Title

Reconstructing and Profiling Extrachromosomal DNA

Author

Luebeck, Jens-Christian

Publication Date

2021

UNIVERSITY OF CALIFORNIA SAN DIEGO

Reconstructing and Profiling Extrachromosomal DNA

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

in

Bioinformatics and Systems Biology

by

Jens-Christian Luebeck

Committee in charge:

Professor Vineet Bafna, Chair
Professor Julie Law, Co-Chair
Professor Prashant Mali
Professor Jill Mesirov
Professor Paul Mischel

2021

The dissertation of Jens-Christian Luebeck is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents, who gave me unwavering support and encouragement as I pursued this dream.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I give my most sincere thanks to my thesis adviser, Professor Vineet Bafna for his mentorship and support throughout the course of my degree. His patient wisdom and his genuine, kind mentoring style are qualities that I and everyone who has had the privilege of having Vineet as a mentor, can attest to.

I'd also like to thank my thesis committee members, particularly Professor Paul Mischel who led the project which comprises the last chapter of the thesis. I also thank the committee co-chair Professor Julie Law, and other committee members Professor Prashant Mali and Professor Jill Mesirov. All of them have given me fantastic feedback and advice for which I am eternally grateful.

I thank the faculty I rotated with prior to joining Vineet's lab, Professor Prashant Mali, and Professor Trey Ideker for their tutelage as a first-year student.

I would also like to acknowledge my fellow lab-mates from the combined Pevzner/Bafna/Bandiera CSE labs. Particularly Viraj Deshpande for allowing me to ask him an innumerable number of questions early on in my degree. I'd also like to thank the fellow members who I shared an office with for the better part of my PhD; Ben Pullman, Duong Nguyen, Niema Moshiri and Miin Lin.

I very gratefully acknowledge one collaborator from within the lab in particular – Siavash Raeisi Dehkordi, whose hard work helped me wrap up many difficult projects. I'd

also like to thank John Pang and Nam Nguyen for their amazing efforts on the HPV ecDNA project too.

I'd like to acknowledge the fellow members of my graduate program cohort who I shared many fond memories with. I also thank the more senior members (now alumni) of my graduate program who helped mentor me, provided me with advice and shared their experiences with me when I began graduate school.

Chapter 1, in full, is a reprint of the material "AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications" by Jens Luebeck, Ceyda Coruh, Siavash R. Dehkordi, Joshua T. Lange, Kristen M. Turner, Viraj Deshpande, Dave A. Pai, Chao Zhang, Utkrisht Rajkumar, Julie A. Law, Paul S. Mischel & Vineet Bafna as it appears in Nature Communications 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material "FaNDOM: Fast nested distance-based seeding of optical maps" by Siavash Raeisi Dehkordi, Jens Luebeck and Vineet Bafna as it appears in Cell Pattens 2021. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3, in full, is a reprint of the material "Extrachromosomal DNA in HPV mediated oropharyngeal cancer drives diverse oncogene transcription" by John Pang, Nam Nguyen, Jens Luebeck, Laurel Ball, Andrey Finegersh, Shuling Ren, Takuya Nakagawa, Mitchell Flagg, Sayed Sadat, Paul S. Mischel, Guorong Xu, Kathleen Fisch, Theresa Guo, Gabrielle Cahill, Bharat Panuganti, Vineet Bafna and Joseph Califano, as it appears in Clinical Cancer Research 2021. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4, in full is currently being prepared for submission for publication of the material and was co-authored by Jens Luebeck, Sihan Wu, Xiaohong Li, Alvin Ng, Patricia Galipeau, Carissa Sanchez, Annalise Katz-Summercorn, Hoon Kim, Sriganesh Jammula, Phoebe He, Howard Y. Chang, Roel Verhaak, Carlo C. Maley, Ludmil Alexandrov, Brian J. Reid, Rebecca Fitzgerald, Thomas Paulson, Vineet Bafna & Paul S. Mischel. The dissertation author was the primaryinvestigator and author of this material.

VITA

2010 – 2015 University of Washington
Bachelor of Science, Applied Computational Mathematical Sciences

2019 University of California San Diego
Master of Science, Computer Science

2015 – 2021 University of California San Diego
Doctor of Philosophy, Bioinformatics and Systems Biology

PUBLICATIONS

K. L. Hung, J. Luebeck, S. R. Dehkordi, C. Coruh, J. A. Law, W. J. Greenleaf, P. Mischel, V. Bafna, H. Y. Chang, Targeted profiling of human extrachromosomal DNA by CRISPR-CATCH. bioRxiv, (2021).

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, (2021).

O. S. Chapman, J. Luebeck, S. Wani, A. Tiwari, M. Pagadala, S. Wang, J. D. Larson, J. T. Lange, I. T.-L. Wong, S. R. Dehkordi, S. Chandran, M. Adam, Y. Lin, E. Juarez, J. T. Robinson, S. Sridhar, D. M. Malicki, N. Coufal, M. Levy, J. R. Crawford, S. L. Pomeroy, J. Rich, R. H. Scheuermann, H. Carter, J. Dixon, P. S. Mischel, E. Fraenkel, R. J. Wechsler-Reya, V. Bafna, J. P. Mesirov, L. Chavez, The landscape of extrachromosomal circular DNA in medulloblastoma. bioRxiv, (2021).

J. Pang, N.-P. 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, J. Califano, Extrachromosomal DNA in HPV mediated oropharyngeal cancer drives diverse oncogene transcription. Clin. Cancer Res., (2021).

E. N. Bergstrom, J.-C. Luebeck, M. Petljak, V. Bafna, P. S. Mischel, R. Harris, L. B. Alexandrov, Comprehensive analysis of clustered mutations in cancer reveals recurrent APOBEC3 mutagenesis of ecDNA. Nature (in press), (2021).

S. R. Dehkordi, J. Luebeck, V. Bafna, FaNDOM: Fast nested distance-based seeding of optical maps. Patterns (New York, NY). 2 (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. Nat. Commun. 11, 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. Nat. Genet., 1–7 (2020).

C. J. Rocca, J. N. Rainaldi, J. Sharma, Y. Shi, J. H. Haquang, J. Luebeck, P. Mali, S. Cherqui, CRISPR-Cas9 Gene Editing of Hematopoietic Stem Cells from Patients with Friedreich's Ataxia. Mol. Ther. - Methods Clin. Dev. 17, 1026–1036 (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 (2019), doi:10.1038/s41586-019-1763-5.

V. Deshpande, J. Luebeck, N. P. D. Nguyen, M. Bakhtiari, K. M. Turner, R. Schwab, H. Carter, P. S. Mischel, V. Bafna, Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat. Commun. 10 (2019), doi:10.1038/s41467-018-08200-y.

J.-C. Luebeck, Alignment Methods for Optical Maps (2019).

N. D. Nguyen, V. Deshpande, J. Luebeck, P. S. Mischel, 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. 46, 3309–3325 (2018).

D. Zhao, M. G. Badur, J. Luebeck, J. H. Magaña, A. Birmingham, R. Sasik, C. S. Ahn, T. Ideker, C. M. Metallo, P. Mali, Combinatorial CRISPR-Cas9 Metabolic Screens Reveal Critical Redox Control Points Dependent on the KEAP1-NRF2 Regulatory Axis. Mol. Cell. 69, 699–708 (2018).

V. E. Gray, R. J. Hause, J. Luebeck, J. Shendure, D. M. Fowler, Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell Syst. 6, 116–124 (2018).

J. P. Shen, D. Zhao, R. Sasik, J. Luebeck, A. Birmingham, A. Bojorquez-Gomez, K. Licon, K. Klepper, D. Pekin, A. N. Beckett, Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. Nat. Methods. 14, 573–576 (2017).

ABSTRACT OF THE DISSERTATION

Reconstructing and Profiling Extrachromosomal DNA

by

Jens-Christian Luebeck

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2021

Professor Vineet Bafna, Chair
Professor Julie Law, Co-Chair

Circular extrachromosomal DNA (ecDNA) are a genomic lesion occurring in tumors, and represent a foundational, growing frontier in cancer biology. The discovery that focal amplifications exist in multiple topologies and arise by different mechanisms has enabled cancer researchers to study the consequences of different types of focal amplification – revealing that focal amplifications like ecDNA lead to worse patient survival. Consequently, there is an urgent need for bioinformatic methods and tools to study these focal amplifications, particularly ecDNA. This thesis describes novel tools and methods which can be used to study ecDNA, and other focal amplifications. It also demonstrates how those tools and methods can be used to profile focal amplifications across different cancer types, ultimately revealing novel biology about the structure, function and genesis of ecDNA in different contexts.

I first present two methods, AmpliconReconstructor (AR) and FaNDOM, which incorporate optical mapping data to resolve the structures of ecDNA and other focal amplifications. AR incorporates both optical mapping and NGS data and builds upon a prior method for ecDNA detection with NGS data, AmpliconArchitect (AA). FaNDOM utilizes optical mapping solely and enables the rapid characterization of large structural variants using assembled OM contigs or individual OM molecules.

I also describe the landscape of ecDNA in oropharyngeal squamous cell carcinoma, demonstrating that both human and hybrid human-viral ecDNA exist and are associated with distinct patterns of transcriptional splicing. Visualizations of the rearranged ecDNA structures and overlaid transcription-level data reveal the overexpression of genes carried on ecDNA.

Lastly, I describe the genesis of ecDNA in Barrett's esophagus, the precursor tissue of esophageal adenocarcinoma. We utilized methods for profiling ecDNA, such as AmpliconClassifier, to demonstrate that ecDNA exist in pre-cancerous tissue, are associated with worse histology, that they are subsequently found again in cancer, and that they tend to undergo positive selection during the malignant transformation. These findings solidify ecDNA as a potent driver of cancer, and not an opportunistic passenger.

INTRODUCTION

Human chromosomes typically come in 23 neat pairs, where an individual's parents are each responsible for one copy. Miraculously, each human cell typically packs two full meters of chromosomal DNA into a nucleus that is only one-hundredth the diameter of a human hair. Despite being exceedingly dense, dividing cells still manage to assort their chromosomes into pairs in a remarkably coordinated biological ballroom dance. This organized condensation of DNA, pairing of chromosomes, followed by cell division and a return to an uncondensed state, is a lively molecular gigue which has served as the basis for life on earth for billions of years.

Molecular pathologists are well aware of the appearance of chromosomes, even from the early 20th century, having examined stained chromosomes through the lenses of microscopes. Today we are already rapidly closing in on the 100th anniversary of the first human "karyotype", describing the appearance of chromosomes, by American zoologist Theophilus Painter in 1922 (subsequently refined by Tjio and Levan in 1956).

However, not all cells conform nicely to this paradigm. The organized nature of chromosomes is completely scrambled in most cancer genomes. Instead of 46 total chromosomes, there are frequently dozens of additional chromosome copies, often consisting of fragments from several chromosomes joined together into a frightening and undecipherable DNA patchwork.

While studying the abnormal karyotypes of cancer cells from pediatric cancers in 1965, Cambridge pathologist Arthur Spriggs observed many tiny DNA particles appearing as conjoined dots. He termed these "double minutes" and speculated that these were likely DNA fragments detached from chromosomes. Only a medical anecdote, the phenomenon remained unstudied and was overlooked as a simple byproduct of cancer genome instability for decades.

In the late 1980's, Salk Institute researcher Geoffrey Wahl demonstrated double minutes were in fact circular DNA fragments which could spontaneously form when genomes were unstable. With advances in microscope imaging and the introduction of methods to fluorescently tag specific genes, researchers opened a new frontier in cancer biology. Research scientist Paul Mischel who pioneered groundbreaking research on these DNA circles began to more generally call them "extrachromosomal DNA", or ecDNA.

Mischel and colleagues discovered that the free floating ecDNA particles which often accumulated in extremely high numbers in cancer actually carried critical genes. These ecDNA-borne genes which would normally be found in a single copy on each chromosome are the same genes that provided instructions telling cells to grow and divide (proto-oncogenes). One key property of ecDNA that distinguishes it from chromosomal DNA, is ecDNA's ability to segregate randomly when cells divide. Unlike chromosomes that so neatly segregate to daughter cells, ecDNA will divide randomly, enabling some cells to end up with hundreds of copies of potent oncogenes after only a few cell divisions. They act as the throttle lever aboard a metaphorical runaway freight train.

In 2015, Mischel and UCSD computer science professor Vineet Bafna began a collaboration which sought to unravel the genomic structure of ecDNA. Bafna, who had helped to develop the algorithms which reconstructed the first human genome in 2001, had ample experience working with the various DNA sequencing technologies required to analyze ecDNA genomes. The goal would not be to just detect ecDNA, but to read the ecDNA sequence itself.

One of Bafna's students, Viraj Deshpande developed a tool for analysis of ecDNA using WGS data, enabling researchers to study and characterize the contents of ecDNA using sequencing data. The tool, called AmpliconArchitect (AA) identifies regions of the genome with genomic copy number increases associated with ecDNA, and forms a graph-based representation from which it can extract the signatures of ecDNA. This tool has served as a cornerstone for our subsequent ecDNA research.

Standard DNA sequencing data only measures small fragments of the genome at once, and due to the complex, rearranged nature of ecDNA - it is often ambiguous as to what the true ecDNA structure is, or if we are actually seeing ecDNA at all. Very difficult ecDNA reconstruction cases are rather common and short DNA sequencing data alone is inadequate. We developed methods which use optical mapping to resolve complex rearrangements such as those seen in ecDNA (Chapters 1-2). A related paper published in Nature which we worked on with Paul Mischel's lab utilized our method to provide a complete ecDNA reconstruction and was the subject of a New York Times Science article entitled "Scientists Are Just Beginning to Understand Mysterious DNA Circles Common in Cancer Cells", by Carl Zimmer.

