UNIVERSITY OF CALIFORNIA SAN DIEGO

Deep Learning in Cancer Biology

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Computer Science

by

Utkrisht Rajkumar

Committee in charge:

Professor Vineet Bafna, Chair
Professor Jingbo Shang, Co-Chair
Professor Taylor Berg-Kirkpatrick
Professor Melissa Gymrek
Professor Paul Mischel
Professor Bing Ren

2022

The dissertation of Utkrisht Rajkumar is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

Dedicated to my parents for their constant love, support, and encouragement
as I pursued my dreams.

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

DEDICATION iv

TABLE OF CONTENTS v

LIST OF FIGURES vii

LIST OF SUPPLEMENTAL TABLES viii

ACKNOWLEDGEMENTS ix

VITA xi

ABSTRACT OF THE DISSERTATION xiii

INTRODUCTION 1

CHAPTER 1. EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA 3

- 1.1 Introduction 3
- 1.2 Results 6
  - 1.2.1 Network Training Procedure 6
  - 1.2.2 Test Set Segmentation Accuracy 8
  - 1.2.3 Comparison of Segmentation Methods 9
  - 1.2.4 ecDNA Heterogeneity 9
  - 1.2.5 Modeling the Effect of Environmental Changes (Drug Treatment) on ecDNA 10
  - 1.2.6 Oncogene Amplification on HSR and ecDNA 12
- 1.3 Discussion 12
- 1.4 Methods 14
  - 1.4.1 Data set 14
  - 1.4.2 Ground Truth Labeling 15
  - 1.4.3 Segmentation 16
  - 1.4.4 Loss Function 17
  - 1.4.5 Accuracy 18
  - 1.4.6 Post-processing of Segmentation 19
  - 1.4.7 Accuracy Metrics 19
  - 1.4.8 Entropy and Entropy Efficiency 20
  - 1.4.9 Drug Treatment Quantification 21
  - 1.4.10 Evolutionary Model for ecDNA Driven Copy Number 21
  - 1.4.11 FISH Analysis 23
- 1.5 Acknowledgements 23
- 1.6 Appendix 24

CHAPTER 2. Pan-Cancer Analysis of Oncogene Amplification in Interphase Cells . 35

- 2.1 Introduction 35

2.2 Methods...36
2.2.1 Dataset Overview...36
2.2.2 Ground Truth Labeling 37
2.2.3 Training and Test Split 37
2.2.4 Pre-process..38
2.2.5 Architecture..38
2.2.6 Training Procedure 39
2.2.7 Image Acquisition Protocols..39
2.3 Results..41
2.3.1 Test set Accuracy 41
2.3.2 Cell Line Analysis...42
2.3.3 Cell Line Evolution 42
2.3.4 Cell-line Amplification Analysis 43
2.3.5 Quantifying Multiple Oncogenes in a Single Cell 44
2.3.6 Patient Tissue Results 45
2.4 Discussion and Conclusion..45
2.5 Acknowledgements 46
2.6 Appendix...47
CHAPTER 3. DeepViFi: Detecting Oncoviral Infections in Cancer Genomes using Transformers...55
3.1 Introduction...55
3.2 Related Works 57
3.3 Method..58
3.3.1 Overview 58
3.3.2 Method Details 58
3.4 Training and Inference 60
3.4 Dataset Generation...62
3.4.1 Training Set...62
3.4.2 Test Set...63
3.4.3 HPV Mediated Primary Oropharyngeal Cancer Samples..........64
3.5 Results..64
3.5.1 Method Comparisons 64
3.5.2 Qualitative Analysis..66
3.5.3 Detecting HPV in Oropharyngeal tumor samples......................66
3.5.4 Detecting HBV in tumor samples. 67
3.6 Discussion and Conclusion..68
3.7 Acknowledgements 69
3.8 Appendix...70
REFERENCES...74

LIST OF FIGURES

Figure 1: Detecting ecDNA in DAPI-Stained Images 24

Figure 2: ecSeg performance and applications 26

Figure 3: ecSeg applications 28

Figure 4: Incorrect classification of chromosomes as nuclei in COLO205 29

Figure 5: Incorrect detection of large ecDNA in COLO205 30

Figure 6: Incorrect false negative calls in cell line CA718 31

Figure 7: Incorrect annotation of ecDNA in cell line CA718 32

Figure 8: Simulating the impact of drug on ecDNA counts and heterogeneity 33

Figure 9: Data Overview 47

Figure 10: ecSeg-i Pipeline 49

Figure 11: Test set Accuracy 50

Figure 12: Amplification Heterogeneity 52

Figure 13: Patient Tissue Results 54

Figure 14: DeepViFi pipeline 70

Figure 15: Test Set Performance and Analysis 71

Figure 16: Tumour Sample Analysis 72

Figure 17: Detailed architecture of DeepViFi transformer 73

LIST OF SUPPLEMENTAL TABLES

Table 1: Performance on different neural architectures 34

Table 2: Precision and recall scores for ecSeg and ecDetect on entire data set 34

Table 3: Performance on test data set 34

Table 4: Entropy... 34

Table 5: Drug treatment ecDNA counts 34

ACKNOWLEDGEMENTS

I first and foremost thank my advisor, Professor Vineet Bafna, for accepting me as his student. Although I often felt like a fish out of water, his encouragement and faith motivated me to persevere. His calm and kind mentorship not only made me a better scientist but also a better human.

I also thank my thesis committee members for guiding me through this difficult process. I thank committee co-chair, Professor Jingbo Shang, for always being approachable and making time to impart his wisdom in natural language processing. I thank Professor Paul Mischel for introducing me to an exciting new world of cancer biology which motivates the first two chapters of this thesis. I also deeply thank committee members Melissa Gymrek, Bing Ren, and Taylor Berg-Kirkpatrick for their insightful feedback and advise.

I am very grateful to my undergraduate advisors, professors Young-Han Kim and Ronald Graham, for their tutelage and motivating me to pursue a PhD. They were my first exposure to the world of research, and it was the best start I could have asked for.

I thank my lab mates for their help, advice, laughs, and memories. Particularly, I thank Nam Nguyen and Jens Luebeck for always answering my endless questions, Sara Javadzadeh for helping me on various projects, and Niema Moshiri for his advice on how to navigate academia.

I am also very grateful to members of Professor Paul Michel and Bing Ren’s labs for being amazing collaborators. They constantly provided motivation, knowledge, and data, for many of the projects throughout my PhD.

Chapter 1, in full, is a reprint of the material “EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA” by Utkrisht Rajkumar, Kristen Turner, Jens Luebeck, Viraj Deshpande, Manmohan Chandraker, Paul Mischel, and Vineet Bafna as it appears in IScience, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full is currently being prepared for submission for publication of the material “Pan-Cancer Analysis of Oncogene Amplification in Interphase Cells” by Utkrisht Rajkumar, Ellis J Curtis, Kristen Turner, Sihan Wu, Justin Wahl, Ivy Wong, Jun Tang, Homa Hemmati, Shailaja Kasibhatla, Paul Mischel, and Vineet Bafna. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is a reprint of the material “DeepViFi: Detecting Oncoviral Infections in Cancer Genomes using Transformers” by Utkrisht Rajkumar, Sara Javadzadeh, Mihir Bafna, Dongxia Wu, Rose Yu, Jingbo Shang, and Vineet Bafna as it appears in ACM BCB, 2022. The dissertation author was the primary investigator and author of this paper.

VITA

2014 – 2017 University of California San Diego
Bachelor of Science, Computer Engineering

2021 University of California San Diego
Master of Science, Computer Science

2018 – 2022 University of California San Diego
Doctor of Philosophy, Computer Science

PUBLICATIONS

U. Rajkumar, E. J. Curtis, K. Turner, S. Wu, J. Wahl, I. Wong, J. Tang, H. Hemmati, S. Kasibhatla, P. Mischel, V. Bafna, EcSeg-i: Pan-cancer Analysis of Oncogene Amplification in Interphase Cells. In preparation.

J. T. Lange, C. Y. Chen, Y. Pichugin, L. Xie, J. Tang, K. L. Hung, K. E. Yost, Q. Shi, M. L. Erb, U. Rajkumar, S. Wu, C. Swanton, Z. Liu, W. Huang, H. Y. Chang, V. Bafna, A. G. Henssen, B. Werner, P. S. Mischel, Principles of ecDNA random inheritance drive rapid genome change and therapy resistance in human cancers. Nature Genetics, (2022).

U. Rajkumar, S. Javadzadeh, M. Bafna, D. Wu, R. Yu, J. Shang, V. Bafna, DeepViFi: Detecting Oncoviral Infections in Cancer Genomes using Transformers. Proceedings of the 13th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, (2022).

S. Javadzadeh, U. Rajkumar, N. Nguyen, S. Sarmashghi, J. Luebeck, J. Shang, V. Bafna, FastViFi: Fast and accurate detection of (Hybrid) Viral DNA and RNA. NAR Genomics and Bioinformatics, 4(2), lqac032, (2022).

K. L. Hung, K. E. Yost, L. Xie, Q. Shi, K. Helmsauer, J. Luebeck, R. Schöpflin, J. T. Lange, R. Chamorro, N. E. Weiser, C. Chen, M. E. Valieva, I. T.-L. Wong, S. Wu, S. R. Dehkordi, C. V. Duffy, K. Kraft, J. Tang, J. A. Belk, J. C. Rose, M. R. Corces, J. M. Granja, R. Li, U. Rajkumar, J. Friedlein, A. Bagchi, A. T. Satpathy, R. Tjian, S. Mundlos, V. Bafna, A. G. Henssen, P. S. Mischel, Z. Liu, H. Y. Chang, ecDNA hubs drive cooperative intermolecular oncogene expression. Nature, 600(7890), 731–736, (2021).

K. Fujimoto, T. Mizugaki, U. Rajkumar, H. Shigeta, S. Seno, Y. Uchida, M. Ishii, V. Bafna, H. Matsuda, A CNN-Based Cell Tracking Method for Multi-Slice Intravital Imaging Data. Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, (2021).

J. Luebeck, C. Coruh, S. R. Dehkordi, J. T. Lange, K. M. Turner, V. Deshpande, D. A. Pai, C. Zhang, U. Rajkumar, J. A. Law, P. S. Mischel, V. Bafna, AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications. Nature Communications, 11(1), 4374, (2020).

H. Kim, N.-P. Nguyen, K. Turner, S. Wu, A. D. Gujar, J. Luebeck, J. Liu, V. Deshpande, U. Rajkumar, S. Namburi, S. B. Amin, E. Yi, F. Menghi, J. H. Schulte, A. G. Henssen, H. Y. Chang, C. R. Beck, P. S. Mischel, V. Bafna, R. G. W. Verhaak, Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nature Genetics, 52(9), 891–897, (2020).

U. Rajkumar, K. Turner, J. Luebeck, V. Deshpande, M. Chandraker, P. Mischel, V. Bafna, EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA. iScience, 21, 428–435, (2019).

S. Wu, K. M. Turner, N. Nguyen, R. Raviram, M. Erb, J. Santini, J. Luebeck, U. Rajkumar, Y. Diao, B. Li, W. Zhang, N. Jameson, M. R. Corces, J. M. Granja, X. Chen, C. Coruh, A. Abnousi, J. Houston, Z. Ye, R. Hu, M. Yu, H. Kim, J. A. Law, R. G. W. Verhaak, M. Hu, F. B. Furnari, H. Y. Chang, B. Ren, V. Bafna, P. S. Mischel, Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature, 575(7784), 699–703, (2019).

S. Chowdhry, C. Zanca, U. Rajkumar, T. Koga, Y. Diao, R. Raviram, F. Liu, K. Turner, H. Yang, E. Brunk, J. Bi, F. Furnari, V. Bafna, B. Ren, P. S. Mischel, NAD metabolic dependency in cancer is shaped by gene amplification and enhancer remodelling. Nature, 569(7757), 570–575, (2019).

N. Alon, S. Butler, R. Graham, U. Rajkumar, Permutations Resilient to Deletions. Annals of Combinatorics, 22(4), 673–680, (2018)

ABSTRACT OF THE DISSERTATION

Deep Learning in Cancer Biology

by

Utkrisht Rajkumar

Doctor of Philosophy in Computer Science

University of California San Diego, 2022

Professor Vineet Bafna, Chair
Professor Jingbo Shang, Co-Chair

Deep learning methods have significantly advanced the state of computer vision and natural language processing. Their ability to discover intricate patterns in ever-expanding datasets is critical in solving cancer biology problems. However, cancer biology poses unique challenges. Typical input data, such as tumor images and DNA sequences, have significantly different semantic contexts than the traditional datasets used to train the deep learning methods. Thus, it is infeasible to leverage large pre-trained networks and requires training from scratch. Moreover, these data types are not human readable, making it difficult to annotate the data and interpret what the model has learned. This thesis aims to resolve these challenges and solve three urgent cancer biology problems using deep learning methods.

Cancer is mediated through various mechanisms. One such mechanism is circular extrachromosomal DNA (ecDNA), one of the primary drivers of oncogene amplification. EcDNA is prevalent across a wide variety of cancer types and leads to worse patient survival. Thus, there is a critical need for tools to study these genomic lesions. However, it is difficult to understand various facets of ecDNA just through sequence-based methods and requires image-based reconstructions.

I first present ecSeg, a deep learning tool to reconstruct ecDNA in images of tumor cells in *metaphase*. EcSeg uses a fully convolutional network and traditional computer vision techniques to semantically segment ecDNA. EcSeg correlates these segmentations with amplification profiles to reveal ecDNA mechanics and their resistance to drug therapy.

To translate ecSeg to clinical practice, I present ecSeg-i to resolve the ecDNA status of *interphase* cells in cancer patient tissue. Tissue images primarily contain interphase cells in which the DNA is loosely wound, making it extremely challenging to distinguish ecDNA. EcSeg-i uses a DenseNet to determine the ecDNA status and amplification profiles of cancer patient tissue.

Lastly, I present DeepViFi to identify oncoviral infections in cancer genomes. Rapidly mutating oncoviruses, such as HPV, can infect the host and disrupt various

biological pathways, sometimes causing hybrid human-viral ecDNA to appear. DeepViFi is a transformer-based tool which uses an openset framework to embed DNA reads and detect oncoviral infections in next-generation sequencing data.

INTRODUCTION

Deep learning methods permeate various fields of study from predicting meteorological events using satellite images to determining the 3D structure of proteins using their amino acid sequences. They can discover complex and often unknown patterns by leveraging ever-expanding datasets. They are composed of simple nonlinear functions that gradually transform the raw input into more abstract representations. With enough layers of transformation, these methods can learn remarkably complex representations.

These qualities are desirable in solving various problems in cancer biology. In this thesis, I investigate cancer biology through two lenses: *imaging* using Computer Vision (CV) techniques and *molecular genetics* using natural language processing (NLP) techniques. In the first two chapters of my thesis, I study ecDNA, a key driver of oncogene amplification, in fluorescently stained *images* of tumor cells using CV. In the concluding chapter of my thesis, I study oncoviral infections through short-read *DNA sequencing* data using NLP.

EcDNA are primarily found in tumor cells and are known to cause poor patient outcomes. Due to limitations in sequence-based methods, the primary way to distinguish ecDNA is through fluorescently stained images of tumor cells. EcDNA in images of tumor cells in metaphase appear as hundreds, sometimes thousands, of tiny faint DNA particles that are easily confounded with salt-and-pepper noise. It is even more difficult to discern ecDNA in images of cells in interphase when the DNA is inside intact nuclei.

Thus, marking true ecDNA requires extremely laborious expert annotations. Secondly, existing deep learning methods cannot be leveraged in this setting, because they are typically trained on everyday images such as hand-written numbers and pets. To overcomes these challenges, we built deep learning-based CV tools, trained from scratch on uniquely crafted datasets, to investigate ecDNA.

In a parallel vein, oncoviruses are a family of viruses that are known to mediate various cancers. Rapidly mutating oncoviruses, such as the human papillomaviruses (HPV), can infect the host and disrupt biological pathways, sometimes causing hybrid human-viral ecDNA to appear. Thus, it is essential to efficiently detect oncoviral infections. When the viral family is known, specific sequences can be probed directly by searching databases of known viral sequences. However, if the viral family shows high divergence between members, detection based on direct sequence match can fail. Traditional deep learning sequence tools are unsuitable in this setting as they are typically trained on human language data which have significantly different semantic contexts than DNA sequences composed of just four characters: A, T, C, G. Thus, we built a deep learning-based NLP tool to understand DNA contexts and identify oncoviral reads belonging to diverged viral families.

