all targets are bound by the probe. Thus, extrachromosomal FISH signals validate ecDNA, but absence of FISH signals is not indicative of a lack of ecDNA.

We cropped these 483 images into 9,660 patches of 256×256. Some patches were purely background and we only included patches with at least 1% of the total area being covered in DAPI. We were left with 5949 usable patches. We split this data set such that 60% was used for training (3570 patches), 20% for validation (1190 patches), and the final 20% for reporting test results (1189 patches).

Ground Truth Labeling. Manual identification of ecDNA can be laborious as a single image can easily contain more than 200 ecDNA elements, sometimes up to 500. Thus, we built a software, using off-the-shelf morphological operations, to toggle a region as being ecDNA or not. The ground truth was then obtained through a manual annotation process using that software. To reduce the annotator's work, we seeded the process by providing ecDetect annotations which the annotator could then toggle on or off.

We used Otsu's thresholding to binarize the gray-scale image (**Otsu1979**). The adaptive method demarcated the nuclei and chromosomes, but the smaller and lower intensity ecDNA were marked as background. We smoothed the edges of the chromosomes and nuclei by performing an *open* operation, which is an erosion of the connected components followed by a dilation. We next used Bradley local thresholding (**Bradley2007**), an adaptive thresholding algorithm, to perform ecDNA annotations. Bradley local thresholding uses a sliding average filter and checks if the brightness of the center pixel is $T\%$ lower than the mean intensity of the pixels in the window. If it is lower, then the pixel is set to black or otherwise set to white. We used a window size of 3×3 pixels and a threshold value of $T = 3\%$. This allowed us to segment the image to a finer resolution with ecDNA predictions. We post-processed ecDNA segmentation by removing stray components that were less than 15 pixels in size, filling in any holes, removing spurs, and performing an *open*

operation on each of the connected components. Notably, the process missed many true ecDNA, but the coarse segmentation was useful for training the U-net.

However, for the 96 test set images (1189 patches), where we needed a more precise accounting of false negative and false positives, we used additional annotators who refined the predictions by manually examining each image and correcting any ecDNA that were falsely classified during the coarse annotation.

Segmentation. Inspired by the U-Net, we used a modified fully convolutional neural network presented in Fig. 1f. We optimized the architecture by performing grid search over the network's hyper-parameters. We varied the number of filters in the first layer ($\{16, 32, 64\}$), input patch sizes ($\{128^2, 256^2, 512^2\}$) and L2 regularization ($\{1, 0.1, 0.01, 0.001, 0.0001\}$). We applied multi-scale context aggregation using dilated convolution (Yu and Koltun 2015). We found that although the chromosomal IoU increased, the ecDNA precision and recall remained the same. We also experimented with pre-trained weights from VGG16 trained on ImageNet. However, because ImageNet contains images of everyday objects, our model had a more difficult time generalizing to the microscopy images. In each case, we minimized loss on the network variants using the Adam optimizer on 8 GeForce GTX 1080 Ti GPUs. We trained the network on the training set and used the validation set to evaluate loss and mIoU. The training was halted if the loss on the validation set did not change for 7 epochs (the 'patience' time). The test data was a "holdout" set that was only used for final quantification of the model and had no direct effect on the training itself. The performance was optimized on a network with 32 filters in the first layer and doubling the number of filters in each layer, input image sizes of 256×256, and a L2 regularization parameter of 0.0001.

We decided not to perform any data augmentation through warping and stretching. The relative size and shapes of ecDNA are very critical, and often times, certain ecDNA are almost the size of chromosomes, such as in the COLO205 cell line Supplementary Fig. 1. Any warping and stretching could cause the ecDNA and chromosomes to be indistinguishable even for the human eye. Rotations were not used either as our images have no rotational significance. All the images were taken from a top-down view with no bias towards orientation. Finally, as we collected data from a large number of cell lines, we had sufficient variation in our dataset.

We denoted each ground truth image as a collection of pixels $\mathcal{P}$ with the goal of classifying the pixels into one class from $\mathcal{C} = \{b, n, h, e\}$, representing background (b), nucleus (n), chromosome (h), and ecDNA (e). The ground truth was described by a binary function $y_c(x) \in \{0, 1\}$ for all $x \in \mathcal{P}$, $c \in \mathcal{C}$. Additionally, $\sum_c y_c(x) = 1$ for all pixels, enforcing a single class assignment. For each $x \in \mathcal{P}$, $c \in \mathcal{C}$, the network outputs a class score, $P_c(x) \in [0, 1]$. We trained the network to minimize a custom loss function defined below.