We are continuing to expand the contexts in which we study ecDNA. I also analyzed how virally-mediated cancers, such as cervical and oropharynx cancer, which are driven by human papillomavirus infection, can form human-viral hybrid ecDNA (Chapter 3). The last chapter of the thesis (Chapter 4) describes analysis which enables us to better understand the early origins of ecDNA in tumors, by studying the frequency of ecDNA in a well-characterized precancerous lesion called Barrett's esophagus. There we are found that ecDNA is a very early event in tumorigenesis, and that it drives the transformation from benign neoplasm to malignancy.

CHAPTER 1. AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications

1.1 Introduction

Oncogene amplification is a major driver of cancer pathogenicity (1-5). Genomic signatures of oncogene amplification include somatic focal copy number amplifications (fCNAs) of relatively short (typically < 10Mbp) genomic regions (5,6). Multiple mechanisms cause fCNAs including, but not limited to, extrachromosomal DNA (ecDNA) formation (5,7,8), chromothripsis (9), tandem duplications (10,11) and breakage-fusion-bridge (BFB) cycles (12-14). EcDNA, in particular, enables tumors to achieve far higher oncogene genomic copy numbers and maintain far greater levels of intratumor genetic heterogeneity than previously anticipated, due to their non-chromosomal mechanism of inheritance - enabling tumors to evolve rapidly (5,15,16). In addition, the very high DNA template level generated by ecDNA-based amplification, coupled to its highly accessible chromatin architecture, permits massive oncogene transcription (17-19).

While ecDNA elements are a common form of fCNA (5), other mechanisms can also result in amplification with different functional consequences (6). Accurate identification and reconstruction of the fCNA structure not only describes the rearranged genomic landscape, but also represents a first step in identifying the generative mechanism – to ultimately gain understanding about an fCNA's biological consequence. Reconstruction of fCNA architecture involves determining the order and orientation of the genomic segments that constitute the amplicon. There are many methods to detect single genomic breakpoints from sequencing data, using a variety of different sequencing technologies (20-24). However, fewer methods are available to handle the more difficult problem of ordering and orienting multiple genomic segments joined by breakpoints into high confidence copy number-aware scaffolds, which are subsequently joined to enable complete reconstructions of complex rearrangements (6,25). This problem represents the key algorithmic challenge addressed by our work.

A previous method for characterizing the identity of focally amplified genomic regions, AmpliconArchitect (AA), generates an accurate breakpoint graph from next-generation sequencing (NGS) data (6). The graph encodes the genomic segments involved in fCNAs, their copy numbers, and breakpoint edges connecting the segments. Unambiguous reconstruction of fCNA architecture requires extracting paths and cycles from the breakpoint graph, to reveal the true structure of the underlying rearranged genome. However, in practice, path/cycle extraction is often confounded by duplications of large genomic regions inside an amplicon (Fig. 1.1a), imperfections in the graph arising from errors in estimation of segment copy numbers, erroneous and/or missing breakpoints.

We hypothesized that an approach combining the strengths of NGS with long-range genome mapping data would enable larger and more unambiguous reconstructions of fCNA architectures. We utilized optical mapping (OM) data, which provides single-molecule information about the approximate locations of fluorescently-labeled sequence motifs on long fragments of DNA (26). Importantly, optical mapping has orthogonal sources of error to DNA sequencing (27,28). Primary sources of error to consider include missing OM labels, uncertainty about the exact location of the label on the imaged molecule, and possible molecular chimerism. The median molecule (map) length used in assembly across all samples present in this study is 244 kbp (molecule N50 340 kbp), while the median segment length in breakpoint graphs in this study is 100 kbp, highlighting that OM data can span multiple junctions in breakpoint graphs derived from focal amplifications. The integrated NGS data and

OM data provide an orthogonal pairing of short- and long-range information about genomic structural variation.

We present a computational method for reconstructing large complex fCNAs, AmpliconReconstructor (AR). AR takes a breakpoint graph and long-range OM data as inputs. We utilize Bionano (Bionano Genomics, Inc., San Diego, CA) whole-genome imaging to generate single-molecule optical maps, which are *de novo* assembled into OM contigs (contig N50 72.8 Mbp). AR produces an ordering and orientation of graph segments, with fine-structure information from the breakpoint graph embedded into the large-scale reconstructions. As output, AR reports large-scale reconstructions of fCNA amplicons. We demonstrate the large-scale and fine-scale accuracy of AR using simulated OM data derived from seven cancer cell lines (6,21) (CAKI-2, GBM39, NCI-H460, HCC827, HK301, K562, T47D). Finally, we validate the fCNA reconstructions using cytogenetics.

1.2 AmpliconReconstructor (AR) Results

1.2.1 Overview of AmpliconReconstructor (AR)

We formulated the problem of fCNA reconstruction in multiple parts. First, alignment of genomic segments with optical map contigs. Second, the reconstruction of a genomic scaffold using OM data as a backbone. Third, the identification of the maximal simple paths in a graph where each node is an OM scaffold, for which the path is not a sub-sequence of another maximal simple path. AR separates these computational tasks into four primary modules (Fig. 1.2a,b). To address the first problem, we designed an OM alignment module, SegAligner, for aligning reference segments to assembled OM contigs generated by either the Bionano Irys or Bionano Saphyr instruments (Fig. 1.1b,c). SegAligner is critical as it can score placements of short genomic segments onto an OM contig, which wasn’t possible with other aligners. To address the second problem, we introduce two modules. First, a scaffolding module, which takes a collection of breakpoint graph segments aligned to OM contigs as input and creates scaffolds represented by directed acyclic graphs (DAGs) (Fig. 1.2c-e, Methods – “Reconstructing amplicon paths with AmpliconReconstructor”). The second module for scaffolding with AR involves a novel scaffold-path imputation technique (Fig. 1.2f-h, Methods – “Imputing paths in the scaffold with AmpliconReconstructor”) to connect breakpoint graph segments that may individually be too small to be informatively labeled and aligned with optical mapping (Fig. 1.2f). We address the final problem with a pathfinding module, which links scaffolds and searches for paths in a copy number (CN)-aware manner, to identify possible reconstructions of the amplicon. AR outputs a collection of sequence resolved paths supported by the linked scaffolds. We implemented a visualization utility, CycleViz, to show the integrated OM- and NGS-derived breakpoint graph data (Fig. 1.3). AmpliconReconstructor is implemented in Python, and SegAligner is implemented in C++. Both tools are available publicly at https://github.com/jluebeck/AmpliconReconstructor.

1.2.2 AR accurately reconstructs simulated amplicons

We utilized multiple simulation strategies to measure the performance of AR (Fig. 1.4). We used 85 non-trivial breakpoint graph paths reported by AmpliconArchitect from 25 cancer cell lines (6) as a ground-truth set of amplicon structures, and a separate simulation of 20 *de novo* simulated circular ecDNA structures. We first present the results of the 85 breakpoint graph paths. These paths included cyclic (37 paths) and non-cyclic paths (48 paths) with lengths varying from 260 kbp to 2.8 Mbp (median 1.1 Mbp) and the number of graph segments varying from 3 to 47 (mean 17.5 segments). These paths were used as a reference from which we simulated OM molecules. (Methods – “Simulation of amplicons to measure AR performance”). Simulated molecules were assembled into contigs using the Bionano Assembler (29,30).

For each of the 85 simulated amplicons, we ran AR on the corresponding breakpoint graph and the *de novo* assembled contigs, and examined four different variables that could affect the performance of AR. First, we tested AR performance using SegAligner for OM alignment, versus AR using other OM alignment tools to replace SegAligner. Second, we evaluated the performance of AR across a range of amplicon copy numbers. Third, we measured performance with false edges present in the breakpoint graph. Finally, we generated and tested mixtures of three similar amplicons from the same samples, simulated with different amplicon copy numbers, to measure the effects of potential amplicon heterogeneity on AR performance.

We measured the accuracy of AR by computing precision and recall across the four simulation conditions. As precision and recall could be quantified in multiple ways when comparing ground-truth and reconstructed simulation paths, leading to different understandings of performance, we described three ways of measuring the similarity of the paths (Length (bp), Nseg, Breakpoint; Methods – “Measuring AR simulation performance”), based on the longest common substring (LCS) between ground-truth and reconstructed path sequences. We report the Length (bp) measurement in the analysis described here, while results with other measurements are presented Figure 1.5.

AR using SegAligner achieved a mean F1 score (harmonic mean of the precision and recall) of 0.88 for the highest copy number level (CN 20) and 0.68 for the lowest copy number level (CN 2) (Fig. 1.2i, Fig. 1.5). In contrast, when OMBlast (31) or Bionano RefAligner (29,32) were used in place of SegAligner, we noticed a decrease in both precision and recall. For RefAligner and OMBlast, respectively, we report mean F1 scores of 0.52, 0.43 for CN 20, and 0.42, 0.41 for CN 2. When imputation was omitted from AR, the mean F1 score for CN 20 decreased from 0.88 to 0.70. We observed similarly consistent trends using other methods of measuring precision and recall – Nseg and Breakpoint (Fig. 1.5). Large duplications inside a rearranged amplicon represent a challenging case to reconstruct. We identified 60 duplications of one or more graph segments (mean length 281 kbp) in the simulated amplicons, and we report that AR resolved 75% (45) of these duplications. We saw some cases of 'assembly failure,' where no paths differing from the reference genome involving the amplicon segments were assembled. Figure 1.2i shows cumulative precision and recall values for AR using SegAligner (with and without imputation), and with assembly failures filtered.

To understand the reasons for loss of performance on a small number of simulation cases, we examined the results from the CN 20 simulation where individual reconstructions showed either precision or recall < 0.6. We manually examined the results from the 85 total cases and found that of 13 amplicons with precision below this threshold, nine cases showed signs of assembly failure, while three had incorrect reconstructions likely on account of graph complexity. The remaining case showed an issue with incorrect scaffold linking. Of 14 amplicons having recall below the threshold, nine cases showed signs of assembly failure, while five had highly segmented breakpoint graphs making it difficult for AR to identify anchoring alignments around the breakpoints, leading to an incomplete reconstruction.

False edges in the breakpoint graph increase the possible number of path imputations that AR considers, potentially leading to erroneous scaffolds. On simulated CN 20 amplicons, we added additional false edges between existing graph segments. We tested three scenarios with the proportion of additional false edges ranging from 0%, 50% and 100% of the number of true graph edges. The three scenarios resulted in nearly identical mean F1 scores of 0.881, 0.880, 0.881 across the 85 amplicon simulations (Fig. 1.6a), highlighting the robustness of the path imputation method.

To understand how AR performed when faced with amplicon heterogeneity, we designed a simulation study involving 123 combinations of breakpoint graph paths where each combination was derived from paths found in a single sample, generated at varying copy number mixtures. We simulated amplicons from heterogeneous mixtures with (1) a single dominant amplicon (CNs 20-2-2); (2) a linear mixture of CNs (CNs 20-15-10); (3) equally abundant amplicons (CNs 20-20-20). We report mean F1 scores of 0.92, 0.89, and 0.91, respectively for the three cases. To explain the increase in performance of the mixture simulations as compared to the single amplicon simulations, we hypothesize that the greater total number of molecules improved the assembly process. Regardless, the high similarity between the precision and recall in each mixture case (Fig. 1.6b) indicates AR can reconstruct an accurate amplicon path even in the context of heterogeneity.

Lastly, we designed a simulation strategy not reliant on prior AA-generated paths. Instead, we generated 20 *de novo* simulated rearranged circular amplicons (median size 2.0 Mbp, mean segments 9.3) and replaced the hg19 reference used to generate background molecules with a simulated tumor genome generated with SCNVSim (33). AR's performance on these cases achieved a mean F1 score of 0.860 (0.731 when assembly failures included). The distributions of F1 scores for the 20 *de novo* cases and the 85 AA-derived cases were not statistically different between the 85 AA-derived amplicons and the 20 *de novo* simulated amplicons (two-tailed Mann-Whitney U-test, p-value = 0.1996, test statistic = 631.0). Based on these results, we found AR to be robust, and to outperform other methods for resolving fCNA architecture.

1.2.3 AR reconstructs ecDNA in multiple forms

Three cell lines in our study were previously reported to contain ecDNA (5) - GBM39, NCI-H460, and HK301. We previously analyzed (17) glioblastoma multiforme (GBM) cell line GBM39 using a preliminary version of AR with Bionano RefAligner (29,32) and manual merging of graph segments. Re-analysis reproduced an unambiguous 1.26 Mbp *EGFRvIII*-containing circular ecDNA identical to the previously published structure (17) (Fig. 1.7). The entire amplicon was captured by a single non-circular OM contig, with circularity confirmed by an overlapping breakpoint graph segment aligned to both ends of the contig.

Prior studies of ecDNA have documented their integration into chromosomes over time, linearizing and appearing as homogeneously staining regions (HSRs), often in non-native locations (5,7,15). In a previous study (5), The GBM cell line HK301 had been cytogenetically determined to have circular ecDNA; however, we observed from FISH (fluorescence in situ hybridization) data that the sample's ecDNA had become HSR-like at the time of this study (Fig. 1.8a). AA reported a breakpoint graph supporting amplification of both *EGFRvIII* and *EGFR* wild-type (Fig. 1.8c), however an unambiguous reconstruction from the graph alone was not possible. The AR reconstruction of the HK301 fCNA indicated a complex cyclic structure supported by three contigs (Fig. 1.8d), which explained 98.1% of the amplified genomic regions. The graph segments came predominantly from chr7, but also included two small regions (2890 bp, 4591 bp) from chr6 (Fig. 1.8c,d). We noted a ~20 kbp deletion inside *EGFR*, showing a lower CN than the surrounding region, but which was still amplified over the baseline regions of chr7. This indicates heterogeneity of *EGFR* wild-type/vIII mutation status. Despite the heterogenous status of this allele, AR reconstructed the *EGFRvIII* version – which is the dominant form of the amplicon (Fig. 1.8d).