CHAPTER 1. EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA

1.1 Introduction

Despite the well-recognized importance of oncogene amplification in cancer pathogenesis [1], 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 [2], breakage fusion breakage cycles [3], aneuploidies [1], chromothripsis [4], and neochromosome formation [5] 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 [6, 7] 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 [8], referred to as double minutes (DMs), but the difficulty in linking these observations with modern cancer genomics led to a massive underestimation of their prevalence [6]. 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 [9] 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 [10] while maintaining their structural features. For example, focal amplifications containing EGFR have identical structures but can be extrachromosomal or integrated into non-native locations within chromosomes (Figure 1a-1d). Therefore, sequence-based reconstructions have limited power in revealing the spatial location of focal amplifications.

The use of 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 enable 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 to chromosomes), present challenges 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 [11] detected only 14 of 25 ecDNA. An earlier tool, ecDetect [7], 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 out 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 (CNNs) have consistently outperformed traditional image processing algorithms on biological data sets [12]. By using a large number of learnable parameters, they can recognize complex underlying patterns in large data sets [13]. In contrast to image classification tasks, fully convolutional networks (FCNs) were constructed to perform pixel-wise classification [14]. 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 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 [15], 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 [16].

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 (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.

1.2 Results

1.2.1 Network Training Procedure

To train ecSeg, we developed and make available a unique data set containing ground truth labeling of nuclei, chromosomes, and ecDNA, starting from 483 unlabeled images of dimensions 1040×1392 from Turner, 2017 [7]. The ground truth labeling was created by multiple scientists involved in independent annotation (Methods). Due to the difficulty of annotating 18.9K ecDNA across 483 images, a decision was made early on to use a coarsely annotated data set which 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 (4760 patches) and test (1189 patches) data sets. Importantly, in order 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 only used 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 comprised of the binary cross entropy (BCE) 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 (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 (Methods). We also tested with different network architectures such as U-net with multi-scale context aggregation using dilated convolution [17], pre-trained weights from VGG16 [18] 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 a 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 (Methods) to measure the fraction of true calls. The mIoU score showed similar convergence behavior on the training data (Figure 2b).

1.2.2 Test Set Segmentation Accuracy

On the test data (1189 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 4).

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 (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 (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.

1.2.3 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 ***2***). 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). ecSeg performance varied across cell-lines (Table ***3***). Thus COLO205, where the ecDNA are notably larger in the 9 images (Figure 5) had worse performance (75% precision, 64% recall) compared to 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 (Figure 6, Figure 7). 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 a accurate estimate of ecDNA abundance (Figure 2f).

1.2.4 ecDNA Heterogeneity

The ecDNA model of focal amplification [6, 9] 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 ***4***). Noting that the entropy value depends upon the number n of sampled cells (images), we also plotted the normalized *entropy-efficiency* value ($\frac{\mathcal{H}_n}{\log_2 n}$) for 40 cell-lines. Interestingly, most (21 of 29) cell-lines whose ecDNA copy numbers exceeded 10 per cell, had entropy-efficiency above 90% (Figure 2g, Table ***4***) suggesting an important role for ecDNA in maintaining copy number heterogeneity.

1.2.5 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 [7]. 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 [10]. To test the effect of drugs and other environmental factors in modulating ecDNA counts, we used ecSeg to quantify ecDNA counts in cells prior to Eb treatment, and followed up 2 weeks, 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 [19, 7] (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

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

is positive, and increasing for small values of k, and decreases 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, Methods).

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

Different choices of the decay parameters r, α all predicted a sharp decrease in ecDNA per cell, and a decrease in heterogeneity (Figure 8) 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 (Figure 3b,c, Table ***5***). 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 $r = 20, \alpha = 0.04$. While 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.

1.2.6 Oncogene Amplification on HSR 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 HSRs [10]. 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, Figure 3e,f). To quantify this phenomenon, we used an EGFR FISH probe and ecSeg analysis to locate EGFR (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 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 [20].

1.3 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 [10, 6]. ecSeg can be used to localize 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 is often confounded by PCR mediated artifacts. Automated cytogenetic analysis allows for an automated measurement of heterogeneity and to understand 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.

1.4 Methods

1.4.1 Data set

We started with a data set from Turner et al. 2017 [7]. 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).

1.4.2 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 [21]. 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, 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.

1.4.3 Segmentation

Inspired by the U-Net, we used a modified fully convolutional neural network presented in Figure 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 [17]. 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 Figure 4. 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 $x \in \mathcal{P}, x \in \mathcal{C}$, the network outputs a class score, $P_c(x) \in [0,1]$. We trained the network to minimize a custom loss function defined below.

1.4.4 Loss Function

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

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

Similarly, we compute Dice loss as:

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

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 $n_b \gg n_n \gg n_h \gg n_e$, we assigned weight w_c to each class $c \in \{b, n, h, e\}$ as follows:

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

Correspondingly, the weight of a pixel was given by:

$$w_x = \sum_c y_c(x)w_c$$

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.

1.4.5 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:

$$M = \frac{1}{|\mathcal{T}|}\sum_{\tau}\frac{1}{|\mathcal{C}|}\sum_{c}\frac{\boldsymbol{\theta}_{c,\tau} \cdot \boldsymbol{y}_c}{\left\|\boldsymbol{\theta}_{c,\tau}\right\|_1 + \|\boldsymbol{y}_c\|_1}$$

1.4.6 Post-processing of Segmentation

Post-training, the network outputs a 256×256 matrix O, with

$$O[x] = \arg\max_{\mathrm{c}} P_c(x)$$

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.

1.4.7 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 ecDNA 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 $\frac{\sqrt{75}}{pi} \simeq 5$ pixels ensures that ecDNA 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:

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

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

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

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

1.4.8 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).

1.4.9 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.

1.4.10 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 [19]. 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

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

$$d_k = 1 - b_k$$

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

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

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 [22], $f_{m,\alpha}(k)$ decreases logistically in value for $k > M_s$ reaching $f_{m,\alpha}(k) \rightarrow 0$ 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 $\max\{1 - f_{m,\alpha}(M_s), f_{m,\alpha}(M_\alpha)\} \rightarrow 0$. 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

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

$f_{r,\alpha}(k)$ provides negative selection pressure causing a steep decline in the average number of ecDNA per cell. We simulated the effect of the drug using $r \in \{5, 20, 50, 100\},\ \alpha \in \{0.07, 0.04, 0.03\}$. Figure 8 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).

1.4.11 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.

1.5 Acknowledgements

Chapter 1, in full, is a reprint of the material “EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA” by Utkrisht Rajkumar, Kristen Turner, Jens Luebeck, Viraj Deshpande, Manmohan Chandraker, Paul Mischel, and Vineet Bafna as it appears in IScience, 2019. The dissertation author was the primary investigator and author of this paper.

1.6 Appendix

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 methods.

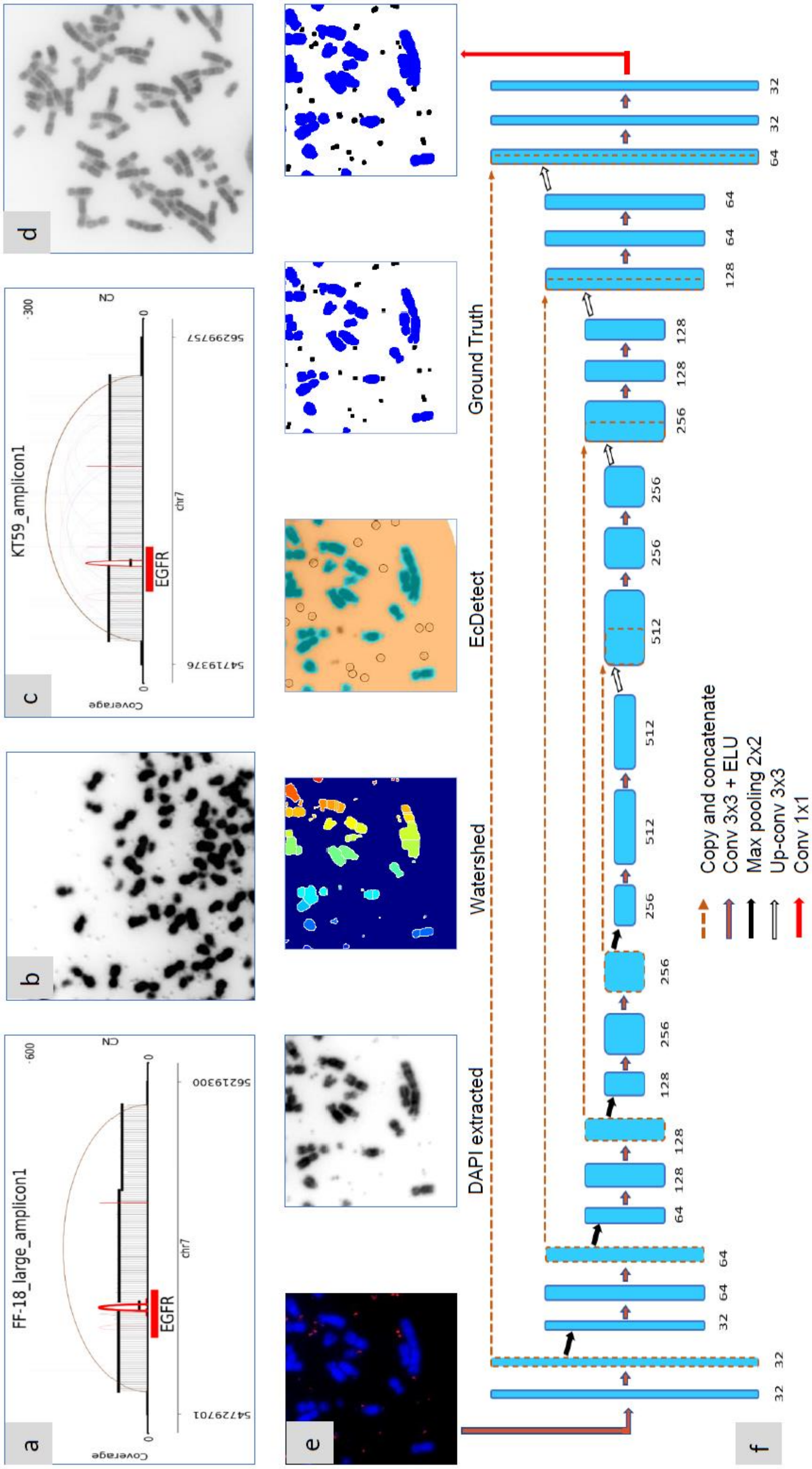
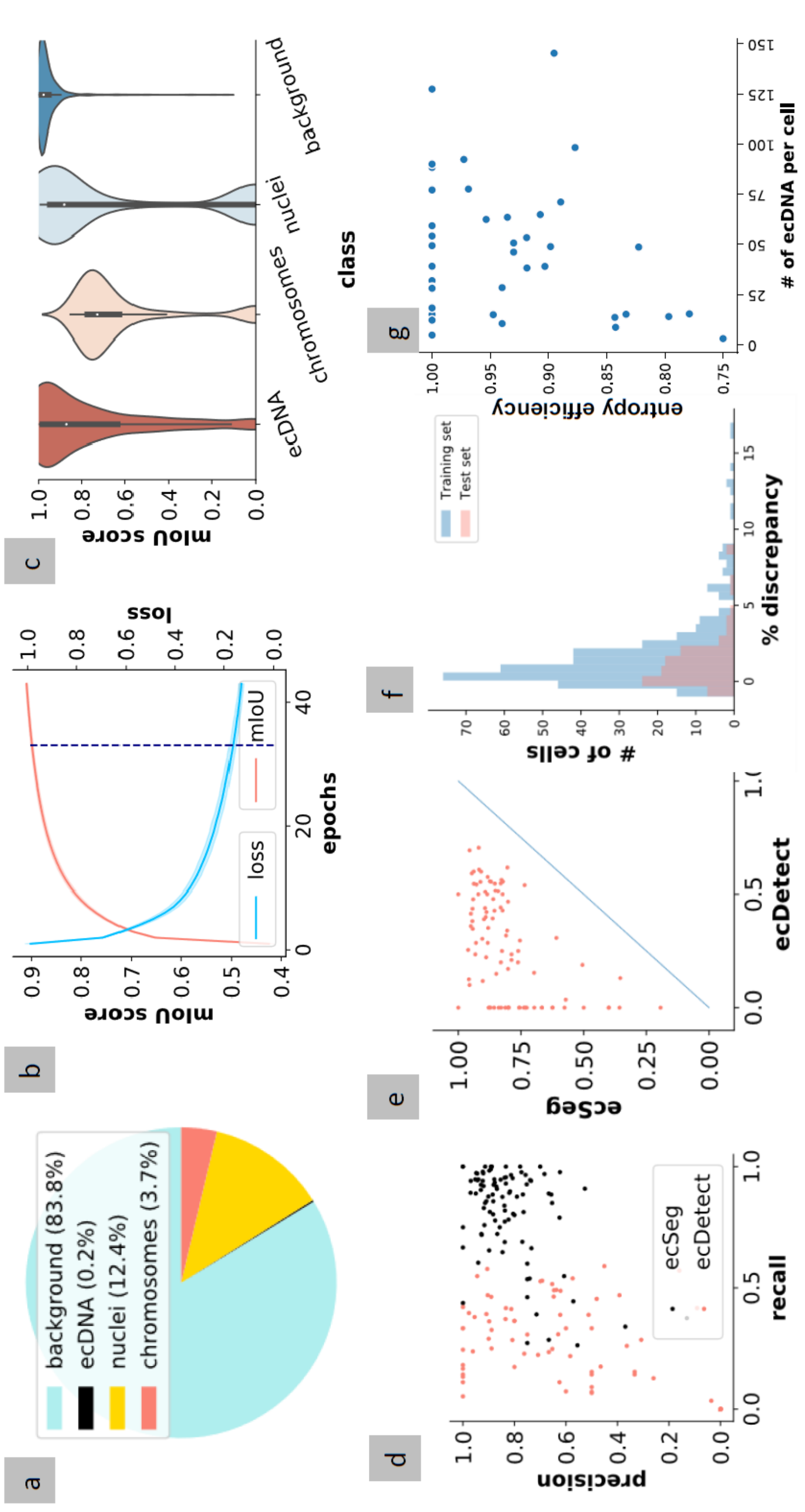

a
FF-18_large_amplicon1
Coverage
CN
600
0
EGFR
chr7
54729701
56219300
b
c
KT59_amplicon1
Coverage
CN
300
0
EGFR
chr7
54719376
56299757
d
e
DAPI extracted
Watershed
EcDetect
Ground Truth
f
32
32
32
64
64
64
128
128
128
256
256
256
512
512
512
256
256
256
128
128
128
64
64
64
32
32
Copy and concatenate
Conv 3x3 + ELU
Max pooling 2x2
Up-conv 3x3
Conv 1x1

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, while the ecDetect F1 scores rarely exceed 0.5 due to low recall, 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.


a
background (83.8%)
ecDNA (0.2%)
nuclei (12.4%)
chromosomes (3.7%)
b
mIoU score
loss
epochs
mIoU
c
mIoU score
ecDNA
chromosomes
nuclei
background
class
d
precision
recall
ecSeg
ecDetect
e
ecSeg
ecDetect
f
# of cells
% discrepancy
Training set
Test set
g
entropy efficiency
# of ecDNA per cell

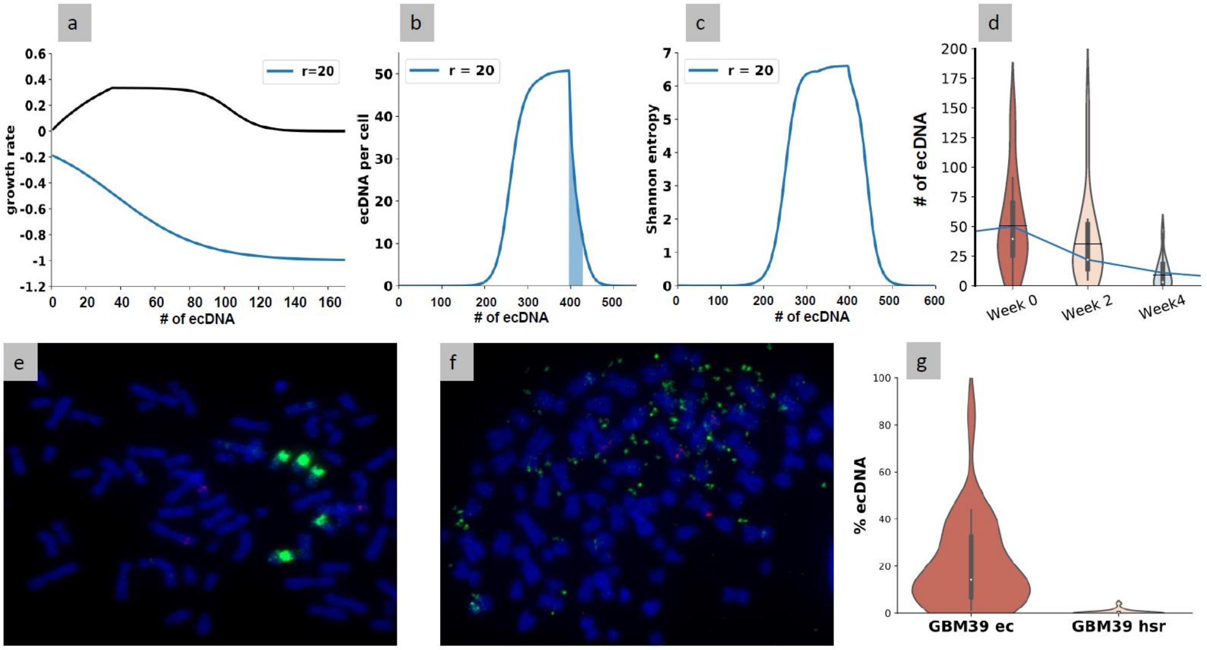


Figure 3: ecSeg applications. (a) The black line shows the growth rate $b_k - d_k$ for ecDNA driven amplification (parameters $\alpha = 0.1, m = 100$), which rises initially and slowly decreases to 0. The effect of a drug on growth rate (blue line) is modeled using a negative selection function $f_{r,\alpha}(k)$ for parameters $\alpha = 0.4, r = 20$. **(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 while white circles and box plots show the median and the middle 50th percentile. The blue line shows mean values of the simulation from panel 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)** A glioblastoma cell-line with EGFR proto-oncogene (stained using green FISH signal) found in homogeneously 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.