Loss function. We defined loss L as a weighted binary cross entropy (BCE) minus the Sørensen-Dice coefficient (Dice). Specifically, the BCE loss for class c was computed using:

$$\mathrm{BCE}[x] = -\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left[y_c(x) \ln \left(\frac{1}{1 + e^{-P_c(x)}} \right) + (1 - y_c(x)) \ln \left(1 - \frac{1}{1 + e^{-P_c(x)}} \right) \right].$$

Similarly, we compute Dice loss as:

$$\text{Dice} = \left[1 - \frac{2\sum_c \mathbf{P_c} \cdot \mathbf{y_c}}{\sum_c (\|\mathbf{P_c}\|_1 + \|\mathbf{y_c}\|_1)}\right]$$

We used weights to boost the under-represented classes. Let n_b, n_n, n_h, n_e denote the total number of pixels belonging to each class in background, nuclei, chromosome, and ecDNA, respectively, for the entire training and validation dataset. As $n_b >> n_n > n_h > n_e$, we assigned weight w_c to each class $c \in \{b, n, h, e\}$ as follows:

$$w_c = \max\left\{1, \frac{n_n}{n_c}\right\}$$

Correspondingly, the weight of a pixel was given by:

$$w_x = \sum_c y_c(x) w_c \tag{3}$$

and the net loss was computed using

$$L = \frac{1}{|\mathcal{P}|} \sum_x w_x(\text{BCE}[x] + \text{Dice})$$

To prevent over-fitting, we trained for 45 epochs with an early stopping "patience" of 7 which stopped training if the loss on the Validation set did not improve for 7 epochs.

Accuracy. For each class c, and threshold $\tau \in \mathcal{T}$, where $\mathcal{T} = \{0.05, 0.1, 0.5\}$, define an indicator $\theta_{c,\tau}(x) = \{1 \text{ if } P_c(x) \geq \tau; 0 \text{ otherwise}\}$. Define the mean Intersection over Union (mIoU) score across all classes as:

$$M = \frac{1}{|\mathcal{T}|} \sum_\tau \frac{1}{|\mathcal{C}|} \sum_c \frac{\boldsymbol{\theta}_{c,\tau} \cdot \mathbf{y_c}}{\|\boldsymbol{\theta}_{c,\tau}\|_1 + \|\mathbf{y_c}\|_1}$$

Post-processing of segmentation. Post-training, the network outputs a 256×256 matrix O, with

$$O[x] = \arg\max_c P_c(x)$$

To filter noise, we computed connected components for each class. Connected components are regions of adjacent pixels with the same class value. We filled all holes in each of the connected components such that the hole is assigned the same class as the surrounding pixels. We performed secondary size thresholding on the ecDNA elements such that all ecDNA components less than 15 pixels are marked as background and those greater than 125 pixels are marked as chromosomes. We also removed any ecDNA that were attached to the edges of chromosomes or nuclei as these regions are often just spurs of the larger class.

Accuracy Metrics. To compute component level accuracy, we computed true positive, false positive, and false negative rates. If the centroid of a predicted ecDNA component was within a 5 pixel euclidean distance of the centroid of a ground truth ecDNA component, we marked this as

a true positive (TP). If there are no ground truth ecDNAs within that distance, we classified the component as a false positive (FP). We found that the average area of ecDNA across our entire dataset was 75 pixels and thus a distance threshold of $\sqrt{75/\pi} \simeq 5$ pixels ensures that ecDNAs detected on the periphery of the boundary from the annotated center pixel is still considered a true positive. Inversely, if there were no predicted ecDNAs within a 5 pixel distance of a ground truth annotation, we classified it as a false negative (FN). We compute our precision and recall for each image as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

We also measured accuracy using the F1 score, a harmonic average of precision and recall.