The lung cancer cell line NCI-H460 has previously been documented to bear *MYC* amplification (34), and our cytogenetic analysis showed evidence for both its HSR-like and ecDNA amplification (Fig. 1.8e,f). Despite the heterogeneous nature of the amplicon's integration status, AA generated a breakpoint graph for a contiguous 2.15 Mbp region of chr8 (Fig. 1.8g). AR reconstructed a single 4.10 Mbp structure supported by five OM contigs (Fig. 1.8h). This structure contained all amplified segments from the breakpoint graph and explained the relative ratios of breakpoint graph segment copy numbers. For example, segment chr8:129,404,278-129,591,422 appeared 4 times, chr8:128,690,200-129,404,277 (carrying *MYC* & *PVT1*) appeared twice, chr8:129,591,423-129,911,811 appeared twice, and chr8:129,911,812-130,640,594 appeared once, making the ratios consistent with the estimated graph segment copy numbers (46, 25, 25, 12, respectively; Fig. 1.8g). The status of the long non-coding RNA *PVT1* (a known regulator of *MYC*) (35) on this amplicon is heterogeneous, as one copy of *PVT1* does not contain breakpoints, while the other shows a disrupted copy of *PVT1*. AR also identified a self-inversion at the end of the amplicon (black arrows in Fig. 1.8h), suggestive of an alternating forward-backward orientation (segmental tandem aggregation with inversion) of the amplicon in the agglomerated ecDNA.

We previously documented a circular amplicon containing an integrated human papillomavirus-16 (HPV16) genome (6), and we hypothesized that AR could help resolve the location of viral insertion in the host genome. We simulated a 1 Mbp circular amplicon with the 7.9 kbp HPV16 genome randomly inserted. AR was able to reconstruct the circular ecDNA structure and identified the integration point of HPV16 (Fig. 1.9) despite the viral genome having no OM labeling sites, suggesting that AR would serve as useful method for validating the existence of genomic oncovirus integrations suggested by NGS data.

In summary, AR reconstructed paths that were consistent with the expected ratios between amplified segment copy numbers and graph structures in GBM39, HK301, and NCI-H460, explaining 99.9%, 98.1%, and 100% of the amplified genomic content in the breakpoint graphs for each cell line, respectively. Furthermore, the AR reconstructions of ecDNA in HSR-like form lend additional evidence to the agglomerative model of ecDNA integration (Fig. 1.8b) (8,36,37).

1.2.4 AR reconstructs a rearranged Philadelphia chromosome in K562

The classical model of the *BCR-ABL1* fusion involves a reciprocal translocation of the q arms of chromosomes 9 and 22 (Philadelphia chromosome) (38). However, this mechanism alone does not explain the copy number amplification of *BCR-ABL1* fusion commonly observed in chronic myeloid leukemia (CML), highlighting a need for methods to better understand the genesis of the *BCR-ABL1* amplification (39,40). To reconstruct the fine structure of a Philadelphia chromosome, we used the CML cell line K562 where a *BCR-ABL1* fusion had previously been reported (41).

The AA-reconstructed breakpoint graph for the *BCR-ABL1* fCNA in K562 contains 8.5 Mbp of amplified genomic segments (Fig. 1.10a). The graph shows signatures of complex rearrangements alongside the *BCR-ABL1* fusion, which AA predicted to have a copy number of 17 (Fig. 1.10a). We generated both Bionano Irys and Bionano Saphyr OM data for K562 cells and observed consistent results in the independent reconstructions of amplicons from both sources (Fig. 1.11a,b). Using the breakpoint graph and OM contigs, AR reconstructed a complex linear structure that chained together 1.7 Mbp from chr22 (containing *BCR*), 548 kbp of chr9 (containing *ABL1*), and multiple regions from chr13 (732 kbp; including a disrupted copy of *GPC5*) (Fig. 1.10b). In Figure 1.10b, we show one possible scaffolding of the given regions, whose structure was reproduced in both Saphyr and Irys datasets. AR also reported junctions between segments in the breakpoint graph where NGS-derived breakpoint edges were not reported, as indicated by the missing half-height grey bars between adjacent genomic segments in the genome tracks of Figure 1.10b. While AR explains many of the amplified segments in this amplicon, we note that there is additional copy number variation in this amplicon it does not explain. For instance, the *BCR* and ABL-containing segments have an elevated CN over the segments on chr13.

We performed FISH experiments using combinations of probes for *BCR*, *ABL1*, *GPC5*, and chr22 centromere probe CEP22. The FISH images confirmed the co-localization of the *BCR-ABL1* fusion and *GPC5* on a common HSR-like structure (Fig. 1.10c) and validated the status of the *BCR-ABL1* fusion as being located on chr22 (Fig. 1.12).

In addition to the reconstruction reported in Figure 1.10b, AR identified other scaffolds, indicating that the genomic structure surrounding the *BCR-ABL1* translocation may be varied across the multiple copies (Fig. 1.11c,d; Fig. 1.13a-f). In particular, the genomic segment bearing *CLTCL1* appears in both forward and reverse directions (Fig. 1.13b,c). Other amplified regions of chr13 include a self-inversion at the 3' end of *GPC5* (Fig. 1.11c,d, Supplementary Fig. 1.13e). A scaffold from the Irys-based reconstruction indicated a secondary reconstruction could be joined with the *BCR-ABL1* reconstruction (Fig. 1.11d; overlap of segment 20). From the AR reconstructions of the *BCR-ABL1* amplicon and the co-existence of *BCR*, *ABL1* and *GPC5* in overlapping locations, as shown by FISH (Fig. 1.10c 'Zoom'), AR enabled us to hypothesize a potential sequence of events by which the fCNA formed. The AR reconstructions support the formation of the *BCR-ABL1* translocation (Fig. 1.13g;i-ii) followed by incorporation of chr13 regions (Fig. 1.13g;iii-iv), which subsequently undergo rearrangement (Fig. 1.13g;v), and ultimately a series of inverted repeats, possibly mediated through dicentrism (Fig. 1.13g;vi). These results are consistent with previous reports using cytogenetic approaches to observe the presence of additional chromosomal segments besides chr9 and chr22 involved in the Philadelphia chromosome (30,31).

1.2.5 AR enabled the reconstruction of a breakage-fusion-bridge

The BFB mechanism of genomic amplification involves the loss of telomeres and subsequent fusion of two sister chromatids (12,13). In subsequent cellular division, the asymmetric breaking of the fused chromosome structure results in one daughter cell acquiring additional pieces of the previously fused chromosome. The structure of various BFBs have been analyzed using cytogenetic techniques (14) and by computational models that predict BFB presence from copy number counts (42,43). Both methods are imprecise, to a degree, and may fail to capture the fine structure of the BFB or handle imprecise copy number counts and/or additional structural variants (SVs) inside the BFB. We deployed AR on the HCC827 lung cancer cell line where AA and cytogenetics suggested a chr7 BFB, though an unambiguous structure was not identifiable (5,6).

We observed a banded pattern of *EGFR* and CEP7 (a chr7 centromeric D7Z1 repeat) in a DNA FISH experiment on HCC827 cells, suggestive of a BFB mechanism (Fig. 1.14a). AA generated a breakpoint graph of a 4.2 Mbp amplified region of chr7 containing *EGFR* (Fig. 1.14b). The amplified BFB segments in the AA output ranged in size from 217 kbp to 1176 kbp. AR enabled the reconstruction of 16 unique OM scaffolds which, when combined, enabled the reconstruction of the BFB structure (Fig. 1.14c,d). The five most informative single scaffolds ranged in size from 750 kbp to 2.3 Mbp, containing multiple junctions which validate the order and orientation of the BFB breakpoint graph segments, resulting in a 9.4 Mbp amplicon, hereafter referred to as a BFB repeat unit. The BFB repeat unit was amplified across the chromosome (Fig. 1.14a, e-f). AR also revealed a region outside the AA amplicon, near the centromere of chr7, which explained the observed *EGFR* and CEP7 repeat (F). In segment B, we observed a 600 bp deletion across the entire BFB repeat unit and an 11 kbp inversion. The latter is labeled throughout Figure 1.14 with a black asterisk and only appears when segment B is duplicated and inverted, suggesting that the SV arose midway through the formation of the BFB. While some BFBs may result in double-minute amplicons (7), AR suggested, and FISH analysis confirmed that the HCC827 BFB does not contain a circular extrachromosomal version of the BFB cycle.

When the AR scaffolds were combined with the copy number data present in the breakpoint graph, we could manually identify a complete BFB structure consistent with the theoretical model of BFB formation (44). A putative sequence of BFB cycles and additional structural variation results in the final BFB structure is shown in Fig. 1.14f (also Fig. 1.15a,b). Without AR, the copy number information and the theoretical model together could not have reconstructed this BFB, as it contains heterogeneous interior structural variants. We further validated the BFB patterning in HCC827 cells with multi-FISH for segments A, C, and D from the BFB, using FISH (Fig.1.14e, Fig. 1.15c). Together, these results show the ability of AR to enable the resolution of a BFB-driven fCNA, even in the presence of additional structural heterogeneity.

In addition to the *EGFR*-bearing amplicon, AA detected five other amplicons containing *MYC* and *NCOA2*, among other oncogenes. The graphs were complex (Fig. 1.16a) and in many cases AA did not identify discordant edges between distinctly amplified regions. Given the dearth of breakpoint edges, we combined the amplicon breakpoint graphs for all six HCC827 amplicons and ran AR on the combined graph, containing 555 segments. AR identified 206 contigs having alignments to one or more graph segments. AR reconstructed multiple possible scaffolds and captured overlapping subsets of amplicon regions from different graphs, suggestive of possible amplicon heterogeneity. One scaffold showed *NCOA2* located on a native region of chr8, while another showed *NCOA2* joined to *MYC* through a segment of chr21 (Supplementary Fig. 12b,c).

1.2.6 Additional focal amplifications reconstructed by AR

In breast cancer cell line T47D, where the AA breakpoint graph suggested amplification of a 634 kbp region, AR reconstructed a 430 kbp segmental tandem duplication, containing oncogene *GSE1* (Fig. 1.17a,b). This highlighted the ability of AR to also reconstruct classes of ultra-large, albeit less-complex SVs.

In renal cancer cell line, CAKI-2, AA generated a breakpoint graph spanning 12.0 Mbp, joining regions from chr3 and chr12 (Fig. 1.17c,d). Despite the lower overall copy number of this amplicon (~5), AR still reconstructed a 13.1 Mbp amplicon explaining 99.9% of the amplified genomic content in the AA-detected fCNA. Both amplicons for CAKI-2 and T47D appear to be intrachromosomal events given the AR results.

Across the focal amplifications we studied in seven cancer cell lines, we report 64 individual amplified breakpoints detected by both AA and validated by AR. Taken together, our data demonstrate the power of AR to combine NGS and OM data to elucidate a variety of complex fCNAs commonly found in cancer - enabling a deeper understanding of the fundamental mechanisms that give rise to fCNAs and promote cancer pathogenesis.

1.2.7 Integration points of focal amplifications

Low frequency breakpoint edges, such as the ones indicating integration points may not appear in the NGS breakpoint graph and may not be seen in assembled OM contigs. Using alignments of single-molecule optical maps, generated by the Bionano RefAligner molecule alignment pipeline, we gathered molecules with split alignments joining a partial alignment inside the amplicon region with a partial alignment outside the amplicon region. For H460, K562, CAKI-2, HCC827 and T47D, OM coverage was deep (> 100) allowing us to cluster split alignments into OM-derived breakpoint clusters suggestive of low-frequency integration points. Requiring that each breakpoint cluster have 10 or more molecules suggesting the same junction (within 25 kbp on either side), we identified four integration point candidates (Table 1.1).

Visualized with MapOptics (45), H460 showed a single integration point between amplicon region chr8:129410000 and non-amplicon region chr12:7660000 (Fig. 1.18a). K562 showed two integration points. The first joined amplicon region chr13:81120000 and non-amplicon region chr1:142890000 (Fig. 1.18b). The second joined amplicon region chr13:93260000 and non-amplicon region chr1:142890000 (Fig. 1.18c). The proximity of these two integration points suggests a left and right boundary for the integration of the K562 *BCR-ABL1* amplicon. CAKI-2 showed one integration point joining amplicon region chr12:88300000 and non-amplicon region chr6:168380000 (Fig. 1.18d).

HCC827 and T47D did not show any such integration points with 10+ molecules of support, which is consistent with the finding that these were chromosomally derived focal amplifications (BFB and segmental tandem duplication respectively).

1.2.8 AR provides a reconstruction improvement over AA

AA can identify some putative paths and cycles in the breakpoint graph only using NGS data. We demonstrated that for complex amplicons, AR provides an improvement to the fraction of the amplified genomic segments in the heaviest reconstruction path or cycle compared to the heaviest path or cycle generated by AA (Fig. 1.19a). OM data may suggest additional amplicon junctions not observed in NGS. The segment junctions observed in the AR output was equal to (GBM39, T47D) or larger than (CAKI-2, H460, HCC827, HK301 and K562) the number of junctions suggested by the AA breakpoint graph alone (Fig. 1.19b).