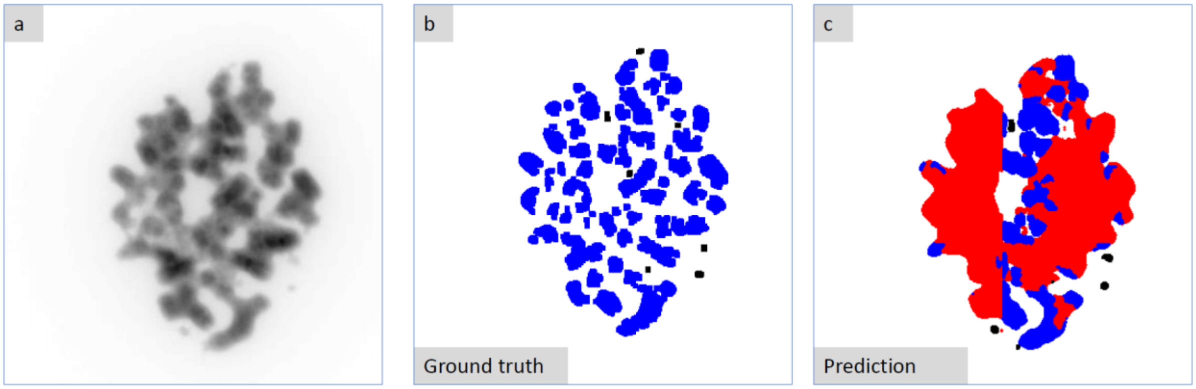


Figure 4: Incorrect classification of chromosomes as nuclei in COLO205. (a) DAPI of original image from cell line COLO205. **(b)** Ground truth annotation with intact nuclei, chromosomes, and ecDNA being represented by red, blue, and black, respectively. **(c)** Segmentation map. COLO205 tumor cell remain 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 Figure 2b.

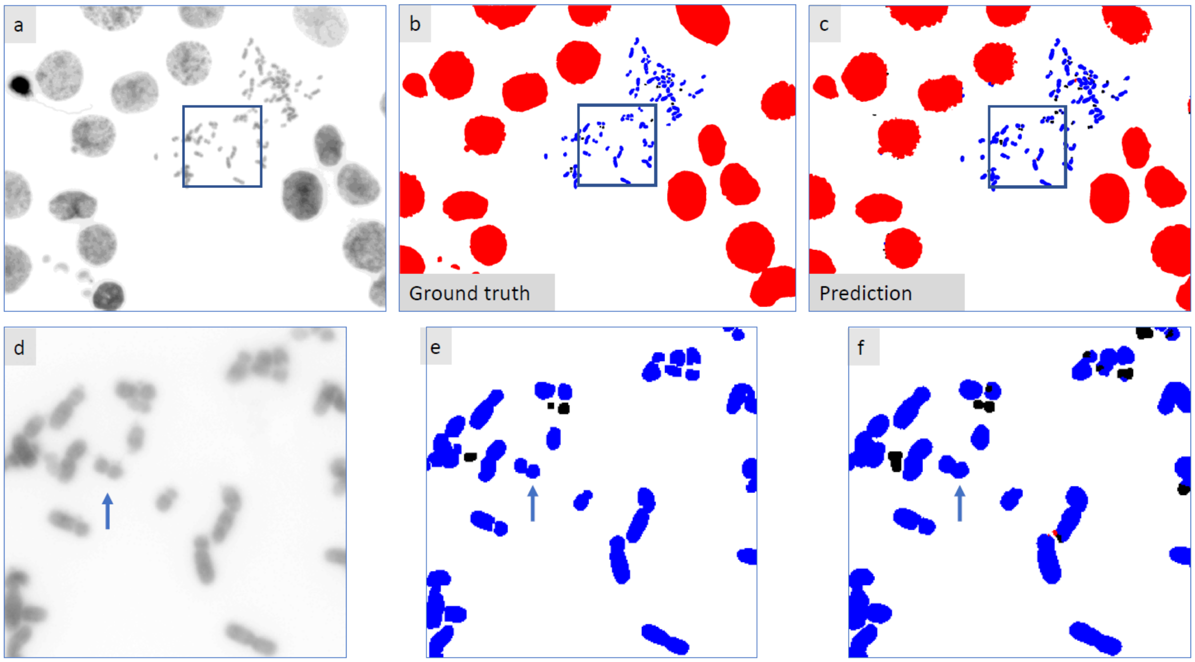


Figure 5: Incorrect detection of large ecDNA in COLO205. (a) DAPI of original image from cell-line COLO205. **(b)** Ground truth annotation with intact nuclei, chromosomes, and ecDNA being represented by red, blue, and black, respectively. **(c)** ecSeg Segmentation map. **(d,e,f)** Crops of DAPI, ground 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 Figure 2b.

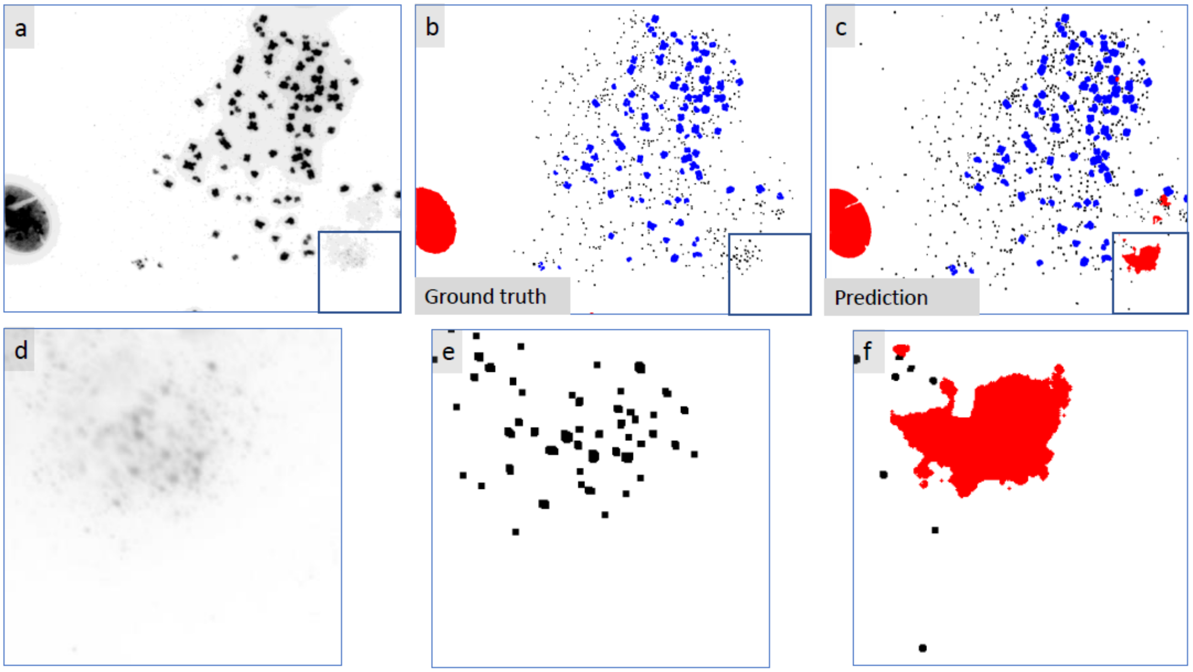


Figure 6: Incorrect false negative calls in cell line CA718. (f) is burst nucleus, but appears to show as ecDNA when zoomed in, and was marked as ecDNA during human annotation. ecSeg correctly annotates it as a nucleus identifying a mistake in the human annotation. Related to main Figure 2b.

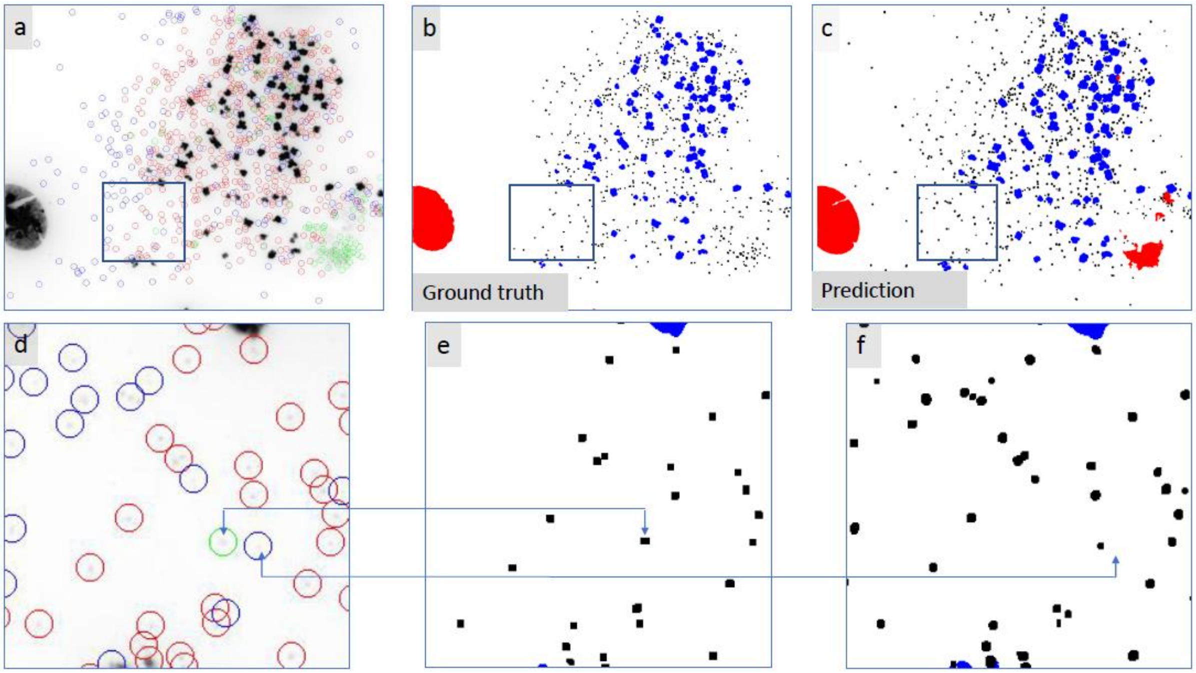


Figure 7: Incorrect annotation of ecDNA in cell line CA718. (a) DAPI of original image from cell-line CA718. **(b)** Ground truth annotation **(c)** ecSeg Segmentation map **(d,e,f)** Crops of DAPI, ground truth annotation, and ecSeg segmentation. Blue circles denote false positives, red circles are 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 Figure 6. Related to main Figure 2d.

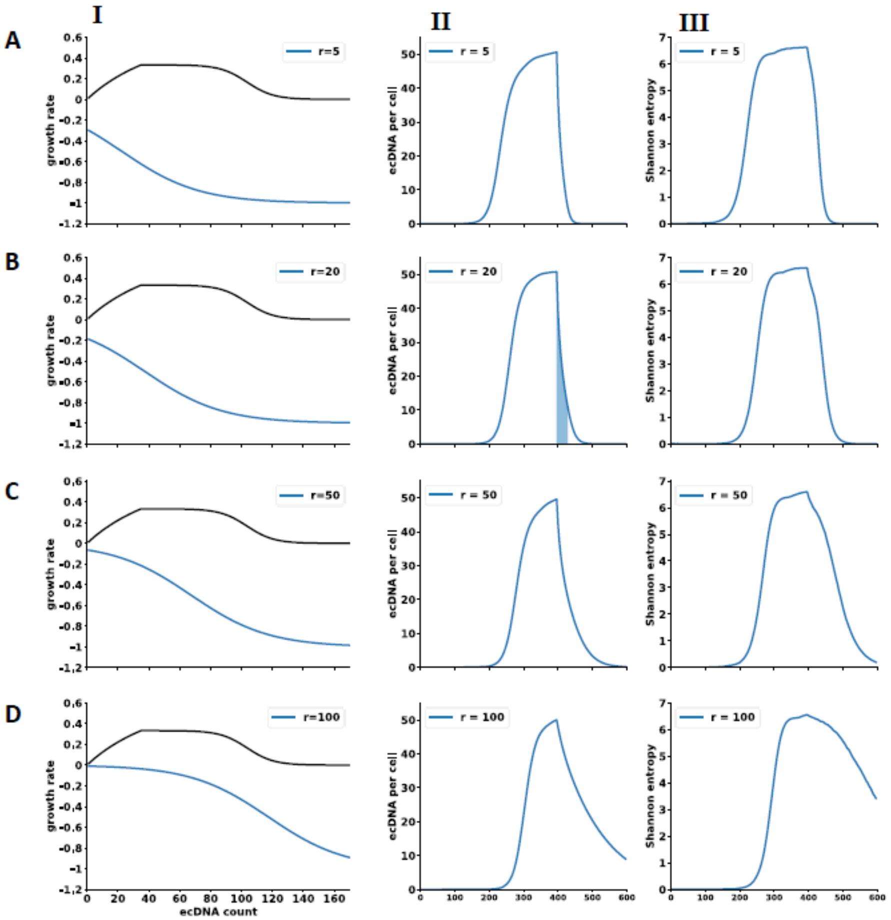


Figure 8: Simulating the impact of drug on ecDNA counts and heterogeneity. Column I shows the modeled growth rates $b_k - d_k$ as a function of ecDNA count (k) for untreated (black line) and drug-treated (blue) lines, for $\alpha = 0.04$, and $r \in \{5,20,50,100\}$ (rows A-D). Columns II and III show 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 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 Erlotinib (related to main Figure 2h).

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.

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.

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.

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.

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.

Please see the Supplemental Files for Deep Learning in Cancer Biology.

CHAPTER 2. Pan-Cancer Analysis of Oncogene Amplification in Interphase Cells

2.1 Introduction

Oncogene amplification is a key driver of cancer pathogenesis. Focal amplifications can occur as chromosomal homogeneously staining regions (HSR) or as extrachromosomal DNA (ecDNA). Amplified ecDNA is present in more than 20 different cancer types and occur especially frequently in glioblastoma, sarcoma, and esophageal carcinoma [7, 23]. Importantly, these focal amplifications are associated with worse patient outcomes [23]. Thus, there is an urgent need for methods and tools to investigate focal amplifications, like ecDNA, in tumor cells.

Sequenced-based methods such as AmpliconArchitect [9] and AmpliconReconstructor [24] aim to reconstruct these focal amplifications. However, ecDNA may integrate into and egress out of chromosomes in response to the cellular environment [10] while maintaining their features. Thus, focal amplifications can be ecDNA or HSR, but can have nearly identical structures. This makes it difficult for sequence-based methods to capture the dynamic nature of ecDNA and the amplification mechanism of a cell's present state.

Image-based reconstructions are *currently* the canonical way of determining the amplification mechanism of a cell's present state. EcDNA can be visually identified in fluorescently stained images of tumor cells in metaphase. They appear as hundreds, sometimes thousands, of tiny faint DNA particles detached from chromosomes. Rajkumar et al. developed ecSeg, a deep learning tool, to semantically segment ecDNA in metaphase cells and correlate the segmentations with amplification profiles to reveal ecDNA mechanics. However, capturing cells in metaphase requires (live-)cell imaging and is typically performed on cultured cell lines. In clinical practice, however, cells are harvested from patient tumor tissue and contain interphase cells in which the DNA is loosely wound and inside an intact nuclear membrane. This makes it extremely challenging to discern ecDNA even for a trained eye.