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Entropy and Entropy efficiency. Consider a sample with n cells. Let n_i (respectively $p_i = \frac{n_i}{n}$ denote the number (respectively, fraction) of cells with i copies. We defined heterogeneity of copy number using Shannon entropy:

$$\mathcal{H}_n = -\sum_i p_i \log_2 p_i,$$

The entropy efficiency, defined by $\frac{\mathcal{H}_n}{\log_2 n}$ normalizes the value between 0 (no heterogeneity) and 1 (maximum heterogeneity).

Drug Treatment Quantification. We cultured GBM39 cells as neurospheres under serum-free conditions (DMEM/F12 basal media with 1X Glutamax, EGF, FGF, and heparin). Cells were cultured in 5 uM Erlotinib. The EGFR-containing ecDNA was quantified via ecSeg at 0, 2, and 4 weeks.

Evolutionary model for ecDNA driven copy number. Consider an initial population of cells, with each cell carrying $k \geq 0$ copies of an oncogene on ecDNA. We modeled the population using a discrete generation Galton-Watson branching process (Bozic et al. 2010). 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

$$\frac{b_k}{d_k} = \begin{cases} 1 + f_{m,\alpha}(k), & 0 \leq k < M_a \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$d_k = 1 - b_k \tag{5}$$

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 ($b_k - d_k \propto f_{m,\alpha}(k)$) is governed by parameters m, α. Initially, the selective advantage

increases with increasing copies, but later diminishes due to increasing metabolic load. We modeled this by defining

$$f_{m,\alpha}(k) = \begin{cases} \frac{k}{M_s} & (0 \leq k \leq M_s)\ , \\ \frac{1}{1+e^{\alpha(k-m)}} & (M_s < k < M_a). \end{cases} \tag{6}$$

Here, parameters m and α are the 'mid-point', and 'steepness' parameters of the logistic function, respectively. Initially, $f_{m,\alpha}(k)$ grows linearly, reaching a peak value of $f_{m,\alpha}(k) = 1$ for $k = M_s$. As the viability of cells with large number of amplicons is limited by available metabolites (**Pavlova2016**), $f_{m,\alpha}(k)$ decreases logistically in value for $k > M_s$ reaching $f_{m,\alpha}(k) \to 0$ for $k \geq M_a$. We model the decrease by a sigmoid function with a single mid-point parameter m s.t. $f_{m,\alpha}(m) = \frac{1}{2}$. The 'steepness' parameter α is automatically adjusted to ensure that $\max\{1 - f_{m,\alpha}(M_s), f_{m,\alpha}(M_a)\} \to 0$. We empirically chose $M_a = 20, m = 100, \alpha = 0.1$ to match a mean copy number of 50 ecDNA per cell observed prior to drug treatment.

The addition of a drug targeting the oncogene provides a disadvantage (negative fitness) to cells carrying extra copies of the oncogene. Therefore, after drug treatment, we used the selective function

$$f_{r,\alpha}(k) = -\frac{e^{\alpha(k-r)}}{1 + e^{\alpha(k-r)}}. \tag{7}$$

$f_{r,\alpha}(k)$ provides negative selection pressure causing a steep decline in the average number of ecDNA per cell. We simulated the effect of the drug using $r \in \{5, 20, 50, 100\}$, $\alpha \in \{0.07, 0.04, 0.03\}$. Supplementary Figure 5 shows the values for $\alpha = 0.04$. We observed that $r = 20, \alpha = 0.04$ best matched the empirical observations with Eb treatment (Figure 3d).

FISH analysis. ecSeg also incorporates FISH analysis. It allows the user to specify the color of the FISH signal used to illuminate the gene of interest and the intensity threshold T ($T = 120$ by default). It then extracts binary images highlighting only the pixels that have the minimum intensity in the appropriate color channel and additionally marks the pixels as either ecDNA or chromosomes. ecSeg outputs a table containing the total number of FISH pixels, the fraction of FISH pixels that are also marked as ecDNA, and the fraction marked as chromosomal for each image in the user-specified file path.

DATA AND SOFTWARE AVAILABILITY

ecSeg is available at https://github.com/ucrajkumar/ecSeg. The accession number for the data reported in this paper is 10.17632 : m7n3zvg539.3.