Dixon et al. (21) used an integrative approach to detect structural variation and associated breakpoints using a combination of NGS, OM data and other sequencing modalities. We identified four cell lines shared between our studies for which Dixon et al. reported breakpoints identified by their integrative approach. We observed that in regions analyzed by AR, more breakpoints were detected with AR than with the integrative approach, though there were some breakpoints indicated by Dixon et al. which were not observed by AR (Fig. 1.19c). In those cases, the majority of breakpoints not observed by AR joined amplicon regions to regions outside the amplicon (CAKI-2: 2 of 3 non-AR breakpoints, H460: 1 of 1 non-AR breakpoints, K562: 11 of 16 non-AR breakpoints). In H460, the one breakpoint not observed by AR was the integration point we later detected, suggesting that these are lower frequency breakpoints perhaps related to integration.

1.3 Discussion

Revealing the architecture of fCNAs, particularly at a large scale, is critical to understanding the functional consequences. For instance, rearrangements present in fCNAs frequently increase oncogene copy number (46), disrupt gene structure (47), and lead to dysregulation of chromatin (17-19). Accurate reconstruction of fCNA architecture can provide insights into the mechanisms of formation, leading to an improved understanding of the biological consequences of fCNA that would not be available solely from methods characterizing individual breakpoints. AR does not yet automatically produce a prediction of the biological mechanism of amplification. Thus, the AR reconstructions still require some manual interpretation based on the visualized results.

While previous methods have characterized complex structural variation using both OM and NGS data (21,48), these methods have typically focused on individual variants and breakpoints[42]. OM tends to detect larger SVs than NGS alone and is less affected by mapping issues on low complexity breakpoints (21,24). We have demonstrated that NGS data, when incorporated with OM can be used to resolve fine-mapped breakpoints suggested by OM. Indeed, some of the individual junctions reported by AR in these cell lines were already known (21) (Fig. 1.19c). However, AR represents a robust, comprehensive algorithmic approach to reconstructing the fine-scale and large-scale structure of an fCNA through the propagation of NGS-derived breakpoint information into larger scaffolds.

Many variables affect the ability to resolve fCNA. Importantly, the complexity and structure of the fCNA, as well as the length of the reads or genome maps. Further compounding the difficulty of fCNA reconstruction, we note that different sequencing modalities do not overlap perfectly in the breakpoints they detect (21,24). Based on our findings, we suggest the resolution of chained breakpoints should be spanned by long-range sequencing data with length sufficient to anchor the chain on both ends. We attribute much of the success of AR for resolving fCNAs to the long molecule length (244 kbp median) in comparison with the length of amplified genomic segments in the breakpoint graph (100 kbp median).

The paths reconstructed by AR represent possible reconstructions of an fCNA and may contain multiple similar explanations for the fCNA architecture. This may be in part due to amplicon heterogeneity, limitations of the optical map assembly process, or errors in linking scaffolds across overlapping graph segments. Despite the integrative nature of AR's inputs, breakpoints may still be missed in the amplicon. One traditionally difficult case to reconstruct involves nested duplication of genomic segments inside an amplicon. Unless a significant fraction of reads or genomic maps have a length greater than the duplicated element, the duplication status may not always be accurately resolved, leading to ambiguity. Multiple tandem duplications can also give rise to a cyclic breakpoint graph structure. However, in that case the same breakpoint would be reused repeatedly, and evidence points against that possibility (5,6,46). Instead, ecDNA-derived mechanisms provide a simpler and arguably more correct interpretation of cyclic graph structures, as validated by cytogenetics and comparison to Circle-seq experiments (46,47).

Genomic structural heterogeneity is problematic for any genome reconstruction, including focal amplifications. Despite the change in topology between linear HSR-like and circular ecDNA fCNAs, the breakpoint graphs between both circular and linear forms of the same samples are highly similar (6), suggesting ecDNA genomic structure is often not altered during reintegration. While we analyzed data from cancer cell lines, sequencing data collected from patients may introduce more sources of complex genomic structural heterogeneity.

Assembled OM contigs may fail to capture rare instances of structural heterogeneity in the genome. However, previous results suggest that focal amplifications conferring a fitness advantage to cancer cells are clonally amplified (5,49), allowing for accurate reconstruction of the dominant structure.

AR produced a high-confidence reconstruction of the K562 *BCR-ABL1* focal amplification yet copy number variance in this amplicon not explained by AR may be due to structural heterogeneity across the many copies of the amplicon. Additional copy number changes in K562 near *BCR* and *ABL1* which are not directly explained by the amplicon through edges identified by NGS reveal limitations to our current method, or possible inaccuracies. Such cases may indicate additional amplicon segments outside the regions reconstructed by AR, suggesting that the true amplicon structure may extend beyond the regions we have captured. Despite the presence of the AR-supported and FISH-validated HSR-like status of the *BCR-ABL1* translocation in K562, there does not exist a completely validated model that explains the increased copy number of *BCR-ABL1* in one single location. We cannot rule out the possibility that the *BCR-ABL1* amplification in K562 is mediated through an ecDNA stage (50), given the transient nature of the emergence and retreat of ecDNA (15) and the highly rearranged genomic landscape surrounding *BCR-ABL1*.

We have not yet adapted AR to accept data generated by other long-range sequencing modalities, breakpoint graphs generated by other tools or to accept breakpoint graphs derived from non-amplified rearrangements. Recent advances in other long-range sequencing technologies (51) highlight the need to adapt the AR algorithm. With modified protocols, nanopore reads may routinely surpass 150 kbp in length - sufficient to frequently chain multiple breakpoints in fCNA. We plan to address this in future methods development. Other sequencing modalities involving NGS with modified sample preparation, such techniques based on Hi-C and linked reads, have shown the ability to reveal additional genomic breakpoints without an additional sequencing instrument (21,24). While *de novo* breakpoint graph construction is not a part of the AR algorithm, we acknowledge that such techniques would be valuable to adapt for breakpoint graph generation.

Methods to accurately characterize fCNAs will enable better classifications of cancer subtypes and their associated prognoses. The accurate, multi-megabase scale, complex fCNAs reconstructed by AR not only describe fine structural features of fCNA architecture, but also reveal mechanistic signatures of fCNA formation, allowing for future interrogation of the relationship between fCNA architecture and the biological consequences of their structure.

1.4 Methods

1.4.1 Cell culture

NCI-H460, K562, and HCC827 cells were obtained from ATCC and cultured in RPMI-1640 media supplemented with 10% FBS. HK301 cells were cultured as neural spheres in DMEM/F12 media supplemented with B27, EGF (20 ng/ml), FGF (20 ng/ml), and heparin (1 ug/ml). All cells were incubated under standard conditions.

1.4.2 Metaphase chromosome spreads

Metaphase cells were enriched by treating cells with Karyomax (Gibco) at a final concentration of 0.1μg ml^{-1}. Cells were collected, washed in PBS, and resuspended in 75mM KCl for approximately 15 minutes at 37°C. Cells were fixed by addition of an equal volume of Carnoy’s fixative (3:1 methanol:glacial acetic acid). Cells were washed three additional times in Carnoy’s fixative and dropped onto humidified glass slides.

1.4.3 FISH

Metaphase spreads were equilibrated in 2x SSC (30mM sodium citrate, 300mM NaCl, pH 7) for approximately 5 minutes. They were dehydrated using successive washes of 75%, 85%, and 100% ethanol for two minutes each and allowed to dry. FISH probes were diluted in hybridization buffer (Empire Genomics) and added to metaphase spreads on slides, along with 22mm^2 coverslips. Samples were denatured at 70-75°C for 30 seconds – 2 minutes. Probe hybridization was performed at 37°C for around 3 hours or overnight in a humid and dark chamber. Samples were washed successively in 0.4x SSC and 2x SSC with 0.1% Tween-20. Samples were incubated with DAPI (0.1μg ml^{-1} in 2x SSC) for 10 minutes, then washed with 2x SSC and briefly rinsed with H_2O. Samples were mounted with Prolong Gold, #1.5 coverslips, and sealed with nail polish. All FISH experiments involved the analysis of at least three independent images and representative results are shown in the figures present in the study.

1.4.4 Microscopy

Confocal microscopy was performed on a Leica SP8 Confocal microscope with white light laser and Lightning deconvolution. Fluorescent microscope images were acquired using an Olympus BX43 microscope with a QiClick cooled camera. Images were subsequently analyzed in ImageJ (52) (using the Bio-Formats plugin (53)), to perform cropping, add scale bars and perform global adjustments to image brightness.

1.4.5 Acquisition of WGS data

We previously published (5,6) WGS data on SRA for six of the seven cancer cell lines (GBM39, NCI-H460, HCC827, HK301, K562, T47D) analyzed here. For CAKI-2, we used WGS data published by the Cancer Cell Line Encyclopedia on SRA.

1.4.6 Breakpoint graph generation

WGS data was aligned to hg19 with BWA-MEM (54) (version 0.7.17-r1188, default parameters), sorted and PCR-duplicate filtered with SAMtools (version 0.1.19-96b5f2294a) (55), and the resulting alignments along with SNV calls produced by Freebayes (56) (version v1.3.1-17-gaa2ace8) were supplied as input to the Canvas (57) CNV caller (version 1.39.0.1598). The alignments and CNV seeds were filtered using AmpliconArchitect's amplified_intervals.py module. Seeds exceeding 40 kbp with copy number 5 were subsequently analyzed with AmpliconArchitect. AmpliconArchitect outputs a breakpoint graph encoding segmented CN calls and the discordant reads connecting the segments. We note that in most cases identical amplicon regions are identified when CNV caller ReadDepth (58) is used for seeding instead.

We standardized the breakpoint graph generation process into a workflow called PrepareAA, available on GitHub: https://github.com/jluebeck/PrepareAA. We used the default parameters specified by PrepareAA in this analysis. To produce *in silico* digestions of breakpoint graph segments into reference optical maps, we used the generate_cmap.py utility in AmpliconReconstructor. This method for *in silico* digestion can produce labeling patterns for the Bionano Saphyr DLE-1 labeling pattern, while many previous methods for *in silico* digestion do not.

1.4.7 OM data generation

High molecular weight (HMW) DNA was extracted from GBM39, HCC827, HK301, and K562 cells using the Bionano Prep Blood and Cell Culture DNA Isolation Kit (Bionano Genomics #80004), with minor modifications to recover good quality HMW gDNA. As detailed below, the Nick, Label, Repair, and Stain (NLRS) and Direct Label and Stain (DLS) reactions were carried out for the Bionano Irys and Saphyr platforms, respectively. To generate the Irys data, DNA was nicked using Nt.BspQI nicking endonuclease (NEB), followed by labeling, repairing, and staining, using the Bionano Prep NLRS DNA Labeling Kit (Bionano Genomics #80001) along with recommended NEB reagents. To generate the Saphyr data, DNA was labeled with DLE-1 enzyme, followed by proteinase digestion and a membrane clean-up step, using the Bionano Prep DLS DNA Labeling Kit (#80005). BspQI-labeled DNA was loaded onto the Irys Chip (Bionano Genomics #20249) and the run conditions were manually optimized on the Irys system (Bionano Genomics #30047) to ensure efficient DNA loading into the nanochannels. DLS-labeled DNA was loaded onto a Saphyr Chip (Bionano Genomics #20319), and run conditions were automatically optimized on the Saphyr system (Bionano Genomics #60239) using the Saphyr Instrument Control Software to maximize DNA loading. Raw images generated by Irys were processed into BNX files using the Bionano software AutoDetect (26). Images from the Saphyr system were processed into digital BNX files via the Saphyr Instrument Control Software. For Irys data, molecules ≥150 kilobase pairs (kbp) were assembled into consensus genome maps using the Bionano Assembler (29,30) (version 5122), using default parameters; for Saphyr data, molecules ≥150 kbp were assembled into maps using Bionano Access (version 1.2.1) (29). Bionano Genomics separately provided Saphyr OM data for cell lines K562, T47D, NCI-H460, and CAKI-2. The methods by which OM data was generated for those four cell lines were previously published (21). All Bionano software utilized alongside this study is available from the Bionano Genomics, Inc. website (https://bionanogenomics.com/support/software-downloads/) under the Bionano Genomics, Inc. software license (https://bionanogenomics.com/company/legal-notices/).

1.4.8 Identifying unaligned amplicon contig regions with AmpliconReconstructor

AmpliconReconstructor coordinates the alignment of in-silico digested breakpoint graph segments to optical map contigs using SegAligner. Alternately, AR can take as input XMAP-formatted alignments produced by other alignment tools. If OM contigs with alignments to graph segments contain unaligned regions with between 20 and 500 unmatched labels, and 200 kbp to 5 Mbp in length, those regions are extracted and searched against the reference genome. The module ARAlignDetect calls SegAligner in the detection mode, which then aligns the extracted unaligned region of the contig(s) to the specified reference genome. If significant alignments are found between unaligned regions of the contig and chromosomal segments in the reference, those segments are extracted, and their identity is added to the collection breakpoint graph segments. Finally, a new breakpoint graph is output containing the newly detected segments.

1.4.9 Reconstructing amplicon paths with AmpliconReconstructor

Optical map alignments of segments with contigs are converted into a scaffold, which we define as a collection of alignments where the genomic distance between each pair of alignment endpoints is known. AR represents the scaffolded alignments as a directed acyclic graph (DAG), where the nodes are an abstract representation of each OM alignment. Directed edges connect adjacent alignment endpoints. Overlapping alignments are connected by special directed edges referred to as forbidden edges (Fig. 1.2h). Two nodes are only connected by a non-forbidden edge if the right endpoint of the source node has one or fewer labels of overlap with the left endpoint of the destination node. Each contig with at least one alignment to a graph segment will comprise an individual scaffold.