In this work, we discern HSR and ecDNA amplifications using the unique fluorescent staining patterns of interphase nuclei. We present ecSeg-i, a deep learning-based tool to cytogenetically determine the amplification status of interphase cells. We perform a pan-cancer study to reveal the amplification profiles across 14 tumor types, including 32 cell lines and 4 patient tumor tissue types. We first show that ecSeg-i achieves an F1-score of at least 0.88 in determining the amplification status of interphase cells. We then present critical use cases of ecSeg-i such as quantifying the amplification heterogeneity between HSR and ecDNA amplified cell lines.

2.2 Methods

2.2.1 Dataset Overview

We obtained images from 14 different tissue types using three different image acquisition protocols which we denote as cultured cell, tissue model, and patient tissue (Figure 9b). We detail the protocols in the Image Acquisition Protocols section (2.2.7). We obtained 450 tissue model images, 257 cultured cell images, and 57 patient tissue images. We generated the tissue model and cultured cell images from 32 unique cell lines (Figure 9c). We collected the 57 tissue images from patients with esophageal cancer (ESC), head and neck squamous cell carcinoma (HNT), non-small cell lung carcinoma (LUC), and non-small cell lung carcinoma squamous (LUM) (Figure 9c).

For several cell lines, we targeted more than one oncogene. For example, we probed for FGFR2 and MYC oncogene for the H716 cell line (Figure 9c). Each oncogene in each cell line has unique amplification profiles. Accordingly, we treated each unique cell line - oncogene pair as a separate sample. In total, we evaluated 39 unique cell line - oncogene pairs.

2.2.2 Ground Truth Labeling

We used whole genome sequencing (WGS) to identify the amplified oncogene in the cultured cell and tissue model cell lines. We then probed for these amplified genes using fluorescence in situ hybridization (FISH) probes and produced metaphase spreads to identify whether the oncogene was amplified on ecDNA or HSR. We grouped the cell line-oncogene pairs into either ec-amplified (ec-amp), HSR-amplified (HSR-amp), or no-amplification (no-amp). All the nuclei belonging to a particular cell line have the same label as their parent cell line. For example, we labeled all the cells in the COLO320HSR cell line as HSR-amp.

2.2.3 Training and Test Split

We split the tissue model images into 75% (343) for training and 25% (98) for testing (Figure 9e). However, we split the cultured cell images into 50% (118) for training and 50% (119) for testing (Figure 9e). We note that COLO320DM is a unique hybrid cell line that contained amplification on ecDNA and HSR. Hence, we did not use any of the COLO320DM images (61 tissue model and 20 cultured cell images) for training. We treated COLO320DM as a special hold test set. We also treated the 57 patient tissue images as a hold test set (Figure 9e).

2.2.4 Pre-process

We pre-processed the images by delineating each individual intact nucleus in the image. We used a package called NuSeT [25] to identify and segment each nucleus. NuSeT utilizes multiple neural networks to identify and separate each nucleus, even in dense, overlapping clusters. We drew a bounding box around each unique nucleus, cropped the bounding box, and resized the crop to a 256×256 patch (Figure 10). We gathered 41698 nuclei from the 461 training images and 25672 nuclei from the 217 test images (Figure 9d-e).

2.2.5 Architecture

The backbone of ecSeg-i is DenseNet-121 [26]. Densenet-121 is a 121 layered convolutional neural network (CNN). The feature maps of all previous layers are concatenated and fed as input to the current layer, making it *densely* connected. The primary benefit of this dense connection is that it enables deeper layers to reuse features learned in earlier layers without having to relearn them. Consequently, a DenseNet uses fewer parameters than an equivalent vanilla CNN.

DenseNet-121 is composed four dense blocks containing, 6, 12, 24, and 16 convolutional blocks, respectively. Each convolutional block is composed of 6 sequential operations: batch normalization (BN), a rectified linear unit (ReLU), 1×1

convolution, BN, ReLU, and a 3×3 convolution. The dimensions of all the feature maps within a dense block are kept the same (i.e. no down-sampling) but the number of filters increases by a growth factor k. This makes it practical to concatenate the feature maps instead of summing them.

We use a growth factor of $k = 32$. Each convolutional block adds 32 additional feature maps. In total, DenseNet-121 has one 7×7 convolutional layer, 58 3×3 convolutional layers, 61 1×1 convolutional layers, 4 averaging pooling layers, 1 max pooling layer, and one fully connected layer.

The original DenseNet-121 used a final classification layer containing 1000 output nodes as it was trying to classify 1000 classes. In this work, we use a final classification layer containing 3 output nodes corresponding to the three output classes: ec-amp, HSR-amp, and no-amp.

2.2.6 Training Procedure

We trained the ecSeg-i on 4 GeForce GTX 1080 Ti GPUs using the Adam optimizer with a learning rate of 0.0001. We used a patience criterion of 7. If the loss did not improve for 7 epochs the training was halted. We minimize the cross-entropy loss function to train our network. We trained the network for 200 epochs and found that the model converged after 120 epochs.

2.2.7 Image Acquisition Protocols

To generate the cultured cell images, we arrested the cells by treating with Colcemid (Karyomax) at a final concentration of 0.1μg/mL for 1-5 hours. Cells were collected, washed with PBS and re-suspended in 75μM KCl for 10-15 minutes at 37 °C. The hypotonic buffer reaction was quenched by adding an equal volume of Cornoy's Fixative (3:1 Methanol:Glacial Acetic Acid). Cells were centrifuged, washed and re-suspended in Cornoy's fixative three more times. Cells were re-suspended in 100-400μL of Cornoy's Fixative and dropped onto non overlapping sections of humidified slides. Slides were equilibrated in 2xSSC and dehydrated in an ascending alcohol series of 70, 85, and 100 percent ethanol for two minutes each. The appropriate DNA FISH (Empire Genomics) probe was added to the sample and placed on a 75 °C slide moat for 3-5 minutes to melt the DNA. Probe hybridization occurred at 37 °C in a humidified slide moat for 4 hours to overnight. Slides were washed for two minutes each in 0.4xSSC and 2xSSC/0.1% Tween20. Slides were stained with DAPI and washed in 2xSSC and ddH2O. Slides were mounted with mounting media (Prolonged Gold or Vectashield). Cover slips were sealed with clear nail polish to prevent drying of the sample. We captured the images using a 63x objective on either an Olympus BX43 wide field fluorescent microscope or a Leica Thunder Imager.

The tissue model images were collected using Y protocol. CytoCell Tissue Pretreatment Kit (LPS 100, Oxford Gene Technology IP Ltd.) was used for heat pretreatment of Formalin-Fixed, Paraffin Embedded (FFPE) tissue prior to Fluorescence in situ Hybridization (FISH). All FISH Probes were purchased from Empire Genomics Inc. FFPE slides were baked at 50 °C overnight, deparaffinized three times with xylene (1330-20-7, Millipore Sigma) for 10 minutes each, and immersed in 100%, and 70% Ethanol (64-17-5, VWR International LLC.) respectively for 2 minutes each. After washing in water for 2 minutes, the slides were incubated in pretreatment solution at 100 °C for 40 minutes. Slides were dehydrated in a graded ethanol series of 70%, 85% and 100% and air dried. Then 10 μL of probe mixture was applied to the hybridization area, cover-slipped and sealed with CytoBond coverslip sealant (2020-00-1, SciGene Corp.). Slides were incubated in ThermoBrite System (Abbott) at 80 °C for denaturation and hybridized at 37 °C for 16 hours. After gently removing the coverslip sealant, the slides were immersed in 2x SSC/0.1% Tween20 (V4261, Promega Corp.) for 3 minutes in the dark. The coverslips were slipped off the slides while still in the SSC buffer. Next, slides were washed in 0.4X SSC solution at 73 °C for 2min, transferred to water for 1 minutes, air dried in darkness, and stained with DAPI (DFS500L, Oxford Gene Technology IP Ltd.), and cover slipped. FISH results were examined with Keyence fluorescence microscope (bz-x800 model, Keyence Corp.).

2.3 Results

2.3.1 Test set Accuracy

We validated ecSeg-i on 10292 nuclei from the 118 cultured cell and 98 tissue model images. Out of the 10292 nuclei, the test set contained 1985 nuclei with no-amp, 3579 nuclei with ec-amp, and 4728 nuclei with HSR-amp. The model obtained an F1-score of 0.92, 0.88, and 0.88 on the no-amp, ec-amp, and HSR-amp nuclei, respectively (Figure 11).

2.3.2 Cell Line Analysis

We evaluated ecSeg-i on the test images for each cell line - oncogene pair. EcSeg-i detected $> 75\%$ of the cells as having ec-amp in all the ecDNA cell lines, $> 85\%$ of the cells as having no-amp in all the no-amp cell lines, and $> 50\%$ of the cells as having HSR-amp in 19 out of 21 HSR cell lines.

We obtained 1596 intact nuclei from the 81 COLO320DM images. COLO320DM is a unique hybrid cell line which not only contains ec-amplified cells but also HSR-amplified cells. Although we did not know a priori the exact ratio of ec-amp to HSR-amp to no-amp nuclei, we expected majority of the cells to contain amplification on ecDNA. We found that 75% of the cells contained amplification on ecDNA and 23% of the cells contained amplification on HSR. This is consistent with our expectations (Figure 11c).

2.3.3 Cell Line Evolution

We used ecSeg-i to determine if the amplification mechanism changes for a cell line as the cells continue to replicate and evolve. We took GBM39HSR as an exemplar. We evaluated the amplification mechanism on 76 GBM39HSR nuclei collected ‘early’ in the cell passage and 24 nuclei collected at a ‘later’ stage in the cell passage. EcSeg-i classified 80% of the ‘early’ cells as HSR-amp but only 37% of the ‘late’ cells as HSR-amp (Figure 12e). This shows that indeed amplification on ecDNA can start to reappear even in HSR cell lines as the cell line continues to evolve.

2.3.4 Cell-line Amplification Analysis

One primary use case of ecSeg-i is determining the amplification heterogeneity interphase cells. We used ecSeg-i to evaluate the heterogeneity between ec-amp, HSR-amp, and no-amp cells. We computed the heterogeneity using two metrics. First, we computed the copy number signal per cell by counting the total number of pixels stained with the color of the oncogene probe divided by the total number of DAPI pixels. We divide by the total number of DAPI pixels as a way of normalizing for the size of the cell. Second, we computed the number of oncogene blobs per cell by counting the number of distinct connected components stained with the color of the oncogene probe.

We evaluated the copy number signal for ec-amp and HSR-amp cells in the COLO320DM cell line. We labeled all cells that had an ec-amp probability of 0.9 or greater as ecDNA-amp cells. Similarly, we labeled all cells that had an HSR-amp probability of 0.9 or greater as HSR-amp cells. We found that ecDNA-amp cells had significantly greater heterogeneity (Figure 12a).

We also evaluated the copy number signal for all the ec-amp, HSR-amp, and no-amp cells in the test set and found that there exists significant difference in amplification heterogeneity between HSR-amp and ec-amp cells. HSR-amp cells had a mean copy number of 7 per cell (Figure 12b). However, cc-amp cells had a mean copy number signal of 15 per cell (Figure 12b), confirming Kim et al.’s findings that ecDNA cells have greater amplification heterogeneity than HSR cells [23].

There was also significant difference in the number of oncogene blobs between ec-amp and HSR-amp cell lines. Ec-amp cells had a mean number of oncogene blobs of 9 per cell, while HSR-amp cells had a mean number of oncogene blobs of 3 per cell (Figure 12c).

Lastly, in Figure 12d we show the mean copy number signal and the copy number signal variance of all HSR-amp and ec-amp cell lines. Although there is visual separation between the two classes, there exists several ec-amp images with low mean and variance while there also exists HSR-amp images with high mean and variance. This indicates that although heterogeneity is an important distinguishing feature of ecDNA and HSR cells, it is not a perfect discriminator. Thus, a tool with deeper capabilities, such as ecSeg-i, is required.

2.3.5 Quantifying Multiple Oncogenes in a Single Cell

Another critical use case of ecSeg-i is cytogenetically reconstructing the amplification profile of multiple oncogenes within the same cell. For example, we tracked the amplification profile of FGFR2 and MYC within the same cell for all cells in the H716 cell line. The mean and median copy number signal of FGFR2 is 0.35 and 0.24, respectively (Figure 12). However, the mean and median of MYC copy number per cell is 0.25 and 0.16, respectively (Figure 12). There also existed regions with both FGFR2 and MYC amplification which are cytogenetically seen with a yellow color (FGFR2-green and MYC-red). We show that different oncogenes have significantly different amplification profiles within the same cell.

2.3.6 Patient Tissue Results

We used ecSeg-i to determine the amplification status on 57 patient tissue images across four tumor types. In Figure 13a, we show an example of an ESC tissue image. Patient tissue images typically consist of greater variance in the number of cells containing each amplification mechanism, unlike pure ecDNA or HSR cell lines. We show that there exist cells with clear ec-amp, HSR-amp, and no-amp signals (**Figure *13***a bottom-row).

We found that 2114 cells out of 6020 cells (35%) in LUC contained amplification on ecDNA (Figure 13b). As expected, LUC also contained the greater heterogeneity in the copy number signal and number of oncogene blobs per cell (Figure 13c-d). However, ecSeg-i classified less than 25% of the cells in LUM, ESC, and HNT as ec-amp. Likewise, there was less heterogeneity in copy number signal and number of oncogene blobs for each of these tissue types.

2.4 Discussion and Conclusion

Cytogenetically identifying the amplification mechanism in interphase cells is an important and incompletely understood problem. Although sequence-based methods can reconstruct focal amplifications, they cannot fully capture the dynamic nature of ecDNA and the amplification mechanism of a cell's *present* state. Image-based tools can accurately reconstruct ecDNA in fluorescently stained images of cells in metaphase in which the ecDNA is clearly visible as tiny DNA particles floating separately from the chromosomes. However, this requires (live-)cell imaging and is difficult to perform on patient tissue images. Patient tissue images primarily contain densely clustered interphase cells, where the DNA is inside an intact nuclear membrane and loosely wound. This makes it extremely challenging to discern ecDNA even for a trained eye.

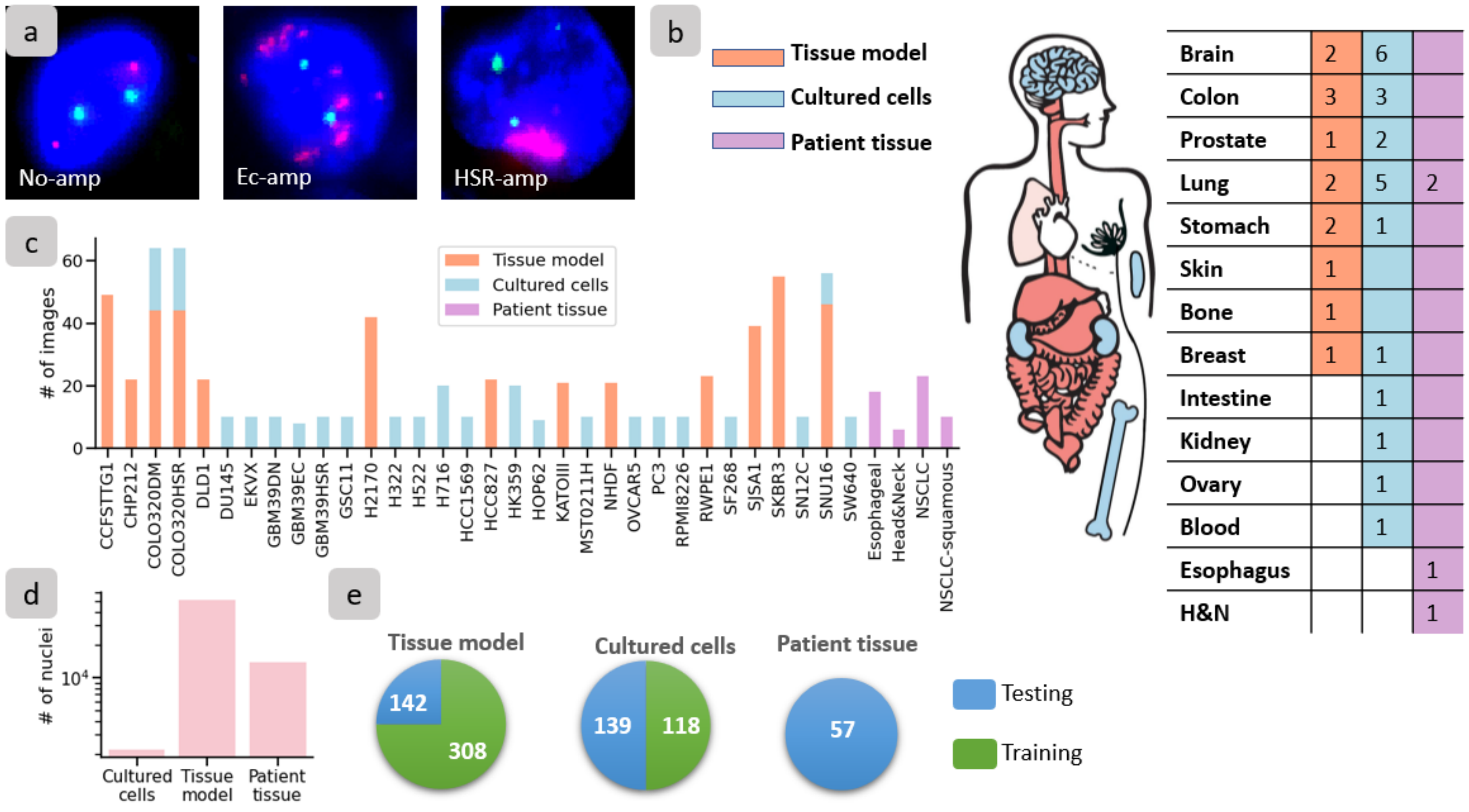
To enable investigation of ecDNA in clinical settings, we present ecSeg-i, a deep learning-based tool, to cytogenetically identify ecDNA and HSR amplifications in interphase cells. We show that ecSeg-i achieves nearly 0.9 F1 score in determining the amplification mechanism across 39 unique cell line – oncogene pairs. We then demonstrated various use cases of ecSeg-i such as capturing the evolution of GBM39HSR cells, amplification heterogeneity between ec-amp, HSR-amp, and no-amp cell lines, and reconstructing the amplification profile of multiple oncogenes within a single cell. Most importantly, we show that ecSeg-i accurately quantifies the amplification mechanism of patient tissue images from various tumor types.