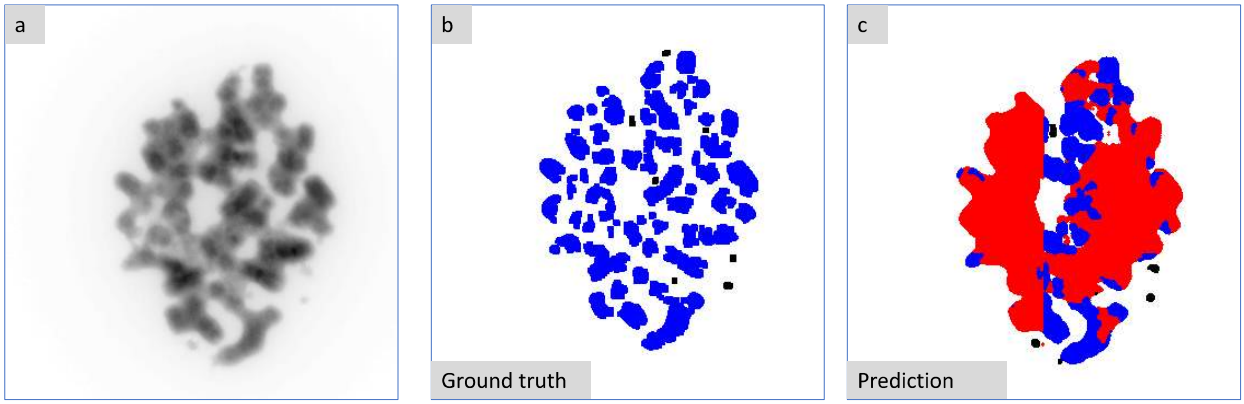
Supplemental References

1. Bradley, D., Roth, G., 2007. Adaptive thresholding using the integral image. J. Graph. Tools 12, 1321.

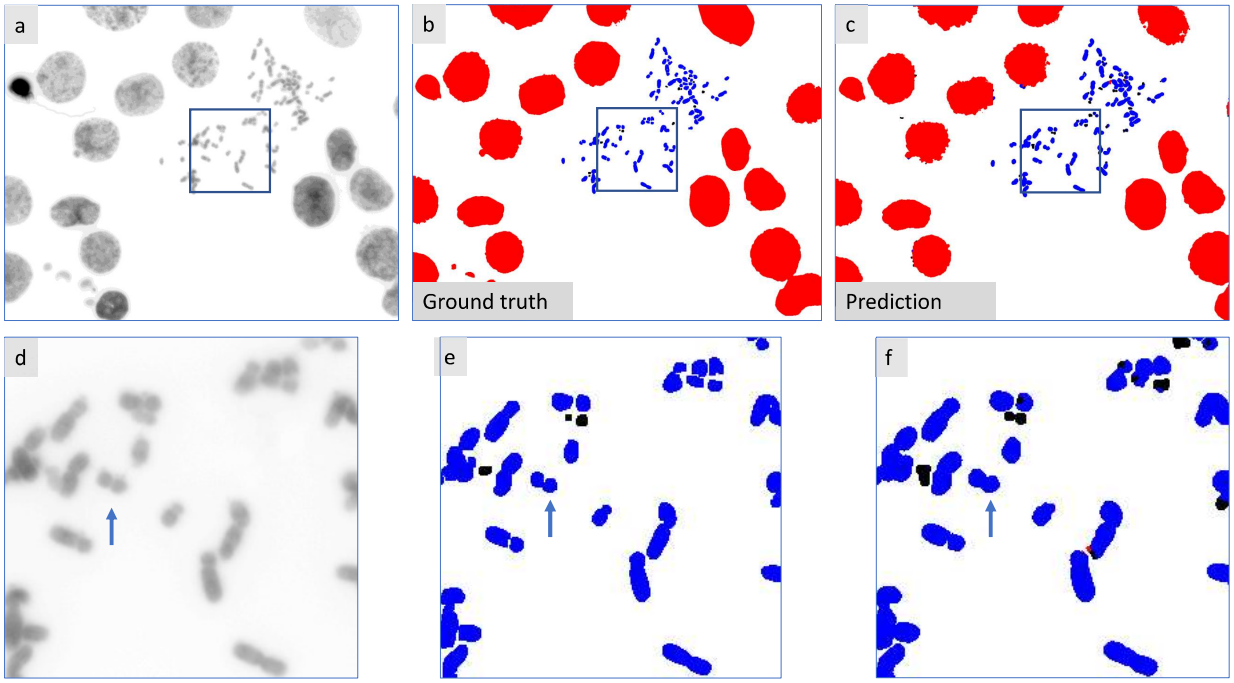
2. Otsu N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Sys. Man. Cyber. 9 (1): 6266.