1.4.10 Imputing paths in the scaffold with AmpliconReconstructor

Some segments in the breakpoint graph may be too short to be uniquely aligned to an OM contig. AR attempts to impute corrected paths in the scaffold using the structure of the breakpoint graph. For every non-forbidden edge in the scaffold graph with a gap size less than 400 kbp, AR identifies breakpoint graph nodes corresponding to the source and destination endpoints, which we will denote as *s,* and *t*. AR then uses a constrained depth-first search (DFS) strategy to identify paths in the breakpoint graph between *s* and *t*. Finding all possible paths between two nodes may produce infinitely many solutions should a cycle exist between the two nodes, so the recursion is constrained to terminate if certain conditions are reached. The constraints used in the search procedure are:

1) The multiplicity of the segments in the candidate path must always remain less than or equal to the copy number of the segment as specified in the breakpoint graph.
2) If a candidate path reaches the destination vertex, its length in base-pair units must not be more than $\min(25000, 10000L_{\mathrm{p}})$ shorter than the distance between the source and destination vertices as expected given the scaffold backbone, where L_p is the length of the path in number of segments.
3) During path construction, the length of a candidate path must not exceed $\min(25000, 10000L_{\mathrm{p}})$ beyond the of the expected distance given the scaffold backbone.
4) The number of valid candidate paths connecting source to destination must not exceed 2^{10}.
5) The path may not form a trivial cycle from ultra-short breakpoint graph segments less than 100 bp long. Such cycles appearing in an NGS-derived breakpoint graph we assumed to be erroneous or artifactual.

As constraint #4 may cause failure of the DFS whereby a tractable number of paths is not found, AR implements a constrained BFS search as a fallback option, which is used when the DFS fails for that reason. By parsimony, shorter paths between two nodes are more likely to be correct, thus AR applies the same set of criteria for the BFS search, with the threshold in constraint #4 increased to 2^{16}.

All valid candidate imputation paths discovered by AR are scored by a fitting alignment procedure using SegAligner. To score a candidate path, the ordered path segments, as well as the first and last labels on the source and destination endpoints, are converted to a compound CMAP composed of the concatenated CMAPs of the individual segments. A fitting alignment is performed between the compound CMAP and the region of the contig between the alignment endpoints, using SegAligner. The path with the alignment score which most improves the junction score is kept. If no valid candidate path improves the score of the junction, it remains unimputed. The scaffold is then updated to contain the imputed breakpoint graph path.

1.4.11 Identifying linked scaffold paths with AR

Given the collection of scaffold DAGs, AR first searches for paths in the individual DAGs which represent heaviest paths in the scaffold DAG, where the weight of a path is the sum of the lengths of its segments in base pairs. AR stores the heaviest path(s) for each scaffold prior to performing scaffold linking.

AR leverages the two independent sources of information encoded in the breakpoint graph and OM contigs to link individual scaffolds. As the breakpoint graph segments are not detected to contain interior breakpoints, two endpoint alignments of the same breakpoint graph segment may be linked across two contigs. AR searches for prefix paths and suffix paths in each DAG. From the collection of prefixes and suffixes, AR searches for overlap between scaffolds generated from different contigs. Given that a contig can be assembled in either direction, overlapping reverse oriented suffixes or prefixes can also be matched. AR exhaustively finds sub-paths hitting either end of a scaffold DAG, which have overlap with other endpoint sub-paths, where the endpoint sequence of the scaffold may be assembled in either direction.

1.4.12 Finding reconstructions in the linked scaffold graph

Given the graph of linked scaffolds, AR searches for paths in the graph which conform to the ratio of estimated copy numbers between the graph's amplified segments. AR starts by searching for all paths in the graph which begin at endpoint nodes in the individual scaffolds. AR then uses a greedy approach to identify the longest unique paths which conform to the copy number restrictions. From the candidate paths, AR checks each path segment's multiplicity against the copy numbers encoded in the breakpoint graph in a ratio-dependent manner.

AR iterates over all the segment multiplicities in the reconstructed path, and at each multiplicity level determines the maximum estimated genomic copy number of path segments with that multiplicity. If a path segment has a multiplicity that is greater than the genomic copy number of that segment divided by the maximum copy number of all segments with multiplicities less than the given segment, then the path violates the copy number ratio check. AR allows each segment in the reconstructed path to exceed by 1 copy the copy number expected given the ratio between breakpoint graph copy numbers and segment multiplicity. If

n_p is the multiplicity of segment n in the candidate path, P, and n_g is the copy number of graph segment n in the breakpoint graph, then n_p must satisfy

$$n_{\mathrm{p}} \leq \max\left(c, \frac{n_{\mathrm{g}}}{m_{\mathrm{g}}}\right) + 1, \qquad \forall\, n \in P$$

where

$$m_{\mathrm{g}} = \max\big(i_{\mathrm{g}}, \forall\, i \in P, i_{\mathrm{p}} == c\big)$$

$$c \in \mathbb{Z}$$

$$n_{\mathrm{p}} > c > 0.$$

If a candidate path passes the copy number ratio check, it undergoes a pairwise comparison with other paths passing this criterion, to check for path uniqueness. A path is unique if it does not represent a subsequence of a previously identified unique path. Furthermore, no rotation of the path sequence may be a subsequence of a previously identified unique path. AR assess subsequence paths by computing a longest common substring between a candidate path and a previously identified unique path (Fig. 1.20). As the paths are first sorted by total alignment score prior to the iterative approach, this method is a greedy algorithm which prioritizes long, heavy paths as being more likely to be identified as unique non-subsequence paths. AR categorizes paths as being cyclic if the first and last scaffold graph node in the path are the same, and the path length is greater than two, as this distinguishes cyclic paths from paths which appear cyclic such as singleton paths or paths which represent segmental tandem duplications. Paths reported by AR are output in the AmpliconArchitect-cycles file format. Default parameters for AR are reported in Table 1.2.

1.4.13 Simulation of amplicons to measure AR performance

We used OMSim (63) (version 1.0) to simulate Bionano Irys OM data from the hg19 reference as well as from 85 non-trivial paths (i.e. not directly consistent with the reference genome) in AA-generated breakpoint graphs from 25 cancer samples and 20 *de novo* simulated ecDNA structures, including both cyclic and non-cyclic breakpoint graph paths (Fig. 4). OM molecules were simulated at 40x baseline coverage for each chromosome arm in hg19. The combined hg19 maps from all arms were assembled into a set of OM contigs using Bionano Assembler (version 5122). A similar process was performed using high-confidence breakpoint graph paths, which were converted to FASTA format and used for map simulation. For each simulated path, molecules were simulated at a range of copy numbers, and simulated molecules from the chromosome arm(s) (downsampled to the appropriate CN) from which the path segments came were combined and *de* novo assembled into OM contigs with Bionano Assembler. The resulting contigs from each amplicon simulation were combined with the previously simulated reference contigs and used as input to AR. For combination sets of three amplicons from the same sample, a similar downsampling and combination strategy was used, where molecules from each of the three amplicon simulations was separately downsampled based on the copy number settings of the mixture then combined. As heterogeneous combinations of amplicons may occur at different ratios, we selected three sets of copy numbers for this combination simulation cases: 20-20-20, 20-15-10, and 20-2-2.

In the simulation of the 20 *de novo* circular amplicons, a simulated tumor reference was generated from hg19 using SCNVSim (version 1.3.1) and simulated amplicon structures were generated using ecSimulator (version 1.0, https://github.com/jluebeck/ecSimulator). OM molecules were generated at baseline 40x coverage and amplicon copy number of 20. The human papillomavirus-16 integration example was performed at the same coverage and copy numbers as the other simulated amplicons.

1.4.14 Measuring AR simulation performance

We computed the longest common substring (LCS) between the AR paths and the ground-truth path and considered only the path having the LCS between AR and AA paths when computing precision and recall. We define the LCS here using the identities of the breakpoint graph segments and their orientations. We pre-filtered some possible assembly error reflected in the paths by removing ends of reconstructed paths which were trivial reconstructions of the reference genome and which were not supported by the AA path. To measure the accuracy of AR-reconstructed paths against the ground truth simulated paths, we developed a set of three measurements which were used in calculating performance and recall.

1) Length (bp): Reports the length of a breakpoint graph path in base pair units.
2) Nsegs: Reports the length of a breakpoint graph path in terms of the number of graph segments (unbiased towards genomic length)
3) Breakpoint: Reports the length of a breakpoint graph in terms of the number of breakpoint graph segment junctions in the path.

We define precision and recall as follows, where *M* is the path measurement function (Length (bp), Nsegs, or Breakpoint), *LCS* is the longest common substring function, P_{AA} is the sequence of segments in the AA path, and P_{AR} is the sequence of segments in the reconstructed AR path:

$$\text{Precision: } \frac{M\big(\mathrm{LCS}(P_{\mathrm{AA}}, P_{\mathrm{AR}})\big)}{M(P_{\mathrm{AR}})}$$

$$\text{Recall: } \frac{M\big(\mathrm{LCS}(P_{\mathrm{AA}}, P_{\mathrm{AR}})\big)}{M(P_{\mathrm{AA}})}$$

To summarize the precision and recall metrics in a single value, we computed a mean F1 score across all the simulated amplicons for a given set of simulation conditions as

$$\text{mean F1} = \frac{\sum_i \left(2\ \frac{\text{precision}_i\ *\ \text{recall}_i}{\text{precision}_i + \text{recall}_i}\right)}{n}$$

1.4.15 Reconstructed path visualizations

We developed a visualization utility, CycleViz (https://github.com/jluebeck/CycleViz), which produces circular and linear visualizations of AR or AA reconstructed amplicons (Fig. 1.3a,b), to create topologically correct visualizations of AR reconstructions. CycleViz accepts inputs including the path files reported by AR (in the AA-cycles format) as well as the path OM alignment files (optional) and produces visualizations which show the reconstructed path, *in silico* digestion of the path segments and the alignments of the digested segments with assembled OM contigs. For circular and linear visualizations, CycleViz places path segments in the visualization based on the length of the segments and their position in the path. For circular visualization layouts, the relative positions are converted to polar coordinates and a circular layout is formed. We also developed a web-based visualization utility, ScaffoldGraphViewer, for visualizing JSON-encoded scaffold graphs generated by AR using CytoscapeJS (64) (Fig. 1.3c). The ScaffoldGraphViewer web utility can be accessed at https://jluebeck.github.io/ScaffoldGraphViewer/.

1.5 Acknowledgements

Chapter 1, in full, is a reprint of the material “AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications” by Jens Luebeck, Ceyda Coruh, Siavash R. Dehkordi, Joshua T. Lange, Kristen M. Turner, Viraj Deshpande, Dave A. Pai, Chao Zhang, Utkrisht Rajkumar, Julie A. Law, Paul S. Mischel & Vineet Bafna as

it appears in Nature Communications 2020. The dissertation author was the primary investigator and author of this paper. The authors thank members of the Bafna and Mischel labs for their input, Dr. Marcy Erb (UCSD SOM Microscopy Core) for advice on FISH experiments, the Ecker Lab at the Salk Institute for use of the Bionano Irys optical mapping instrument, and Bionano Genomics, Inc., (Alex Hastie, Ernest Lam, Andy Pang, Jian Wang, Heather Mashhoodi and others) for supplying data and providing input on this project.

1.6 Appendix

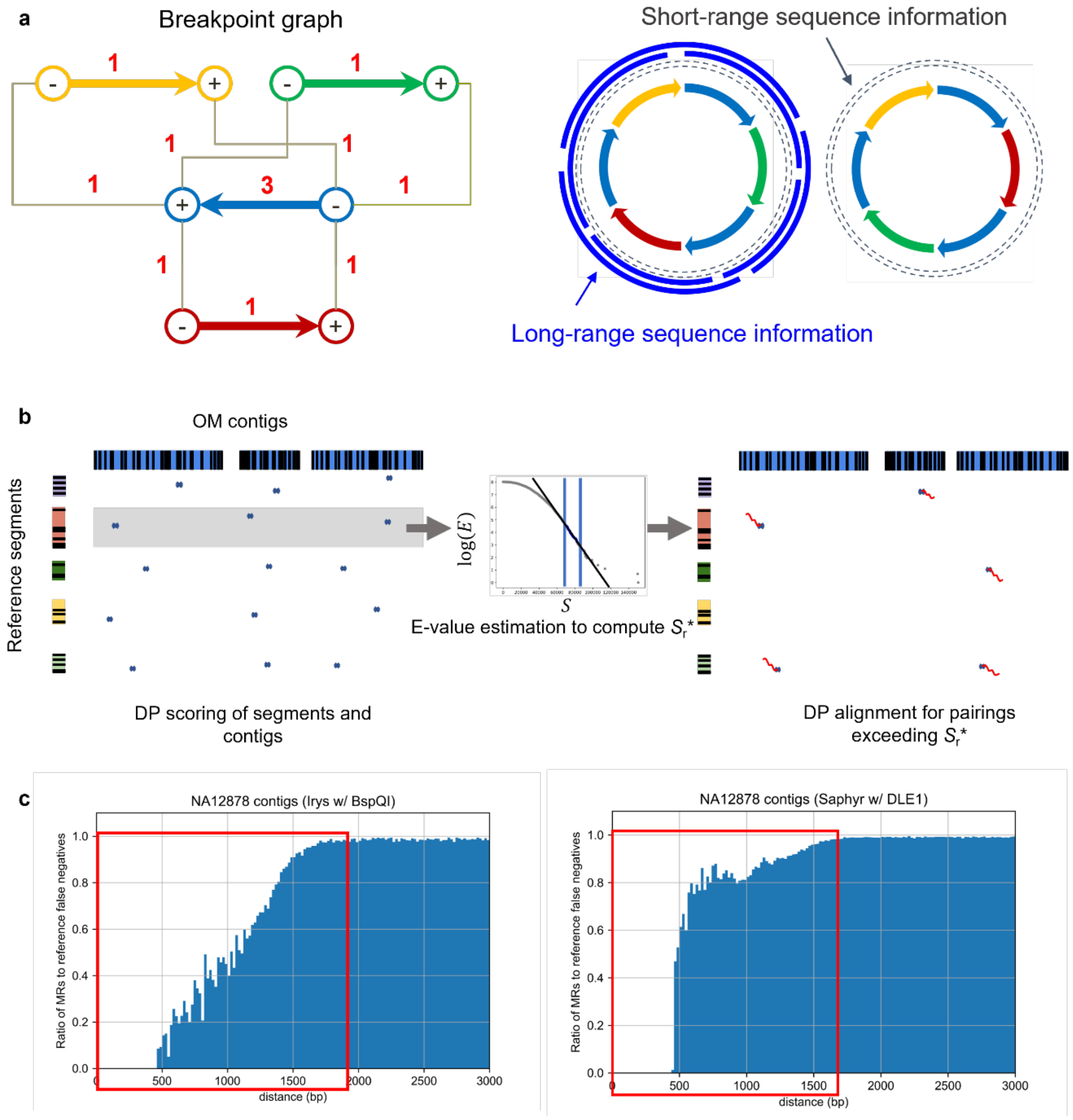
Table 1.1: Discovered integration points and OM supports.