2.5 Acknowledgements

Chapter 2, in full is currently being prepared for submission for publication of the material “Pan-Cancer Analysis of Oncogene Amplification in Interphase Cells” by Utkrisht Rajkumar, Ellis J Curtis, Kristen Turner, Sihan Wu, Justin Wahl, Ivy Wong, Jun Tang, Homa Hemmati, Shailaja Kasibhatla, Paul Mischel, and Vineet Bafna. The dissertation author was the primary investigator and author of this material.

2.6 Appendix

Figure 9: Data Overview. (a) Examples of interphase cells with no-amp, ec-amp, HSR-amp. **(b)** Types of tissue separated by image acquisition protocols denoting the number of unique cell lines and patient tissue. **(c)** Number of images for each cell line and patient tissue type. **(d)** Total number of nuclei across all the images from each image acquisition protocol. **(e)** Training and test split for each acquisition protocol.



Brain	2	6	
Colon	3	3	
Prostate	1	2	
Lung	2	5	2
Stomach	2	1	
Skin	1		
Bone	1		
Breast	1	1	
Intestine		1	
Kidney		1	
Ovary		1	
Blood		1	
Esophagus			1
H&N			1

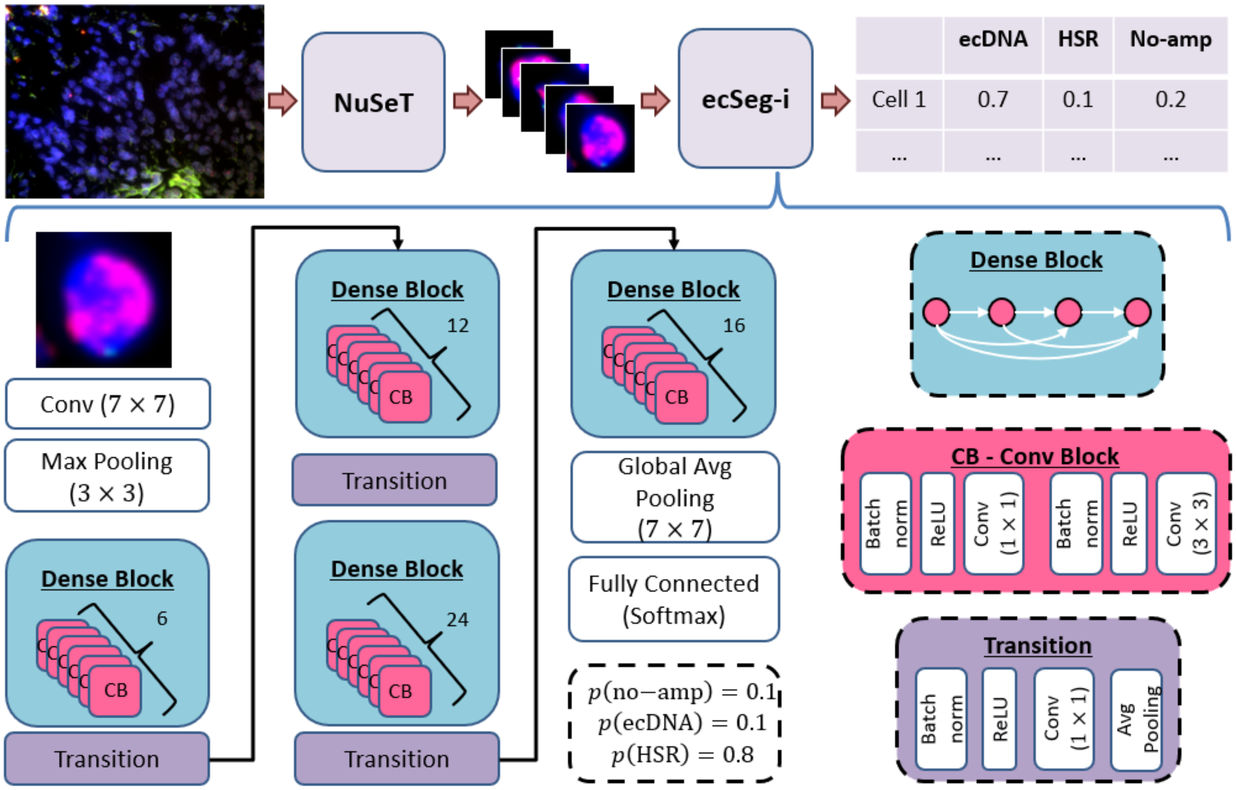


Figure 10: ecSeg-i Pipeline. The top row shows the ecSeg-i pipeline. We feed the tissue images to NuSeT which crops out each nucleus and feeds to ecSeg-i. EcSeg-i then determines the amplification mechanism probability for each nucleus. EcSeg-i uses DenseNet-121 as its backend.

Figure 11: Test set Accuracy. (a) F1-score on test set, where n is the number of cells in each class. **(b)** Distribution of cells predicted as no-amp, ec-amp, and HSR-amp in COLO320DM, a hybrid cell line. **(c-e)** Distribution of cells predicted as no-amp, ec-amp, and HSR-amp across all HSR, no-amp, and ecDNA cell lines.

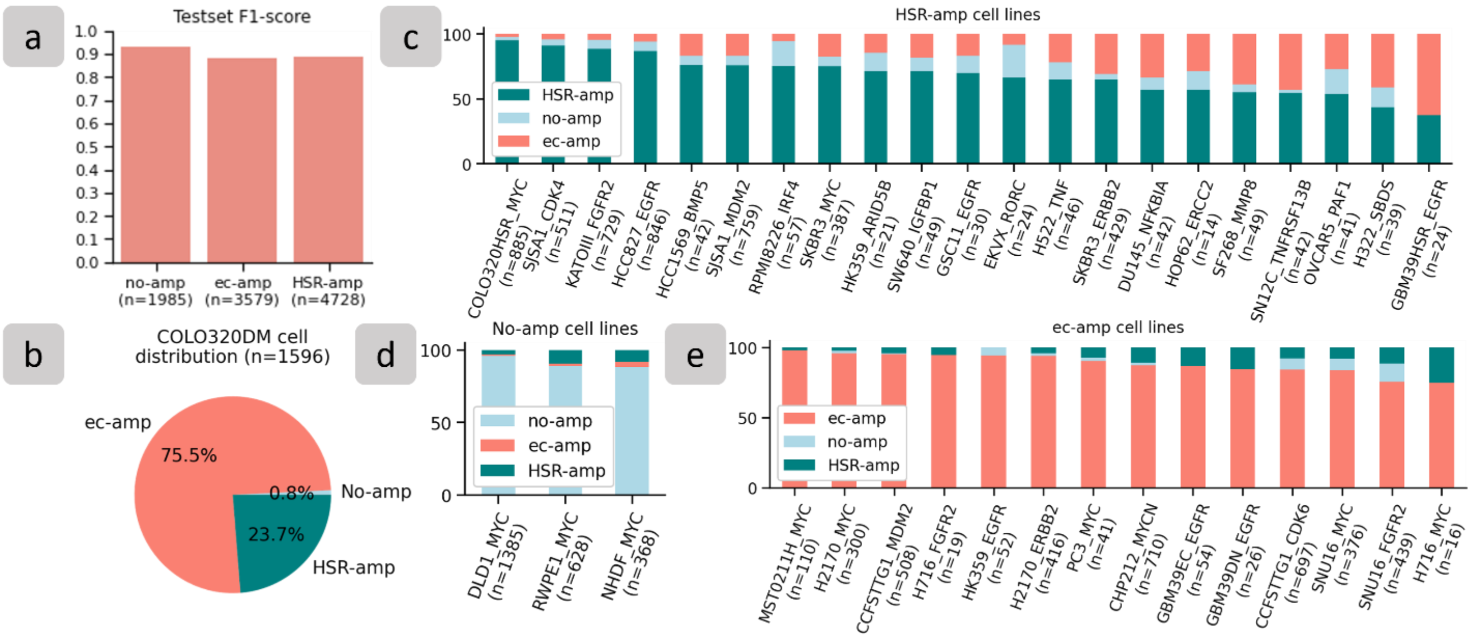
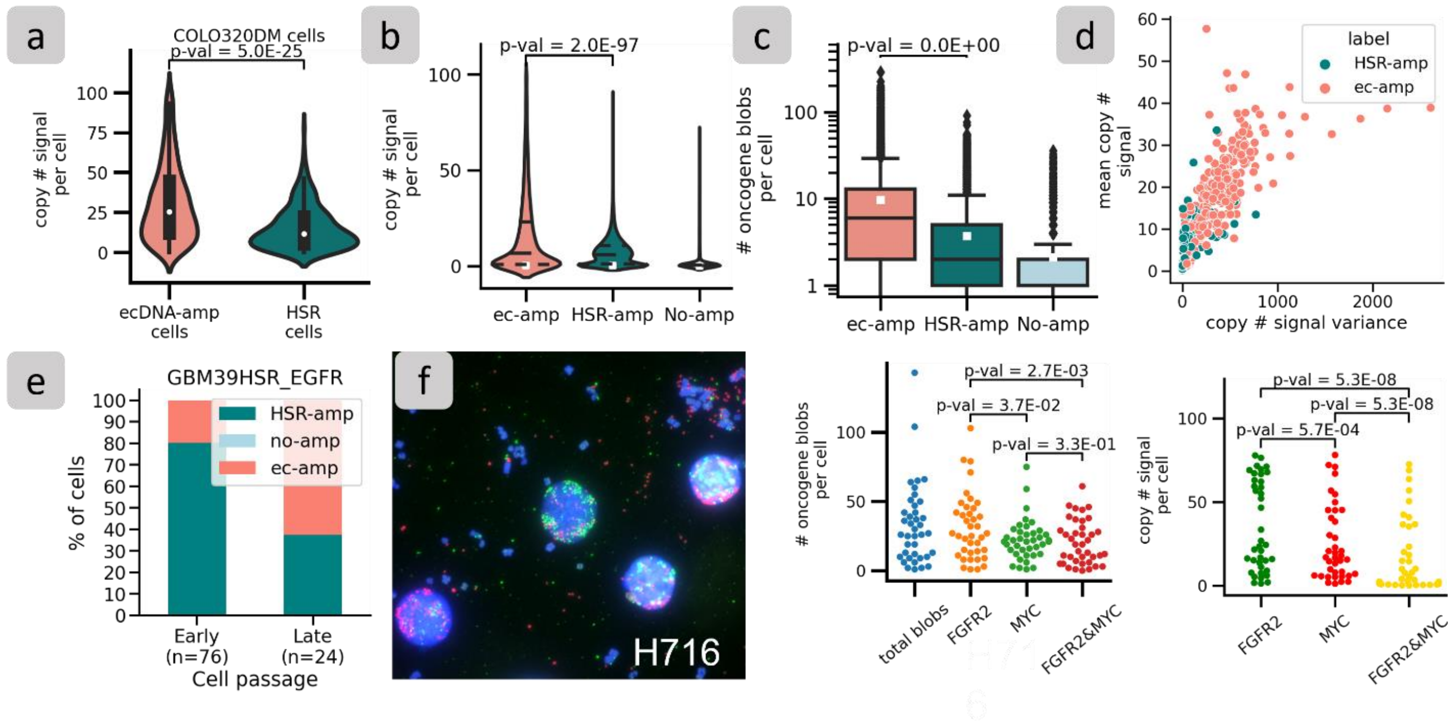

a
Testset F1-score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
no-amp (n=1985)
ec-amp (n=3579)
HSR-amp (n=4728)
b
COLO320DM cell distribution (n=1596)
ec-amp
75.5%
0.8%
No-amp
23.7%
HSR-amp
c
HSR-amp cell lines
100
50
0
HSR-amp
no-amp
ec-amp
COLO320HSR_MYC (n=885)
SJSA1_CDK4 (n=511)
KATOIII_FGFR2 (n=729)
HCC827_EGFR (n=846)
HCC1569_BMP5 (n=42)
SJSA1_MDM2 (n=759)
RPMI8226_IRF4 (n=57)
SKBR3_MYC (n=387)
HK359_ARID5B (n=21)
SW640_IGFBP1 (n=49)
GSC11_EGFR (n=30)
EKVX_RORC (n=24)
H522_TNF (n=46)
SKBR3_ERBB2 (n=429)
DU145_NFKBIA (n=42)
HOP62_ERCC2 (n=14)
SF268_MMP8 (n=49)
SN12C_TNFRSF13B (n=42)
OVCAR5_PAF1 (n=41)
H322_SBDS (n=39)
GBM39HSR_EGFR (n=24)
d
No-amp cell lines
100
50
0
no-amp
ec-amp
HSR-amp
DLD1_MYC (n=1385)
RWPE1_MYC (n=628)
NHDF_MYC (n=368)
e
ec-amp cell lines
100
50
0
ec-amp
no-amp
HSR-amp
MST0211H_MYC (n=110)
H2170_MYC (n=300)
CCFSTTG1_MDM2 (n=508)
H716_FGFR2 (n=19)
HK359_EGFR (n=52)
H2170_ERBB2 (n=416)
PC3_MYC (n=41)
CHP212_MYCN (n=710)
GBM39EC_EGFR (n=54)
GBM39DN_EGFR (n=26)
CCFSTTG1_CDK6 (n=697)
SNU16_MYC (n=376)
SNU16_FGFR2 (n=439)
H716_MYC (n=16)

Figure 12: Amplification Heterogeneity. (a) Amplification heterogeneity presented as copy number signal between cells predicted as ec-amp and HSR-amp. **(b)** Cell level copy number signal across all cell lines separated by ec-amp, HSR-amp, and no-amp. **(c)** Heterogeneity presented as number of oncogene blobs in each cell across all cell lines. **(d)** Image-level mean and variance of copy number signal for ec-amp and HSR-amp images. **(e)** Number of cells predicted as no-amp, ec-amp, and HSR-amp for GBM39HSR cells collected at different stages in the cell passage. **(f)** Copy number of signal and number of oncogene blobs for FGFR2 and MYC oncogene for each cell in H716.


a
COLO320DM cells
p-val = 5.0E-25
copy # signal per cell
100
75
50
25
0
ecDNA-amp cells
HSR cells
b
p-val = 2.0E-97
copy # signal per cell
100
50
0
ec-amp
HSR-amp
No-amp
c
p-val = 0.0E+00
# oncogene blobs per cell
100
10
1
ec-amp
HSR-amp
No-amp
d
label
HSR-amp
ec-amp
mean copy # signal
60
50
40
30
20
10
0
0
1000
2000
copy # signal variance
e
GBM39HSR_EGFR
HSR-amp
no-amp
ec-amp
% of cells
100
90
80
70
60
50
40
30
20
10
0
Early (n=76)
Late (n=24)
Cell passage
f
H716
p-val = 2.7E-03
p-val = 3.7E-02
p-val = 3.3E-01
# oncogene blobs per cell
100
50
0
total blobs
FGFR2
MYC
FGFR2&MYC
p-val = 5.3E-08
p-val = 5.3E-08
p-val = 5.7E-04
copy # signal per cell
100
50
0
FGFR2
MYC
FGFR2&MYC