3. Pavlova, N.N., Thompson, C.B., 2016. The emerging hallmarks of cancer metabolism. Cell Metab. 23, 2747.

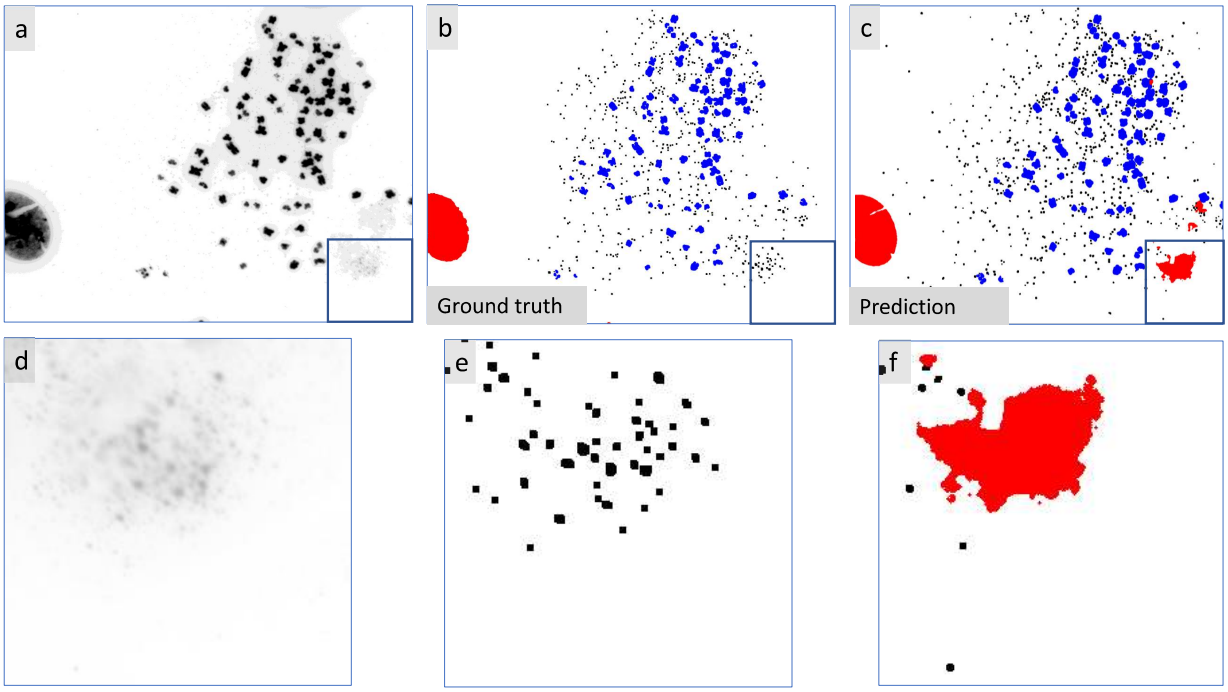
Supplemental Figures



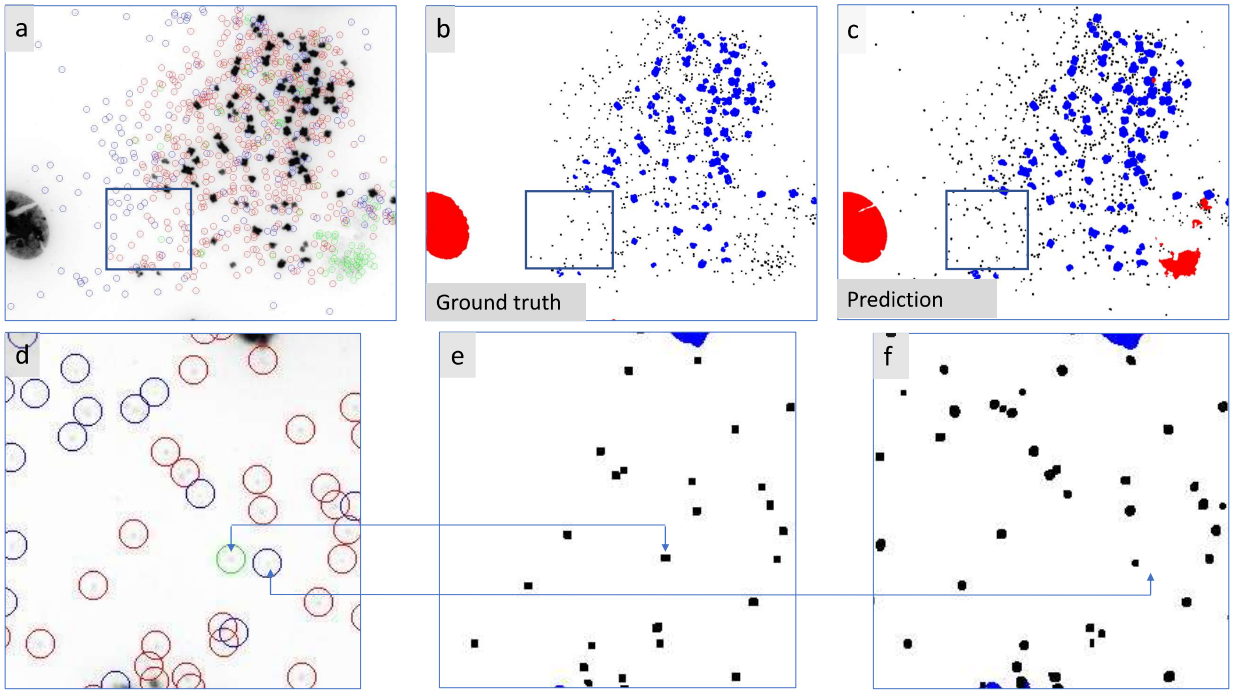
Supplemental Figure 1: **Incorrect classification of chromosomes as nuclei in COLO205.** (a) DAPI of original image from cell line COLO205. (b) Ground truth annotation with intact nuclei, chromosomes, and ecDNA being represented by red, blue, and black, respectively. (c) Segmentation map. COLO205 tumor cell remain tightly clumped even after the nucleic membrane has disintegrated. The network mis-classifies these chromosomes as nuclei due to the tight clustering. Related to main Fig. 2b.