Sample	Location1 chromosome	Location1	Location2 chromosome	Location2	Support (# mols)
H460	chr12	7660219.28	chr8	129414270.10	18
K562	chr1	142891183.20	chr13	81123336.19	326
K562	chr1	142888366.90	chr13	93260823.16	50
CAKI2	chr12	88302814.50	chr6	168382883.20	25

Table 1.2: AR default parameters.

Minimum unaligned contig labels for reference detection	20
Maximum unaligned contig labels for reference detection	500
Minimum unaligned contig size for reference detection	200000
Maximum unaligned contig size for reference detection	5000000
Maximum path imputation gap size	400000
DFS max search depth	64
Maximum candidate paths	1024
Maximum copies of BPG edge in candidate path	20
maximum linked scaffold graph paths to keep	500
Maximum distance of alignment to contig end to be classified as alignment end (base-pairs)	100000
Maximum distance of alignment to contig end to be classified as alignment end (labels)	12

Figure 1.1: Motivation for AR and the SegAligner methods. **a**, Long-range sequencing enables accurate disambiguation of large duplications. A breakpoint graph is represented on the left, with colored arrows representing oriented genomic segments, and with edges connecting the segments between source (+) and destination (-) breakpoints. The reconstructions on the right demonstrate that a graph containing duplicated segments can produce multiple possible reconstructions. Short reads provide information about individual junctions in the graph, but long-range sequence information can span and disambiguate multiple junctions. **b**, The SegAligner method for computing statistically significant scoring alignments by estimating parameters in an E-value model from alignments of segments against all contigs present in a sample. An all vs. all representation of best alignment scores of genomic reference segments and OM contigs is shown on the left. The collection of scores of each reference segment are used to build a scoring distribution for the E-value model. In the diagram on the right, alignments are generated from the segment-contig pairings with significantly high-scoring alignment scores. **c**, Ratio of matching regions (MRs) to false-negative reference labels in Bionano NA12878 data indicates the distance-dependent rate of label collapse for consecutive labels in the hg19 reference genome. Two distinct patterns of label collapse exist between Bionano Irys (left) and Bionano Saphyr (right) platforms. Red boxes indicate the regions in the respective technologies that are considered by SegAligner to have non-zero probability of label collapse.


a
Breakpoint graph
1
1
1
1
1
3
1
1
1
1
Short-range sequence information
Long-range sequence information
b
OM contigs
Reference segments
log(E)
S
E-value estimation to compute S_r*
DP scoring of segments and contigs
DP alignment for pairings exceeding S_r*
c
NA12878 contigs (Irys w/ BspQI)
Ratio of MRs to reference false negatives
distance (bp)
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
3000
NA12878 contigs (Saphyr w/ DLE1)
Ratio of MRs to reference false negatives
distance (bp)
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
3000

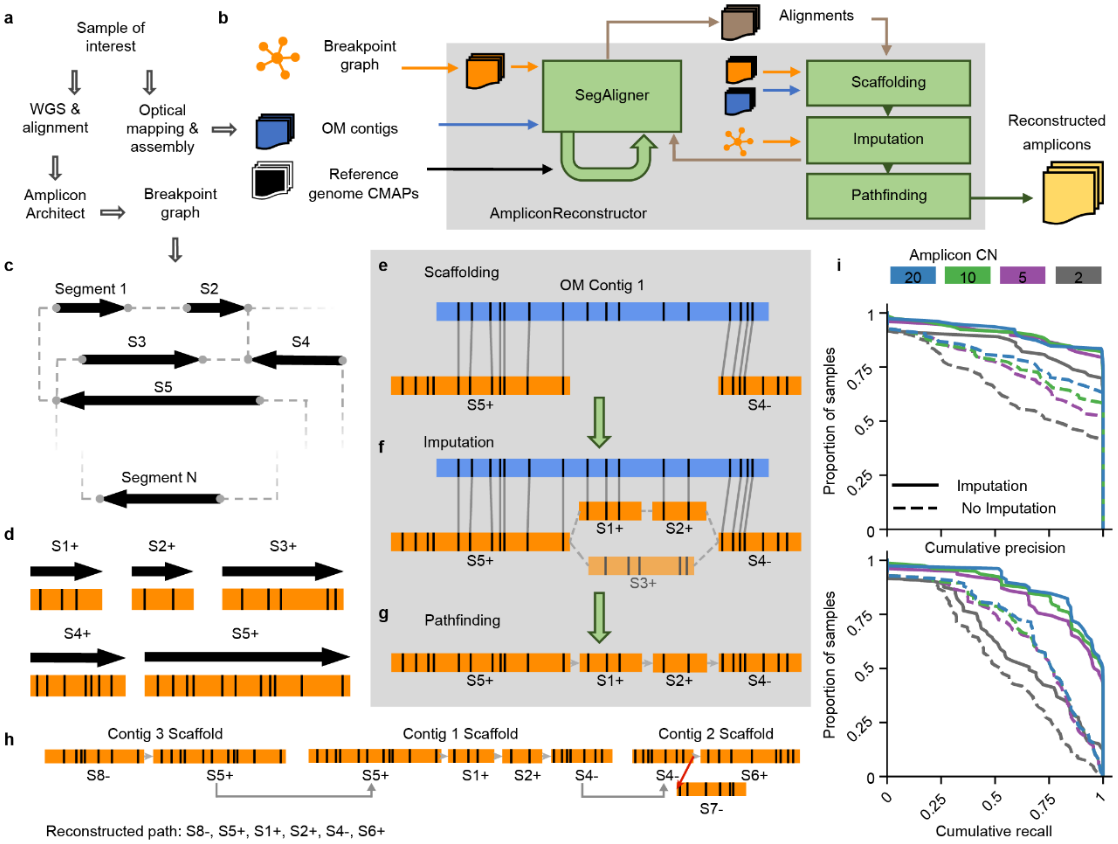


Figure 1.2: AmpliconReconstructor (AR) overview**.** **a**, Workflow to produce the necessary inputs for AR. AR accepts OM data in the consensus map (CMAP) format. **b**, High-level overview of the AR method, where the inputs and outputs are shown outside the grey box representing the AR wrapper. The green loop-back arrow on the SegAligner module represents the identification of reference segments not encoded in the breakpoint graph. **c**, A breakpoint graph with *N* segments. **d**, *In silico* digestion of breakpoint graph segments (orientation given by +/-) from **c** to produce graph OM segments. **e**, Alignment of graph OM segments to OM contigs produces a scaffold of segment-contig alignments. **f**, AR uses the structure of the breakpoint graph to identify paths between scaffold alignment endpoints which are also paths in the breakpoint graph. AR generates composite optical maps from combined path segments to score each candidate path against the gap in the scaffold. **g**, AR identifies a candidate path with maximum score out of the possible imputed paths between two alignments. **h**, AR links individual scaffolds sharing overlap between graph segments. The resulting graph has two types of edges, allowed (grey) and forbidden (red). **i**, Cumulative precision and recall curves based on simulated OM data for AR using SegAligner, calculated with the Length (bp) LCS metric. Line color indicates the copy number (CN) of the simulated amplicon.

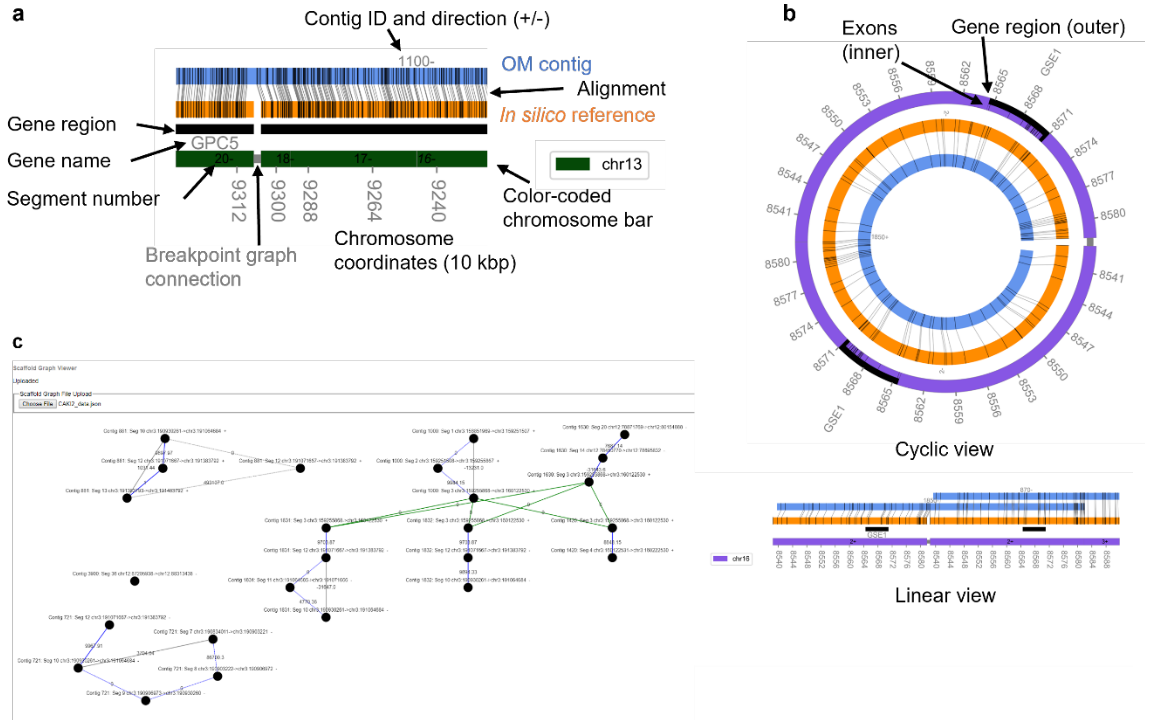
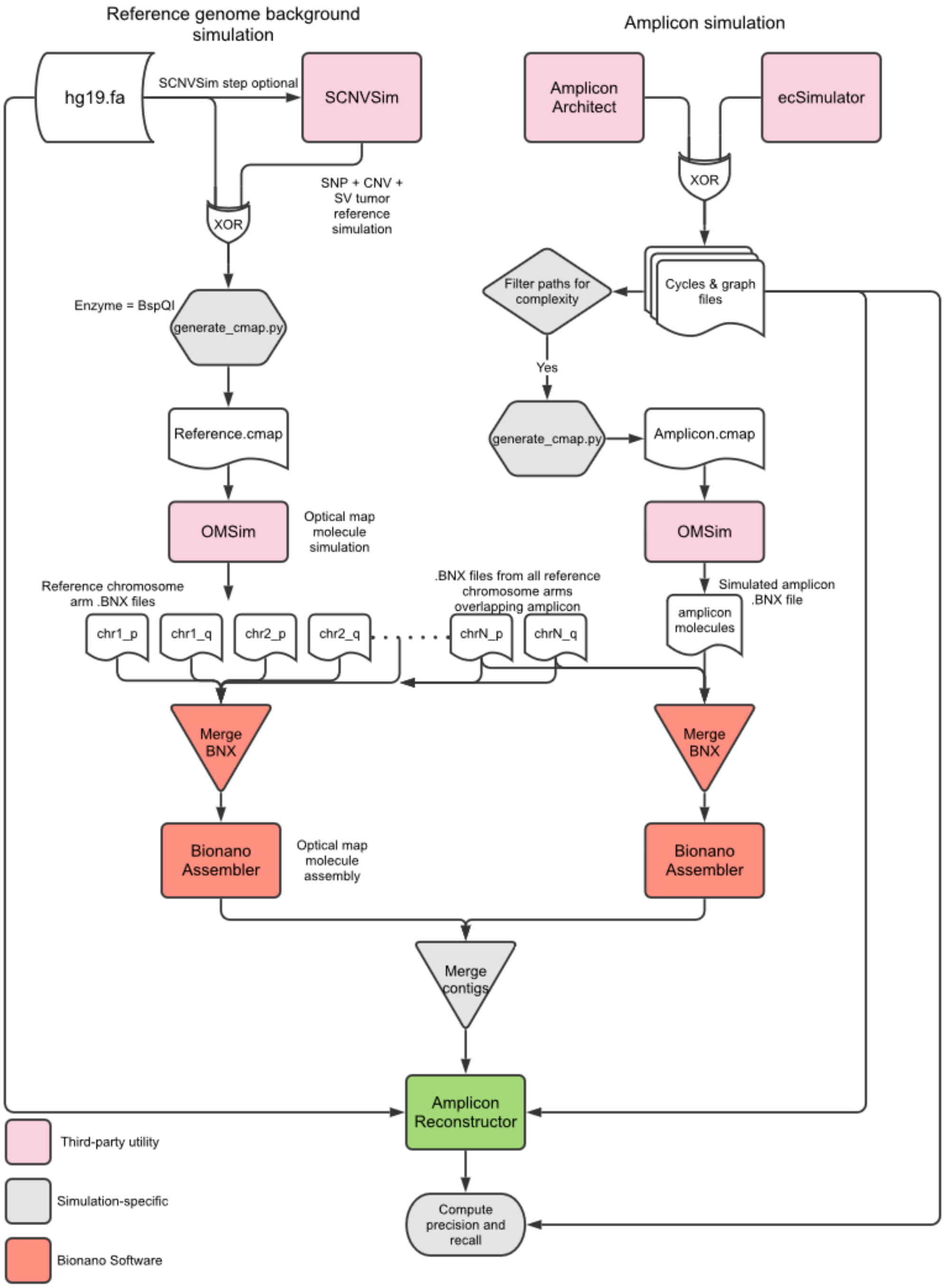