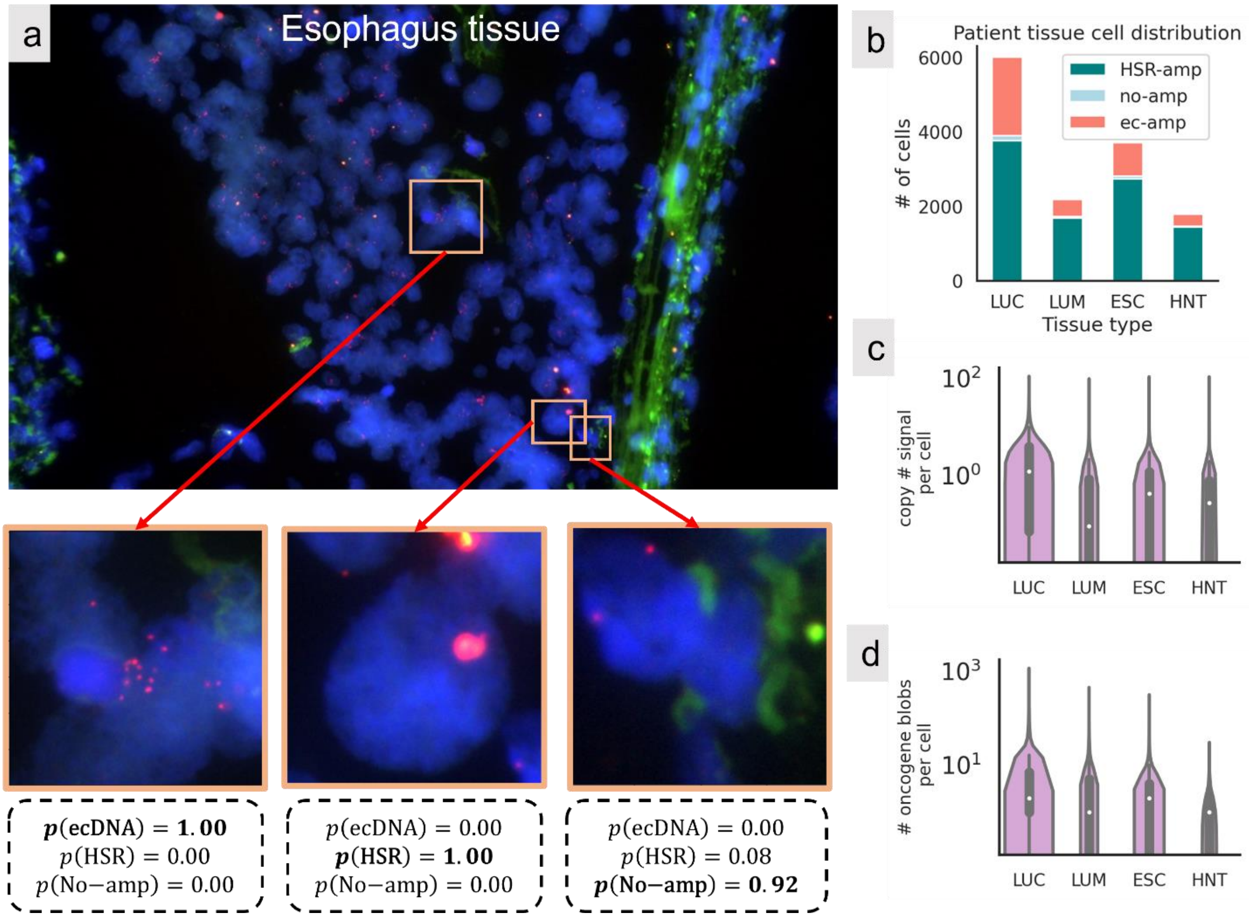


Figure 13: Patient Tissue Results. (a) Image of esophageal tumor tissue with examples of cells various amplification mechanisms. **(b)** Distribution of cells based on their predicted amplification mechanism. **(c)** Distribution of copy number signal per cell. **(d)** Distribution of number of oncogene blobs per cell.

CHAPTER 3. DeepViFi: Detecting Oncoviral Infections in Cancer Genomes using Transformers

3.1 Introduction

Viral infections in (human) hosts are pervasive and occur through a variety of mechanisms. Viral genomes may be encoded using RNA (e.g., Hepatitis C Virus, influenza viruses, Coronavirus) or DNA (e.g. Hepatitis B, Papilloma virus) [27]. Retroviruses like HIV convert their RNA genomes into DNA and then back into RNA for transcription [28]. In all cases, the virus utilizes the host machinery to express viral genes and allow the virus to replicate in the host. Viral infections are directly responsible for many human diseases, and new strains may lead to epidemics or pandemics when introduced into an immunologically naive population. Thus, rapid detection of a viral infection is important.

When the viral family is known, specific sequences can be probed directly by searching databases of known viral sequences. If the viral family shows high divergence between members, detection based on direct sequence match can fail. Here, we address the following question: if training sequences from a diverged oncoviral family are provided, can we learn a latent representation of the sequence that allows us to determine if a query sequence belongs to that viral family without a database search.

Specifically, we take the Papilloma virus (PVs) as an exemplar of a diverged oncoviral family. Human Papillomaviruses (HPV), especially HPV16 and HPV18 are important mediators of cervical and oropharyngeal cancers [29, 30, 31]. HPV mediated oropharyngeal cancers are reaching epidemic proportions, accounting for nearly 5% of all cancers [32]. In 2017, cervical cancer was the second most common cause of cancer related death for women across the world. Other Papilloma viruses (PVs), albeit less well understood, have also been implicated in human diseases including skin warts and rare diseases such as epidermodysplasia verruciformis. Furthermore, there is huge diversity of PVs with hundreds of strains identified [33].

Given its clinical importance, many tools have been developed to identify viral sequences in human cancer sequencing data [34, 35, 36, 37, 38], as well as tools focused on detecting integration into the host genome which is known to increase pathogenicity [39]. Even these specialized methods have suboptimal sensitivity for highly diverged sequences [40, 38]. To address this, ViFi [38] utilized an ensemble of hidden Markov models to identify viral sequences with high sensitivity. ViFi is sensitive but slow and the high runtime is especially burdensome for analyzing large datasets. More importantly, ViFi requires specialized training, including phylogenetic reconstruction followed by construction of an ensemble of hidden Markov models for each sub-family.

While other classification-based approaches might be utilized, we also consider that human (host) genome samples often contain significant numbers (up to 5%) of uncharacterized microbial sequences [41]. Therefore, a strict classification-based learning using a ‘closed set’ approach might not work. Her, we present DeepViFi, a Transformer-based pipeline, to identify oncoviral reads in an open-set learning framework. We show that at 90% precision, DeepViFi achieves 90% recall and is better than other neural network-based tools that use a ‘closed-set’ approach.

Additionally, we demonstrate DeepViFi's efficacy in identifying HPV reads and the viral sub-family of the infecting strain in 9 oropharyngeal tumor NGS datasets. Finally, DeepViFi can be retrained for other viral families without the need for the host, or contaminant genomes.

3.2 Related Works

Recently, deep learning tools have made tremendous progress in various biological applications such as protein folding [42], variant detection [43], and cell segmentation in images [15]. DeepVirFinder [44] and ViraMiner [45] leverage supervised learning with convolutional neural networks (CNNs) to address the kingdom-membership problem, with the goal of identifying viral sequences in metagenomic samples. They make the ‘closed-set’ assumption that the training and test sequences have the same label space.

In a different setting, DNABERT [46] uses the transformer architecture [47] to analyze human DNA contigs and produce latent representations of features on the human genome. These representations can be used for various downstream tasks such as predicting promoter regions and identifying transcription factor binding sites. However, DNABERT cannot be readily applied to short reads as it was trained on large contigs and tokenized at 3,4,5,6-mer level. Finally, it was trained only on human DNA. In this paper, we apply a similar framework to address the family-membership question for viruses using short read sequences.

3.3 Method

3.3.1 Overview

DeepViFi consists of three components: a transformer to produce latent representations of NGS short-reads, a random-forest (RF) model to classify the viral status of the latent representations, and a LightGBM model to identify the sub-family of the viral latent representations (Figure 14).

3.3.2 Method Details

Input Pre-processing. Given a read r of n base pairs, we tokenize each base-pair as a token, which empirically works better compared to tokens of larger substrings. We encode each token using the following mapping function $t : (A, C, G, T, N) \rightarrow (0, 1, 2, 3, 4)$ to obtain an encoded read vector $\boldsymbol{t}_r \in \mathbb{R}^{n\times 1}$ for each read r, where row i represented an encoding of the i-th token. We additionally define a vector, $\boldsymbol{p} = (1, 2, \dots, n-1, n) \in \mathbb{R}^{n\times 1}$, to encode the position of each base pair in r.

Transformer with Self-attention Heads. DeepViFi utilizes a transformer to learn embedding matrices $\mathbf{M}_\mathrm{t}, \mathbf{M}_\mathrm{p} \in \mathbb{R}^{1\times d}$, where the embedding dimension d is a user-defined parameter, to obtain a dense representation combining $\boldsymbol{t}_r$ and $\boldsymbol{p}$**.** The initial encoding, denoted by $\mathbf{X}^{(0)} \in \mathbb{R}^{\boldsymbol{n}\times \boldsymbol{d}}$, is obtained using

$$\mathbf{X}^{(0)} \in \boldsymbol{t}_r\mathbf{M}_\mathrm{t} + \boldsymbol{p}\mathbf{M}_\mathrm{p}.$$

The transformer architecture has $\ell = 8$ encoders and $h = 16$ attention heads per encoder, where ℓ and h are hyperparameters. Each encoding-layer i $(1 \leq i \leq \ell)$

transforms the input $\mathbf{X}^{(\mathrm{i}-1)}$ (where $\mathbf{X}^{(0)}$ is the initial input encoding) from the previous layer to $\mathbf{X}^{(\mathrm{i})}$ using self-attention heads as follows.

We denote $\mathbf{X}$(dropping the super-script) as the input to the encoders. We denote the weights of an attention head as **W$_{\mathbf{Q}}$, W$_{\mathbf{V}}$, W$_{\mathbf{K}}$**, without additional subscripts, for ease of exposition.

Let **V=XW$_{\mathbf{V}}$** denote a learned representation of the input where **W$_{\mathbf{V}}$** $\in \mathbb{R}^{d\times\frac{d}{h}}$. The transformer outputs **Z = SV.** Each resulting token $\boldsymbol{v_k}$ is mapped to $\boldsymbol{z_k} = \sum_j S_{kj}\boldsymbol{v_j},$ where $\sum_j S_{kj} = 1$. Intuitively, S_{kj} corresponds to the importance or attention of j-th token for the k-th token. To compute **S**, we use the following:

1) **Q = X W$_{\mathbf{Q}}$**, where **Q** $\in\mathbb{R}^{n\times\frac{d}{h}}$, **W$_{\mathbf{Q}}$** $\in \mathbb{R}^{d\times\frac{d}{h}}$
2) **K = X W$_{\mathbf{K}}$**, where **K** $\in\mathbb{R}^{n\times\frac{d}{h}}$, **W$_{\mathbf{K}}$** $\in \mathbb{R}^{d\times\frac{d}{h}}$
3) **D =** $\frac{\mathbf{Q}\,\mathbf{K}^{\mathbf{T}}}{\sqrt{\frac{d}{h}}}$; **D** $\in \mathbb{R}^{n\times n}$
4) **S =** Softmax(**D**)

where the Softmax operator is applied along the row dimension with $S_{ij} = \frac{e^{D_{ij}}}{\sum_\ell e^{D_{i\ell}}}$ so that $0 \leq S_{ij} \leq 1$ for all i, j, and $\sum_j S_{ij} = 1$ for all i.

The h outputs are concatenated and transformed using a dense-layer and supplied to a final feed-forward network to produce the input for the next encoder. The output of the final encoding layer ($\mathbf{X}^{(\ell)} \in \mathbb{R}^{n\times d}$) represents a transformation of the original input read r. The complete architecture is shown in (Figure 17).

Random Forest Classification of Viral Reads. We used a random forest model to determine if the read was PV positive or negative by classifying its latent representation from the transformer (Figure 14a). Specifically, we used an ensemble of 500 individual decision trees. Each individual tree outputs a class (HPV+ or HPV-) prediction of the input. The class with the most votes is the final prediction.

LightGBM for Viral Sub-family Classification. We used a LightGBM model to further segregate the detected viral-reads into sub-families after experimenting with other classification methods, including a RF model (Figure 14a). We finetuned the hyperparameters and found that a tree depth limit of 5 and the maximum number of leaves of 31 worked best. The model classified the latent representation of reads into one of Alpha, Beta, Gamma, or 'Other.'

3.4 Training and Inference

We trained the transformer using the masked language modeling paradigm. Random tokens are masked or replaced in the input and the transformer computes the likelihood for each token (A,T,C,G,N) in each position. We appended a fully connected layer with Softmax activation to the final encoder to produce

$$\mathbf{O} = \text{Softmax}(\mathbf{X}^{(\ell)}\mathbf{W_O})$$

where $\mathbf{O} \in \mathbb{R}^{n\times 5}$ represents the likelihoods of each basepair at each position. The ground truth to the model is the unmasked read. We computed the loss by comparing the predicted tokens and the ground truth.

Masking. For masking a viral read of 150bp, we randomly chose 20% (30 positions) of the tokens for a masking procedure. Of these 30 chosen positions, we replaced 80% (24) of the tokens with a [MASK] token, 10% (3) with a random token, and 10% (3) with the original token (i.e. no change). Had we replaced all 20% of the to-be-masked tokens with a [MASK] token, the encoder would have learned to only observe the [MASK] tokens and assumed that all non-masked tokens were correct. Hence, we replaced some of the to-be-masked tokens with a random or original token, forcing the encoder to keep a distributional contextual representation of every input token.

Hyper-parameter optimization. We optimized for the spare categorical Cross-Entropy loss function using the Adam optimizer with a dynamic learning rate [47]. We trained on 8 GPUs for 150 epochs with early stopping with a patience of 10 epochs. We experimented with various other masking ratios. We masked 30% of the input sequence. However, this did not significantly change our loss convergence. We also tried masking contiguous regions in the input sequence instead of selecting random positions. In this case however, the loss did not converge despite experimenting with very low learning rates. When 30 base pairs (20%) of the sequence were contiguously masked, the network did not have enough context with the remaining 120 base pairs (80%) to accurately predict the continuous missing sequence. We also experimented with tokenizing 2-mers and 3-mers instead of single base pairs. In both experiments, the loss never converged despite various network configurations and learning rates.

We experimented with different values of the hyper-parameters $\ell \in \{6,8,10\}$, $d \in \{128, 256, 384\}$, and $h \in \{8, 16\}$, and empirically settled on $\ell = 8, d = 256, h = 16$.

Inference. For inference, we removed the final fully connected layer from the transformer and used the output of the final encoder ($\mathbf{X}^{(\ell)}$) as the latent representation of the input sequence. Recall that $\mathbf{X}^{(\ell)} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{d}}$, where we chose $n = 150$ and $d = 256$. For inference, we averaged the latent representation along the column dimension to produce a single vector of dimensionality 256. We treated this vector as the final latent representation of the input read.

3.4 Dataset Generation

In a typical NGS experiment, bacteria and fungi can contaminate the target sequenced human and viral reads [48]. It is not possible currently to model all the contaminants, and instead we used an ‘open-set’ approach. Specifically, we trained DeepViFi exclusively on viral reads but tested on unseen classes such as contaminant and human reads.

3.4.1 Training Set.

We completely separated training and testing data by restricting training to reads generated from 337 PV reference genomes identified *prior to* 2018 from PaVE [49]. The training reference PV genomes ranged in length from 6953-8607 bp. We simulated reads of length 150 bp at $0.5 \times$ coverage, resulting in 1,145,800 reads.

While only the viral reads were used in the transformer to generate latent representations, we also used a negative dataset for training the random-forest classifier. In keeping with the open-set paradigm, we used 5,000 *randomly generated* reads for the negative set, but tested using real contaminant reads that were not part of training. These reads were combined with 5,000 HPV reads and used for classification using random forests.

We further classified the 337 pre-2018 references into alpha, beta, gamma, and “other” categories and randomly generated 6808, 4324, 5980, and 10856 reads, respectively. The imbalance in reads reflects the uneven number of references in each category.