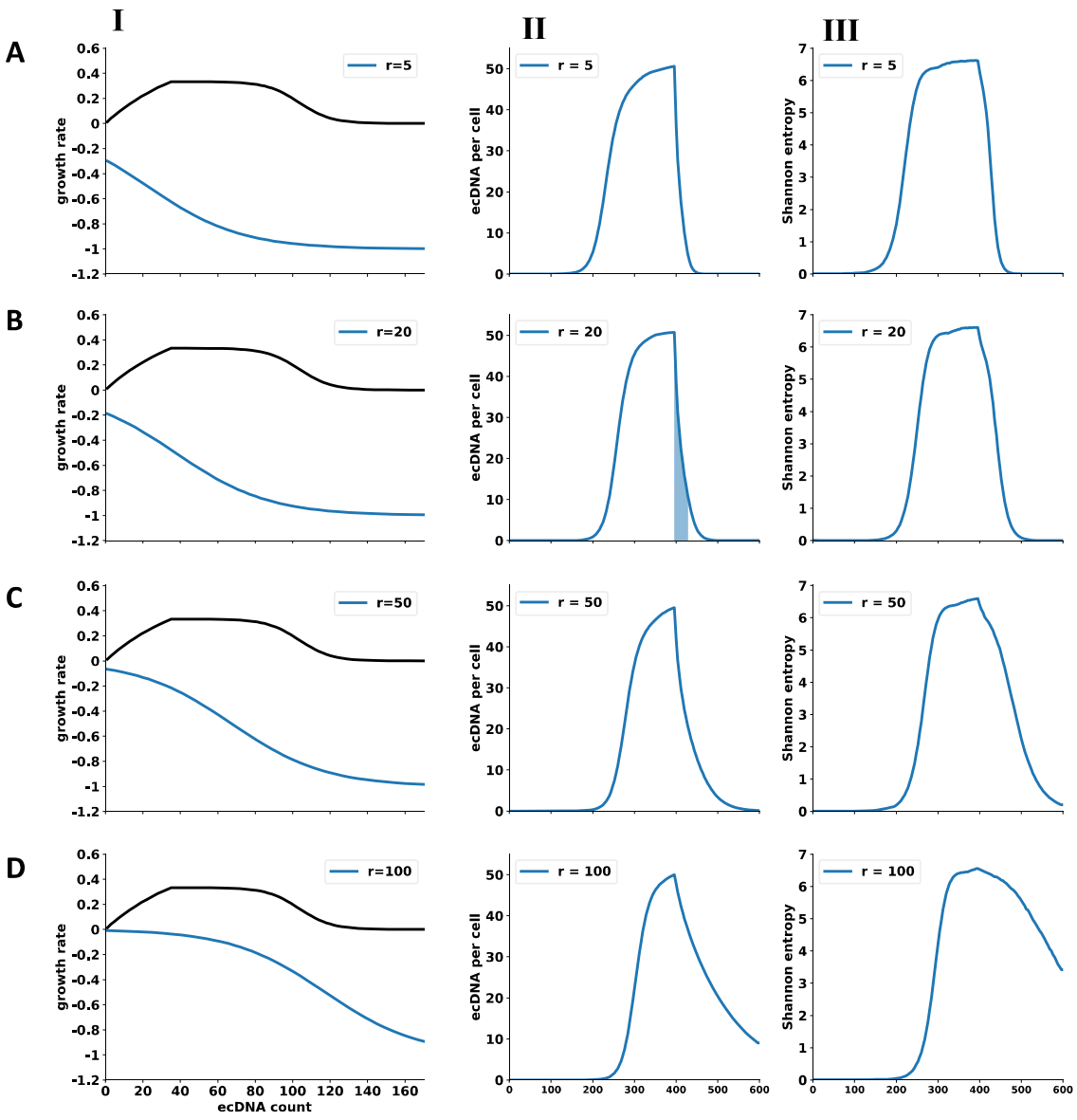
Supplemental Figure 2: **Incorrect detection of large ecDNA in COLO205.** (a) DAPI of original image from cell-line COLO205. (b) 'Ground truth' annotation with intact nuclei, chromosomes, and ecDNA being represented by red, blue, and black, respectively. (c) ecSeg Segmentation map. (d,e,f) Crops of DAPI, ground truth annotation, and ecSeg segmentation. In COLO205, replicating ecDNA structures (double minutes) often closely resemble chromosomes, making it difficult to identify. These structures are marked as chromosomes in both the ground truth and the segmentation map. Related to main Fig. 2b.



Supplemental Figure 3: **Incorrect false negative calls in cell line CA718.** (f) is burst nucleus, but appears to show as ecDNA when zoomed in, and was marked as ecDNA during human annotation. ecSeg correctly annotates it as a nucleus identifying a mistake in the human annotation. Related to main Fig. 2b.



Supplemental Figure 4: **Incorrect annotation of ecDNA in cell line CA718.** (a) DAPI of original image from cell-line CA718. (b) 'Ground truth' annotation (c) ecSeg Segmentation map (d,e,f) Crops of DAPI, ground truth annotation, and ecSeg segmentation. Blue circles denote false positives, red circles are true positives, and green circles are false negatives. As can be verified by looking at the DAPI image, many of the annotated false positives are actually true ecDNA with low-intensity DAPI signals. These ecDNA were missed during the ground truth annotation. False negatives are rare, and often indicate a problem with the ground truth annotation, as shown in Supplementary Fig. 3. Related to main Fig. 2d.



Supplemental Figure 5: **Simulating the impact of drug on ecDNA counts and heterogeneity.** Column I shows the modeled growth rates $b_k - d_k$ as a function of ecDNA count (k) for untreated (black line) and drug-treated (blue) lines, for $\alpha = 0.04$, and $r \in \{5, 20, 50, 100\}$ (rows A-D). Columns II and III show simulated changes in the mean copy number and Shannon entropy as a function of time, when the drug is applied at day 400. Upon drug application, the ecDNA counts and heterogeneity both decline in a manner dependent upon the the strength of selection modeled using α, r. Panel B.II ($r = 20, \alpha = 0.04$; shaded region) best fit the experimental data of GBM cells treated with Erltonib (related to main Figure 2h).