Figure 1.3: Visualization of AR results. **a,** Diagram of features in CycleViz visualization and associated feature tracks. **b**, CycleViz can generate both cyclic (top) and linear (bottom) visualizations. **c**, ScaffoldGraphViewer visualization of AR scaffold graphs using CytoscapeJS. Grey edges indicate connections in the scaffold directed acyclic graph (DAG). Dotted edges indicate imputation. Blue edges indicate edges used in the scaffold's heaviest path. Green edges indicate a link between two scaffolds.

Figure 1.4: Process diagram and control flow for the AR simulation study. The simulations used two primary choices of reference genome, hg19 or hg19 simulated tumor (using SCNVSim). Two choices of amplicon sources are provided; those generated by AA on cancer cell lines (85 cases) or circular ecDNA simulated by ecSimulator (20 cases). Molecules are simulated from background reference and amplicon separately using OMSim. The resulting BNX files are merged such that the amplicon molecules also get the simulated reference molecules from all overlapping chromosome arms. The merged BNX files for amplicon and background reference are assembled independently. This way the background reference assembly only needs to be generated once for all amplicons used. The contigs from the background reference and amplicon are then merged. AR is run on the merged contigs, the breakpoint graph that was either simulated or selected from AA as well as the *in silico* reference genome. Pink filling indicates “third-party” utilities, which are not part of AR or the simulation pipeline. Grey filling denotes tools and scripts developed specifically for the simulation process. Salmon-colored filling indicates tools released by Bionano Genomics, Inc. (San Diego, CA).


Reference genome background simulation
Amplicon simulation
hg19.fa
SCNVSim step optional
SCNVSim
Amplicon Architect
ecSimulator
SNP + CNV + SV tumor reference simulation
XOR
XOR
Enzyme = BspQI
generate_cmap.py
Filter paths for complexity
Cycles & graph files
Yes
Reference.cmap
generate_cmap.py
Amplicon.cmap
OMSim
Optical map molecule simulation
OMSim
Reference chromosome arm .BNX files
.BNX files from all reference chromosome arms overlapping amplicon
Simulated amplicon .BNX file
amplicon molecules
chr1_p
chr1_q
chr2_p
chr2_q
chrN_p
chrN_q
Merge BNX
Merge BNX
Bionano Assembler
Optical map molecule assembly
Bionano Assembler
Merge contigs
Amplicon Reconstructor
Compute precision and recall
Third-party utility
Simulation-specific
Bionano Software

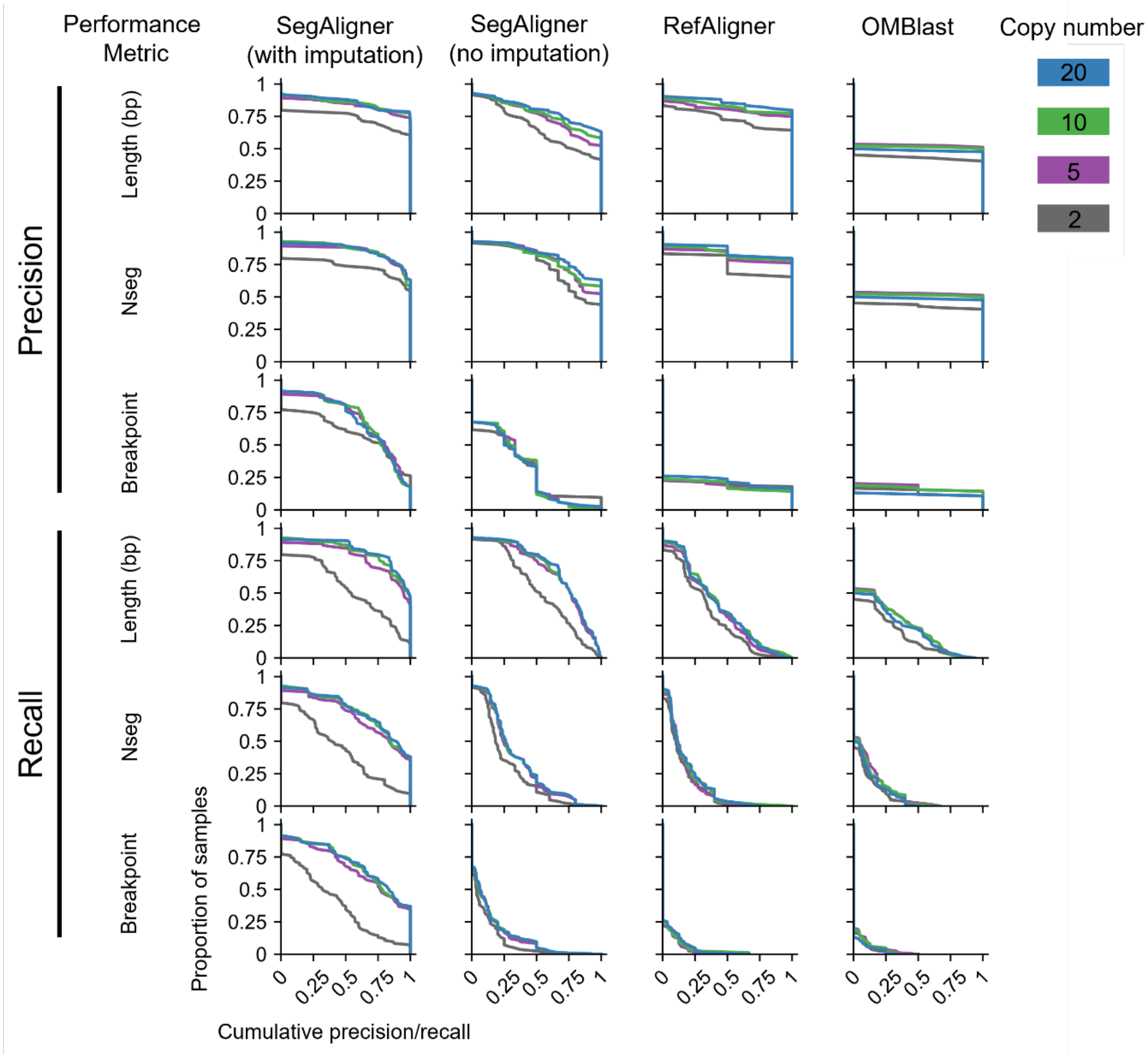
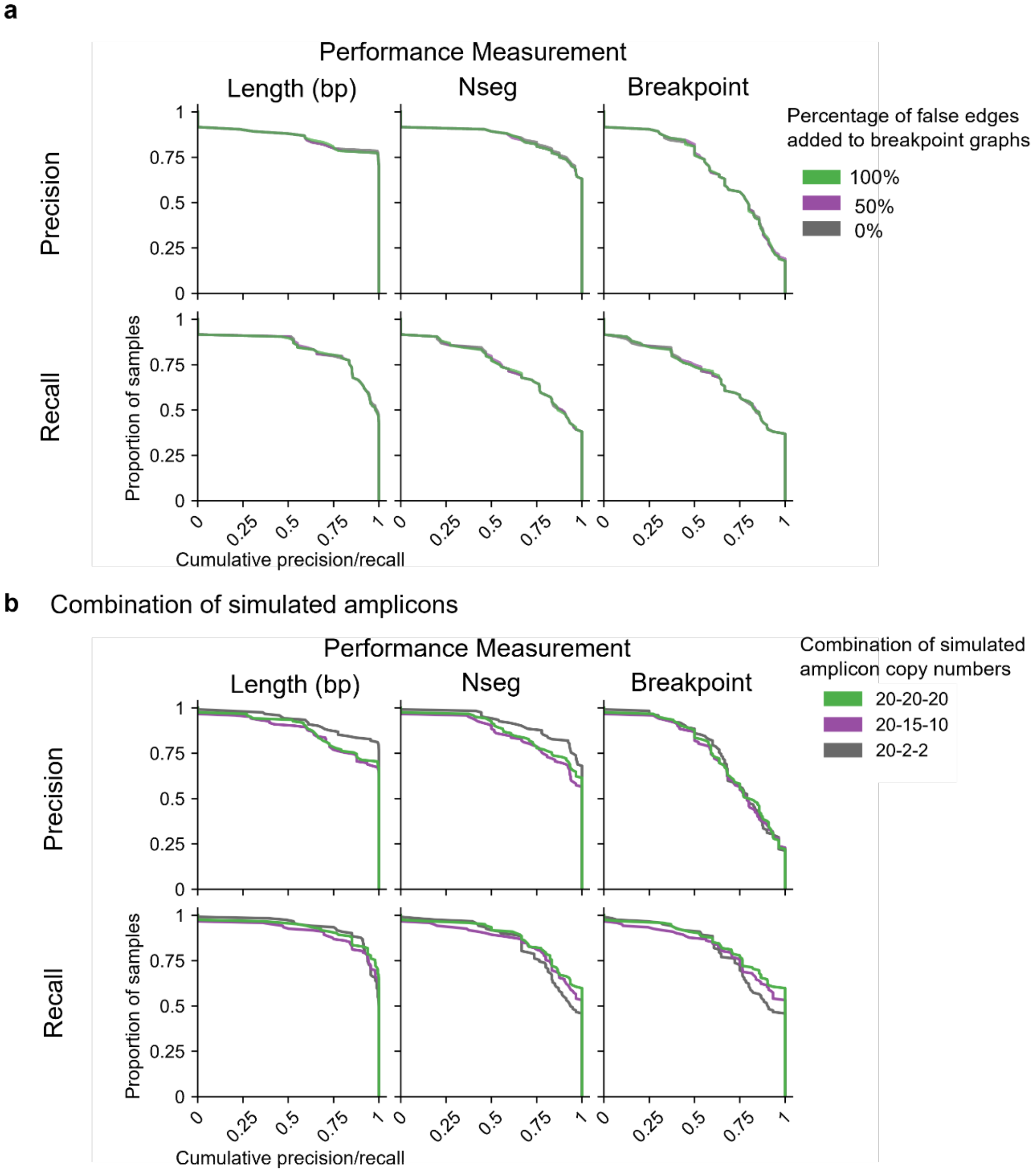


Figure 1.5: AR performance on simulated data using different OM alignment techniques. Cumulative precision and recall curves measuring performance of AR reconstructions on simulated amplicon OM data using different OM alignment methods and precision/recall measurement methods. OM assembly failures were not filtered from this figure.



b Combination of simulated amplicons

Figure 1.6: AR performance on simulated data with false graph edges and amplicon mixtures. **a**, Cumulative precision and recall curves measuring the performance of AR, using SegAligner for OM alignment, on simulated OM data with added false positive edges into simulated amplicon breakpoint graphs. OM assembly failures were not filtered from this panel. **b**, Cumulative precision and recall curves measuring the performance of AR using SegAligner for simulated heterogeneous mixtures of similar amplicons. OM assembly failures were not filtered from this panel.