3.4.2 Test Set.

We exclusively used PV genomes from PaVE deposited on or after 2018 for testing. We simulated reads from each test genomes and generated 4 test sets. Each test set contained reads from 10 viral strains with similar genomic distances from the training genomes. We labeled the test sets as easy, intermediate, hard, and non-human, based on their increasing genomic distance from the training genomes. To maintain an open set paradigm, we added contaminant and human reads to each test set. We randomly chose the contaminant and human reads from known contaminant and human genomes. The contaminant and human reads are considered non-viral.

We evaluated the LightGBM model on the 318 post-2018 references. We generated 0 Alpha, 150 Beta, 1200 Gamma, and 570 ‘other’ reads. There were an uneven number of references for each category in the post-2018 strains. Notably, there were no alpha strains in this (realistic) test set which also represents a harder scenario as alpha strains are the easiest to identify. To rectify the balance, we also evaluated subfamily classification on HPV-mediated tumor patient data, where the alpha subfamily was over-represented.

3.4.3 HPV Mediated Primary Oropharyngeal Cancer Samples.

We also evaluated on 9 oropharyngeal tumor samples from a recent study by Pang et al. [50], where the HPV status for each sample was previously determined. Each sample was HPV-16 positive and contained on average 800 million reads. After alignment filtering, each sample contained approximately 11.5 million reads on average. The ratio of viral to non-viral reads in each sample was .7% on average.

3.5 Results

3.5.1 Method Comparisons

We compared DeepViFi against ViraMiner, DeepVirFinder, ViFi, and an off-the-shelf seq2seq model (Figure 15a). DeepViFi achieved a precision-recall AUC of 0.94, 0.94, 0.91, and 0.16 for detecting HPV reads on the easy, intermediate, hard, and non-human test sets, respectively. We retrained DeepVirFinder and ViraMiner model on a custom training set before evaluation (Methods). Despite retraining, ViraMiner and DeepVirFinder both achieved an AUC value of less than 0.5 on all 4 test sets (Methods).

We also trained and tested an off-the-shelf seq2seq model using eight bidirectional long short term memory (LSTM) encoders [51]. Similar to the transformer, the seq2seq model also generates latent representations which we used to detect viral sequences. We found that although the seq2seq model outperforms DeepVirFinder and ViraMiner on the easy and intermediate test set it still performs worse than DeepViFi. It also underperforms DeepViFi on the hard and non-human test sets (Figure 15a).

On the other hand, ViFi had high precision and recall values, achieving (0.996, 1.0), (0.996, 0.983), (0.996, 0.992), and (0.991, 0.481) on the intermediate, validation, difficult, and non-human test sets. ViFi utilizes HMM ensembles to learn representations that lead to highly accurate classification. While ViFi consistently achieved the highest accuracy, it also required prior construction of a phylogeny of the PV family, followed by a selection of clades to make an ensemble of HMMs. Therefore, DeepViFi reduces some major bottlenecks of ViFi: computational resources, expertise in setup/execution, and difficulty in repurposing to other applications---while maintaining comparable accuracy.

Sub-family Classification Accuracy. We trained sub-family identification by using a LightGBM classifier on the learned representations. The training data had 6,808, 4,324, 5,980 and 10,856 reads, respectively, in the four classes. We used a 70/30 split into training and validation, and the F1-score (harmonic mean of precision and recall) to measure accuracy of sub-family classification. The accuracy on the validation data was high at 0.88, 0.82, 0.83, and 0.9 for the four sub-families.

In contrast, the test dataset (drawn from strains discovered after 2018) had 0, 150, 1200, and 570 reads in the four classes, which did not match the training distribution. Nevertheless, the LightGBM achieved an overall F1-score of 87%, with accuracies of 0.63, 0.87, and 0.93 on Beta, Gamma, and Other classes.

3.5.2 Qualitative Analysis

At large genomic distances, viral reads are far enough apart that they cannot be distinguished from random reads, based solely on percent identity. HMMs address this by assigning different weights to different genomic locations. To understand what DeepViFi is learning, we plotted the distribution of the starting positions of all HPV reads in the easy testset (Figure 15c; top-panel) and compared them to the distribution of start positions of the viral reads that were *separated* from non-viral reads---specifically, reads that had first PC value greater than 1, second PC value > 0, and third PC value > 2 (Figure 15c; bottom-panel). The sharp distinction between the two plots suggests that the discriminating reads are drawn from specific locations of the HPV genome.

We then tested if the representations learned by DeepViFi could distinguish between PV sub-families Alpha, Beta, Gamma, and `Other'. A PCA plot of the latent representations labeled by viral sub-family showed 4 visually distinct (although not linearly separable) clusters for each sub-family (Figure 15d).

3.5.3 Detecting HPV in Oropharyngeal tumor samples.

The tumor WGS (whole genome sequencing) experiments contained ~800M paired-ends per sample on the average and were available in the form of mappings to the human genome using the Burrows-Wheeler Aligner (BWA) [52]. All tumor samples were positive for HPV-16, which belongs to the Alpha subfamily of HPV. We filtered reads where both ends mapped to human sequence, and ran DeepViFi on the remaining reads using ViFi results as the ground truth. Each read from the paired-end was analyzed separately. For the 9 samples, we achieved an average precision-recall AUC of 0.90723.

DeepViFi also classified over 90% of the reads as belonging to the Alpha subfamily in each of the samples. The results are consistent with HPV-16 infection as HPV-16 belongs to the Alpha family. As low levels of other strains might be present, it was not possible to tell if the small number of misclassifications were due to classification error or the presence of other strains.

We performed PCA on the non-human reads in sample T49 to visualize if the representations of the HPV reads mapped to the same latent space as the representations of viral reads from the simulated test sets (Figure 16b). We demonstrate that representations were well separated from other reads and had a third PC value > 2, consistent with PC representations of the test sets.

DeepViFi took 12 hours to process 2 million reads on a CPU with 16 GB of memory. The time reduced to 50 minutes on a single Titan X GPU with 12 GB of memory. This was a significant speedup over the 48 hours taken by ViFi.

3.5.4 Detecting HBV in tumor samples.

As an additional exemplar, we also trained and evaluated DeepViFi to detect HBV reads. We trained the DeepViFi pipeline on 73 known HBV genomes. We then evaluated the pipeline on three HBV-negative and three HBV-positive tumor samples. DeepViFi detected less than 30 reads as viral per million on the HBV negative samples. However, it detected more than 100 reads as viral per million on the HBV positive samples.

3.6 Discussion and Conclusion

The identification of genomic sequences from a taxonomic group is an important problem that is not completely addressed. With highly diverged sequences, sequence-based database search methods may not work. Hidden Markov models improve sensitivity by focusing the scoring on specific, conserved positions. However, they are a challenge to build, as they require extensive feature engineering that have to be tuned for each taxonomic group. Therefore, HMMs are not widely utilized, and sequence-based searches continue to be widely used.

Recently, deep learning methods have provided many breakthroughs, especially in vision and natural languages. Once an architecture is specified, the training does not require domain specific expertise, making them very attractive for multiple tasks. Here, we show that the taxonomic family identification is not successful using a closed-set modeling with neural architectures, because most real-life examples provide instances of open-set learning.

In the context of viral family identification, we achieved very significant improvements by employing a transformer to learn latent representations of PV sequences. While our results easily outperformed closed-set learning using CNNs, they were still lower in sensitivity to a carefully trained ensemble of HMMs. This suggests that additional training using better sampling of the PV sequences is needed to improve representations.

Along the same vein, we can also group the methods surveyed in this paper as supervised and semi-supervised methods. The CNN based methods used here represent end-to-end supervised methods while seq2seq and DeepViFi represent semi-supervised methods. The supervised methods prioritize learning a classifier based primarily on the annotations to differentiate inputs. Meanwhile, the semi-supervised methods learn features to *characterize* the input. We show that simply learning a classifier is insufficient for problems such as identifying viral sequences in NGS data. Our results also match earlier observations on supervised and semi-supervised methods [53].

The representations of the viral sequences were so well separated from the other sequences that a simple, random-forest classifier was sufficient to identify viral reads. However, sub-family classification is a harder problem, and we had to use more sophisticated gradient boosting methods to achieve good results.

In summary, DeepViFi provides a framework for rapidly learning of representations from families, and a fast test for quickly and accurately identifying the target sequences in a larger dataset. The methods presented here are easily adaptable to a multitude of viral families and likely to help with many tasks, including identification of novel, pathogenic viruses, and removal of contaminant reads from whole genome sequencing runs. DeepViFi is available at https://github.com/UCRajkumar/DeepViFi.

3.7 Acknowledgements

Chapter 3, in full, is a reprint of the material “DeepViFi: Detecting Oncoviral Infections in Cancer Genomes using Transformers” by Utkrisht Rajkumar, Sara Javadzadeh, Mihir Bafna, Dongxia Wu, Rose Yu, Jingbo Shang, and Vineet Bafna as it appears in ACM BCB, 2022. The dissertation author was the primary investigator and author of this paper.

3.8 Appendix

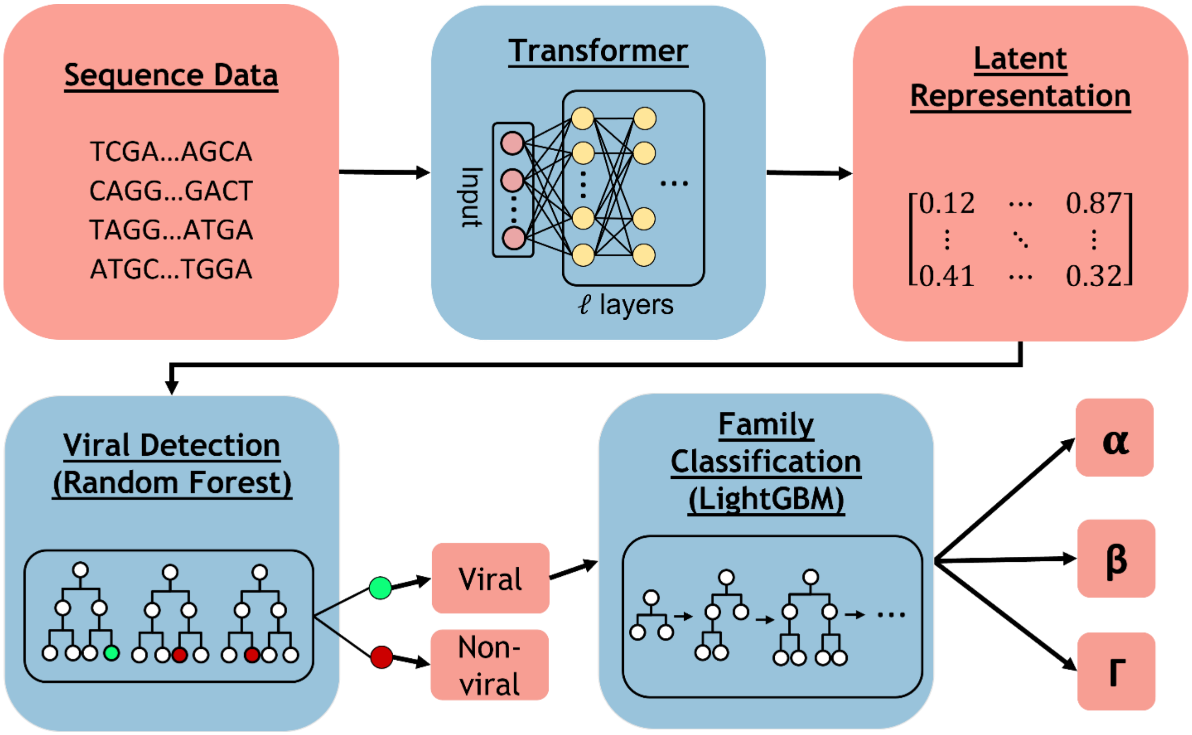


Figure 14: DeepViFi pipeline. The input to DeepViFi is DNA sequencing short reads. The transformer produces latent representations of the reads. These latent representations are fed to a random forest to determine if the read is HPV positive. The latent representations of the HPV positive reads are fed to a lightGBM model to determine their HPV subfamily.

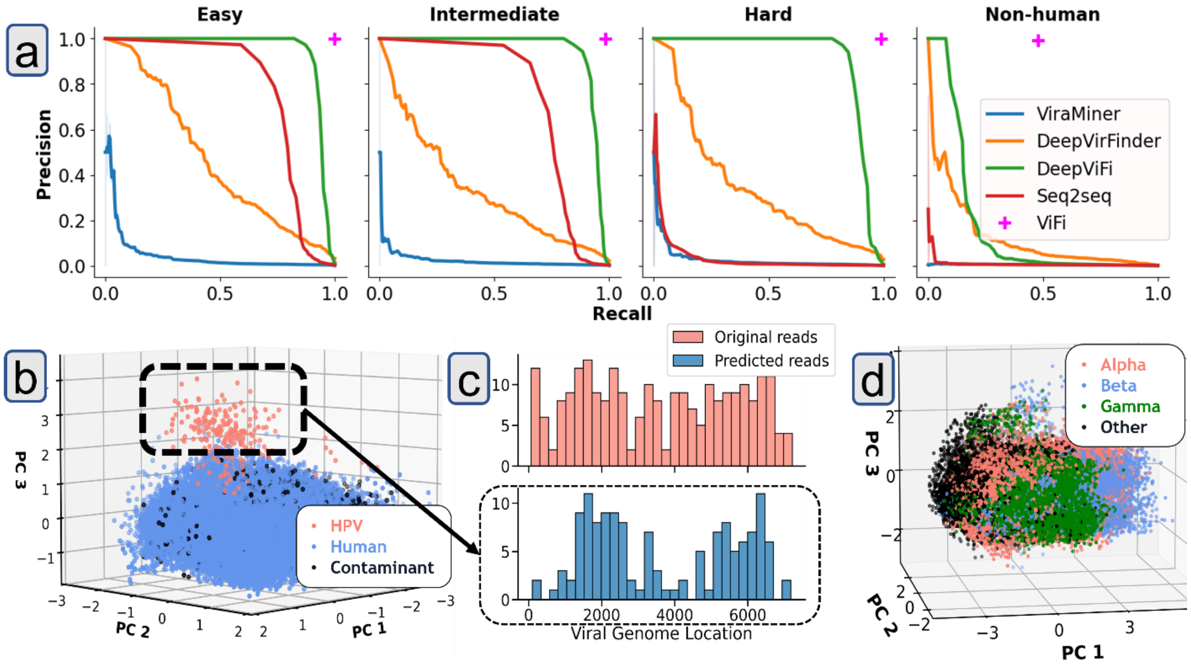


Figure 15: Test Set Performance and Analysis. (a) Precision-recall curves comparing different methods on the four test datasets. **(b)** PCA plot of latent representations of easy test set. **(c)** Read start locations of viral reads in easy test set. The top panel shows the start locations of all viral reads in the easy set. The bottom panel shows the start locations of the subset of viral reads that were highly separable in the PCA plot; i.e. the transformer was most confident of these reads as HPV fragments.

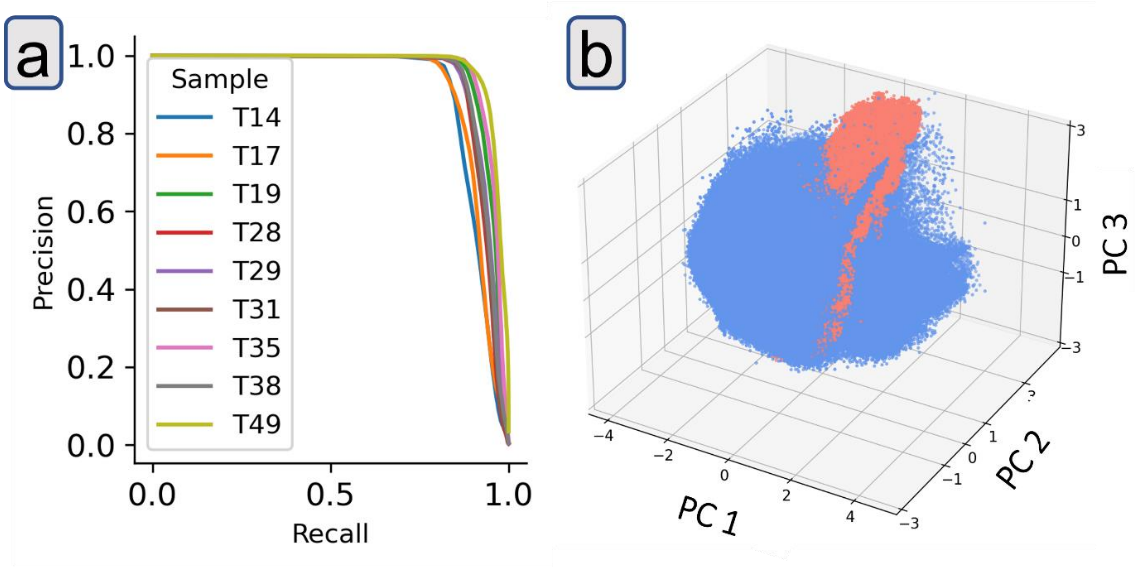


Figure 16: Tumour Sample Analysis. (a) Precision-recall curves for detecting HPV reads in tumor samples. **(b)** PCA of latent representations of T49 reads.

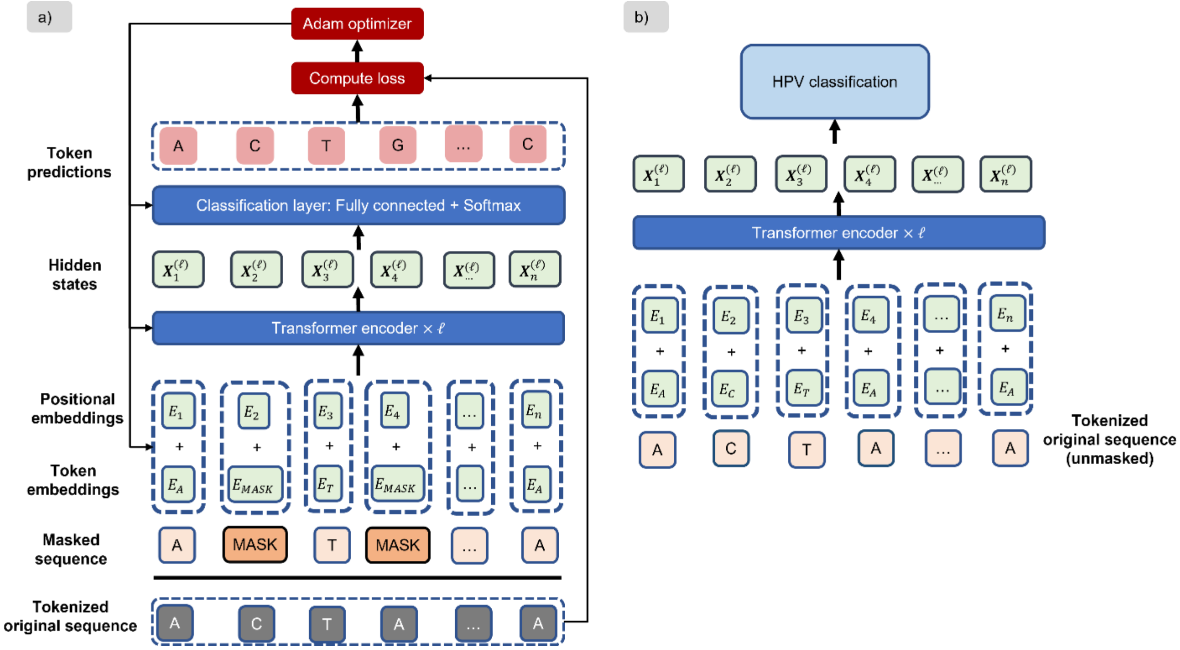


Figure 17: Detailed architecture of DeepViFi transformer. Left panel presents the training pipeline for DeepViFi's transformer. Right panel presents the inference pipeline of the transformer.

REFERENCES

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

[2] F. Menghi, F. P. Barthel, V. Yadav, M. Tang, B. Ji, Z. Tang, G. W. Carter, Y. Ruan, R. Scully, R. G. W. Verhaak, J. Jonkers and E. T. Liu, "The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations," *Cancer Cell,* vol. 34, pp. 197-210.e5, 2018.

[3] K. Kitada and T. Yamasaki, "The complicated copy number alterations in chromosome 7 of a lung cancer cell line is explained by a model based on repeated breakage-fusion-bridge cycles," *Cancer Genetics and Cytogenetics,* vol. 185, p. 11–19, August 2008.

[4] P. Ly and D. W. Cleveland, "Rebuilding Chromosomes After Catastrophe: Emerging Mechanisms of Chromothripsis," *Trends in Cell Biology,* vol. 27, p. 917–930, 2017.

[5] D. W. Garsed, O. J. Marshall, V. D. A. Corbin, A. Hsu, L. D. Stefano, J. Schröder, J. Li, Z.-P. Feng, B. W. Kim, M. Kowarsky, B. Lansdell, R. Brookwell, O. Myklebost, L. Meza-Zepeda, A. Holloway, F. Pedeutour, K. H. Choo, M. Damore, A. Deans, A. Papenfuss and D. Thomas, "The Architecture and Evolution of Cancer Neochromosomes," *Cancer Cell,* vol. 26, pp. 653-667, 2014.

[6] R. G. W. Verhaak, V. Bafna and P. S. Mischel, "Extrachromosomal oncogene amplification in tumour pathogenesis and evolution," *Nature Reviews Cancer,* 2019.

[7] K. M. Turner, V. Deshpande, D. Beyter, T. Koga, J. Rusert, C. Lee, B. Li, K. Arden, B. Ren, D. A. Nathanson, H. I. Kornblum, M. D. Taylor, S. Kaushal, W. K. Cavenee, R. Wechsler-Reya, F. B. Furnari, S. R. Vandenberg, P. N. Rao, G. M. Wahl, V. Bafna and P. S. Mischel, "Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity," *Nature,* vol. 543, p. 122–125, March 2017.

[8] D. Cox, C. Yuncken and A. Spriggs, "Minute chromatin bodies in malignant tumors of childhood," *The Lancet,* vol. 286, p. 55–58, July 1965.

[9] V. Deshpande, J. Luebeck, N. D. Nguyen, M. Bakhtiari, K. M. Turner, R. Schwab, H. Carter, P. S. Mischel and V. Bafna, "Exploring the landscape of

focal amplifications in cancer using AmpliconArchitect," *Nature Communications,* vol. 10, p. 392, January 2019.

[10] D. A. Nathanson, B. Gini, J. Mottahedeh, K. Visnyei, T. Koga, G. Gomez, A. Eskin, K. Hwang, J. Wang, K. Masui, A. Paucar, H. Yang, M. Ohashi, S. Zhu, J. Wykosky, R. Reed, S. F. Nelson, T. F. Cloughesy, C. D. James, P. N. Rao, H. I. Kornblum, J. R. Heath, W. K. Cavenee, F. B. Furnari and P. S. Mischel, "Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA," *Science,* vol. 343, p. 72–76, 2014.

[11] S. Beucher and C. Lantuejoul, "Use of watersheds in contour detection," *International Workshop on image processing Real-Time Edge and Motion Detection Estimation,* p.˜ 17–21, 1979.

[12] B. A. Hamilton, *Kaggle 2018 Data Science Bowl,* 2018.

[13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25,* p. 1097–1105, 2012.

[14] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] O. Ronneberger, P.Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[16] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis,* vol. 42, p. 60–88, 2017.

[17] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *International Conference on Learning Representations,* November 2015.

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, 2014.

[19] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein and M. A. Nowak, "Accumulation of driver and passenger mutations during tumor progression," *Proceedings of the National Academy of Sciences of U.S.A.,* vol. 107, pp. 18545-18550, October 2010.

[20] D. A. Haber and R. T. Schimke, "Unstable amplification of an altered dihydrofolate reductase gene associated with double-minute chromosomes.," *Cell,* vol. 26, p. 355–62, November 1981.

[21] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 9, p. 62–66, 1979.

[22] N. N. Pavlova and C. B. Thompson, "The Emerging Hallmarks of Cancer Metabolism," *Cell Metab.,* vol. 23, p. 27–47, January 2016.

[23] H. Kim, N.-P. Nguyen, K. Turner, S. Wu, A. D. Gujar, J. Luebeck, J. Liu, V. Deshpande, U. Rajkumar, S. Namburi and others, "Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers," *Nature genetics,* vol. 52, p. 891–897, 2020.

[24] J. Luebeck, C. Coruh, S. R. Dehkordi, J. T. Lange, K. M. Turner, V. Deshpande, D. A. Pai, C. Zhang, U. Rajkumar, J. A. Law and others, "AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications," *Nature communications,* vol. 11, p. 1–14, 2020.

[25] L. Yang, R. P. Ghosh, J. M. Franklin, S. Chen, C. You, R. R. Narayan, M. L. Melcher and J. T. Liphardt, "NuSeT: A deep learning tool for reliably separating and analyzing crowded cells," *PLoS Comput Biol,* vol. 16, p. e1008193, September 2020.

[26] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*, 2017.

[27] J. T. Schiller and D. R. Lowy, "Virus infection and human cancer: an overview," *Viruses and human cancer,* p. 1–10, 2014.

[28] W.-S. Hu and S. H. Hughes, "HIV-1 reverse transcription," *Cold Spring Harbor perspectives in medicine,* vol. 2, p. a006882, 2012.

[29] "Integration of human papillomavirus genomes in head and neck cancer: Is it time to consider a paradigm shift?," *Viruses,* vol. 9, 2017.

[30] I. J. Groves and N. Coleman, "J PatholHuman papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us?," *J Pathol,* vol. 245, p. 9–18, May 2018.

[31] "Human papillomavirus and rising oropharyngeal cancer incidence in the United States," *Journal of Clinical Oncology,* 2011.

[32] T. A. Berman and J. T. Schiller, "Human papillomavirus in cervical cancer and oropharyngeal cancer: one cause, two diseases," *Cancer,* vol. 123, p. 2219–2229, 2017.

[33] A. A. McBride, "Human papillomaviruses: diversity, infection and host interactions," *Nat Rev Microbiol,* September 2021.

[34] Q. Wang, P. Jia and Z. Zhao, "VERSE: a novel approach to detect virus integration in host genomes through reference genome customization," *Genome Med,* vol. 7, p. 2, 2015.

[35] Q. Wang, P. Jia and Z. Zhao, "VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data," *PLoS One,* vol. 8, p. e64465, 2013.

[36] J. W. Li, R. Wan, C. S. Yu, N. N. Co, N. Wong and T. F. Chan, "ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution," *Bioinformatics,* vol. 29, p. 649–651, March 2013.

[37] D. W. Ho, K. M. Sze and I. O. Ng, "Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability," *Oncotarget,* vol. 6, p. 20959–20963, 2015.

[38] N. D. Nguyen, V. Deshpande, J. Luebeck, P. S. Mischel and V. Bafna, "ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer," *Nucleic Acids Res.,* vol. 46, p. 3309–3325, April 2018.

[39] D. L. Cameron, N. Jacobs, P. Roepman, P. Priestley, E. Cuppen and A. T. Papenfuss, "VIRUSBreakend: Viral Integration Recognition Using Single Breakends," *Bioinformatics,* May 2021.

[40] Y. Hirose, M. Yamaguchi-Naka, M. Onuki, Y. Tenjimbayashi, N. Tasaka, T. Satoh, K. Tanaka, T. Iwata, A. Sekizawa, K. Matsumoto and others, "High Levels of Within-Host Variations of Human Papillomavirus 16 E1/E2 Genes in Invasive Cervical Cancer," *Frontiers in microbiology,* vol. 11, 2020.

[41] A. Gihawi, G. Rallapalli, R. Hurst, C. S. Cooper, R. M. Leggett and D. S. Brewer, "SEPATH: benchmarking the search for pathogens in human tissue

whole genome sequence data leads to template pipelines," *Genome Biol,* vol. 20, p. 208, October 2019.

[42] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature,* vol. 596, p. 583–589, August 2021.

[43] R. Poplin, P. C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean and M. A. DePristo, "A universal SNP and small-indel variant caller using deep neural networks," *Nat Biotechnol,* vol. 36, p. 983–987, November 2018.

[44] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, R. Poplin and F. Sun, "Identifying viruses from metagenomic data using deep learning," *Quant Biol,* vol. 8, p. 64–77, March 2020.

[45] A. Tampuu, Z. Bzhalava, J. Dillner and R. Vicente, "ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples," *PLoS One,* vol. 14, p. e0222271, 2019.

[46] Y. Ji, Z. Zhou, H. Liu and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics,* February 2021.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[48] S.-J. Park, S. Onizuka, M. Seki, Y. Suzuki, T. Iwata and K. Nakai, "A systematic sequencing-based approach for microbial contaminant detection and functional inference," *BMC biology,* vol. 17, p. 1–15, 2019.

[49] K. Van Doorslaer, Q. Tan, S. Xirasagar, S. Bandaru, V. Gopalan, Y. Mohamoud, Y. Huyen and A. A. McBride, "The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis," *Nucleic acids research,* vol. 41, p. D571–D578, 2012.

[50] J. Pang, N. Nguyen, J. Luebeck, L. Ball, A. Finegersh, S. Ren, T. Nakagawa, M. Flagg, S. Sadat, P. S. Mischel, G. Xu, K. Fisch, T. Guo, G. Cahill, B. Panuganti, V. Bafna and J. Califano, "Extrachromosomal DNA in HPV-Mediated Oropharyngeal Cancer Drives Diverse Oncogene Transcription," *Clin Cancer Res,* vol. 27, p. 6772–6786, December 2021.

[51] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, 2014.

[52] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics,* vol. 25, p. 1754–1760, July 2009.

[53] A. Chatterjee, O. S. Ahmed, R. Walters, Z. Shafi, D. Gysi, R. Yu, T. Eliassi-Rad, A.-L. Barabási and G. Menichetti, *AI-Bind: Improving Binding Predictions for Novel Protein Targets and Ligands,* 2021.

[54] S. Wu, K. M. Turner, N. Nguyen, R. Raviram, M. Erb, J. Santini, J. Luebeck, U. Rajkumar, Y. Diao, B. Li and others, "Circular ecDNA promotes accessible chromatin and high oncogene expression," *Nature,* vol. 575, p. 699–703, 2019.

[55] U. Rajkumar, K. Turner, J. Luebeck, V. Deshpande, M. Chandraker, P. Mischel and V. Bafna, "EcSeg: semantic segmentation of metaphase images containing extrachromosomal DNA," *Iscience,* vol. 21, p. 428–435, 2019.

[56] P. Mischel, V. Bafna, J. Ko, W. Zhang and U. Rajkumar, *Methods of treating extrachromosomal dna expressing cancers,* 2019.

[57] J. T. Lange, C. Y. Chen, Y. Pichugin, L. Xie, J. Tang, K. L. Hung, K. E. Yost, Q. Shi, M. L. Erb, U. Rajkumar and others, "Principles of ecDNA random inheritance drive rapid genome change and therapy resistance in human cancers," *bioRxiv,* 2021.

[58] S. Javadzadeh, U. Rajkumar, N. Nguyen, S. Sarmashghi, J. Luebeck, J. Shang and V. Bafna, "FastViFi: Fast and accurate detection of (Hybrid) Viral DNA and RNA," *NAR genomics and bioinformatics,* vol. 4, p. lqac032, 2022.

[59] K. L. Hung, K. E. Yost, L. Xie, Q. Shi, K. Helmsauer, J. Luebeck, R. Schöpflin, J. T. Lange, R. Chamorro González, N. E. Weiser and others, "ecDNA hubs drive cooperative intermolecular oncogene expression," *Nature,* vol. 600, p. 731–736, 2021.

[60] K. Fujimoto, T. Mizugaki, U. Rajkumar, H. Shigeta, S. Seno, Y. Uchida, M. Ishii, V. Bafna and H. Matsuda, "A CNN-based cell tracking method for multi-slice intravital imaging data," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021.

[61] S. Chowdhry, C. Zanca, U. Rajkumar, T. Koga, Y. Diao, R. Raviram, F. Liu, K. Turner, H. Yang, E. Brunk and others, "NAD metabolic dependency in cancer is shaped by gene amplification and enhancer remodelling," *Nature,* vol. 569, p. 570–575, 2019.

[62] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of graphics, gpu, and game tools,* vol. 12, p. 13–21, 2007.

[63] N. Alon, S. Butler, R. Graham and U. C. Rajkumar, "Permutations resilient to deletions," *Annals of Combinatorics,* vol. 22, p. 673–680, 2018.

[64] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol,* vol. 215, p. 403–410, October 1990.

[65] L. J. de Oliveria Andrade, A. D'Oliveira, R. C. Melo, E. C. De Souza, C. A. Costa Silva and R. Paraná, "Association between hepatitis C and hepatocellular carcinoma," *J Glob Infect Dis,* vol. 1, p. 33–37, January 2009.

[66] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, 2017.

[67] S. Rampersad and P. Tennant, "Replication and Expression Strategies of Viruses," *Viruses,* pp. 55-82, 2018.

[68] S. Sarmashghi, K. Bohmann, M. T. P. Gilbert, V. Bafna and S. Mirarab, "Skmer: assembly-free and alignment-free sample identification using genome skims," *Genome biology,* vol. 20, p. 1–20, 2019.