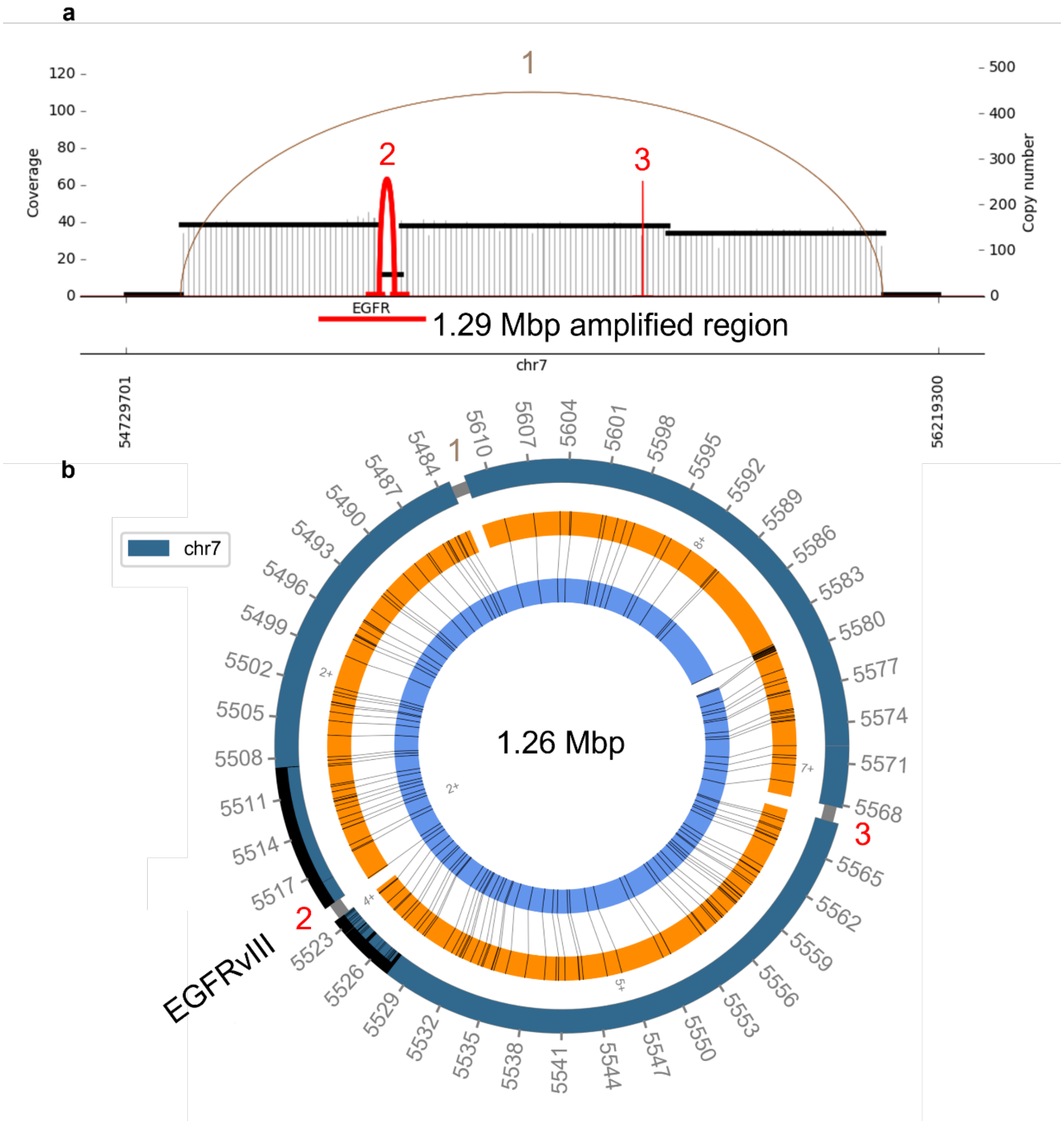
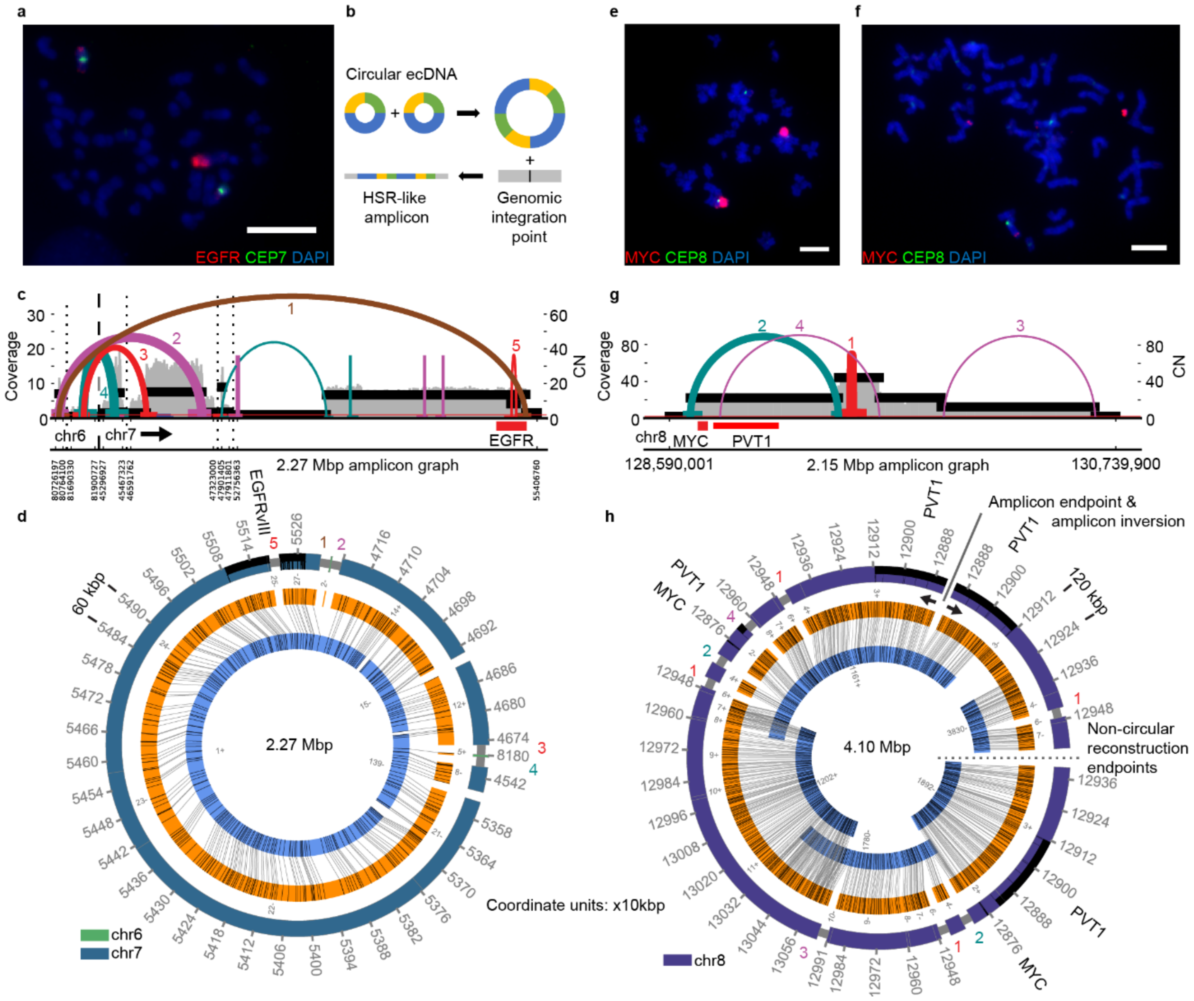


Figure 1.7: AR reconstructs a circular ecDNA in GBM39. **a**, AA detects a 1.29 Mbp amplified region of chr7 bearing *EGFR* and indicates three breakpoint edges. Both *EGFR* and *EGFRvIII* appear to be amplified. **b**, AR (with SegAligner) reconstructs a circular 1.26 Mbp ecDNA in GBM39 using a single Bionano Irys contig. A legend for reading CycleViz AR figures can be found in Figure 1.3a,b. Coordinate units on the labeled figure are scaled by 10 kbp. AR reconstructs a form of EGFR carrying the ~20 kbp vIII deletion, which is supported by the breakpoint graph.

Figure 1.8: Reconstruction of extrachromosomal DNA (ecDNA). **a**, FISH with DAPI (4′,6-diamidino-2-phenylindole)-stained metaphase chromosomes in HK301 showing an HSR-like amplicon containing *EGFR*. Scale bar indicates 10 µm. **b**, Theoretical model for the integration of circular extrachromosomal DNA into HSR-like amplicons, preserving the structure of breakpoint graph. **c**, AA-generated breakpoint graph for HK301 containing *EGFR* and segments from chr6. The coloring of the graph edges represents the orientation of the junction between the two segments. Edge thickness indicates AA-estimated breakpoint copy number. Vertical dashed lines separate segments from different chromosomes while dotted lines indicate distinct genomic regions from the same chromosome. Numbering of breakpoint edges corresponds with AR reconstruction breakpoint numbering. **d**, Cyclic AR reconstruction of HK301 amplicon containing *EGFRvIII*. Breakpoint graph edges supported by the AA graph are numbered in a manner corresponding to the numbering in panel **c**. **e**, FISH with DAPI-stained metaphase chromosomes in NCI-H460 shows HSR-like *MYC* amplicon. Scale bar indicates 7.3 µm. **f**, FISH with DAPI-stained metaphase chromosomes in NCI-H460 showing an extrachromosomal *MYC* amplicon. Scale bar indicates 7.3 µm. **g**, AA-generated breakpoint graph for NCI-H460 containing *MYC* and *PVT1*. **h**, AR reconstruction of the NCI-H460 amplicon. Indicated in this figure is an amplicon inversion point (top right) where the reconstruction explaining the full amplicon ends, and then the structure begins to repeat in the opposite direction (solid line & opposing black arrows). Also indicated is an endpoint for the non-circular reconstruction (center right) where the AR reconstruction and full amplicon structure both stop (dotted line).


a
EGFR CEP7 DAPI
b
Circular ecDNA
HSR-like amplicon
Genomic integration point
e
MYC CEP8 DAPI
f
MYC CEP8 DAPI
c
Coverage
CN
1
2
3
4
5
chr6
chr7
EGFR
2.27 Mbp amplicon graph
80726197
80784100
81690330
81900727
45296927
45467323
46591762
47323000
47901405
47911801
52756363
55406760
g
Coverage
CN
1
2
3
4
chr8
MYC
PVT1
128,590,001
2.15 Mbp amplicon graph
130,739,900
d
EGFRvIII
60 kbp
2.27 Mbp
Coordinate units: x10kbp
chr6
chr7
h
Amplicon endpoint & amplicon inversion
120 kbp
4.10 Mbp
Non-circular reconstruction endpoints
PVT1
MYC
chr8

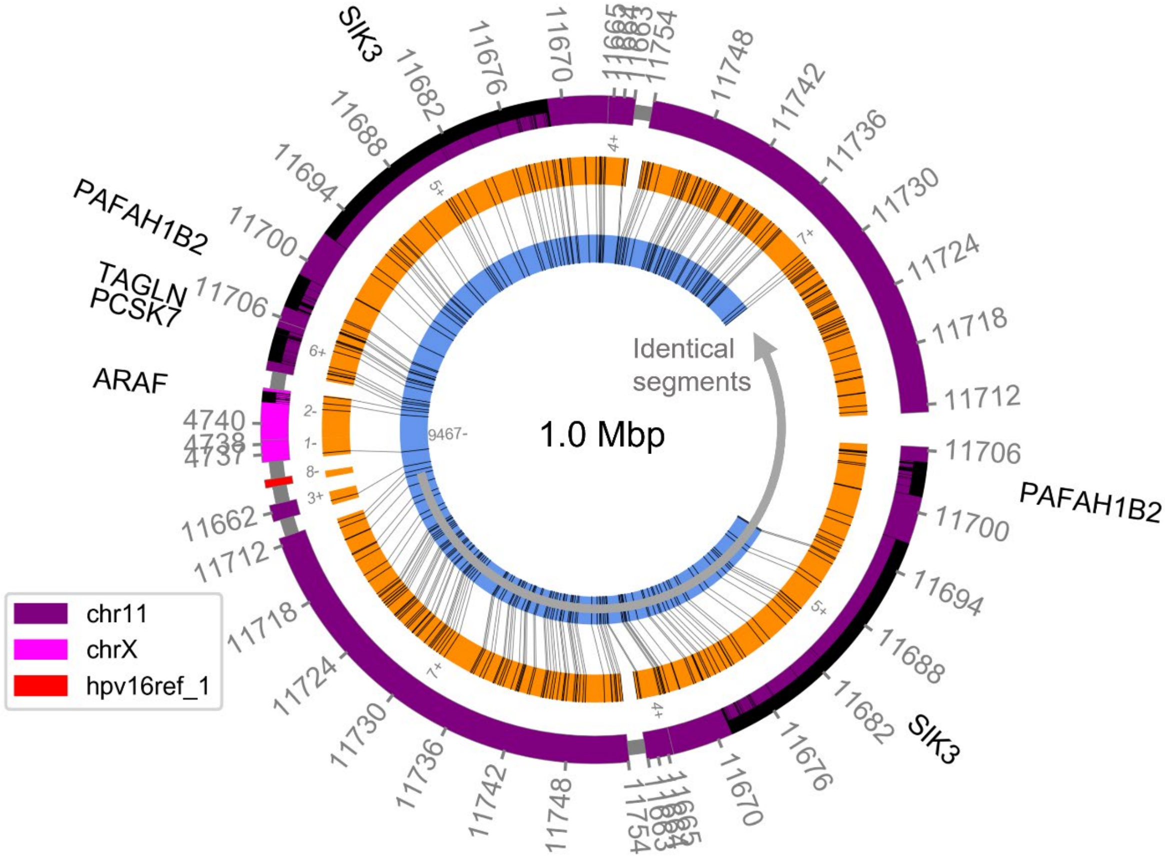


Figure 1.9: AR reconstructs a simulated human-viral amplicon. AR reconstruction of a simulated 1.0 Mbp circular ecDNA containing an integrated human papillomavirus-16 (HPV16). A legend for reading CycleViz AR figures can be found in Figure 1.3. Coordinate units on the labeled figure are scaled by 10 kbp. The grey arced arrow indicates identical overlap between segments assembled at the end of the structure which overlap to form the circular ecDNA structure.

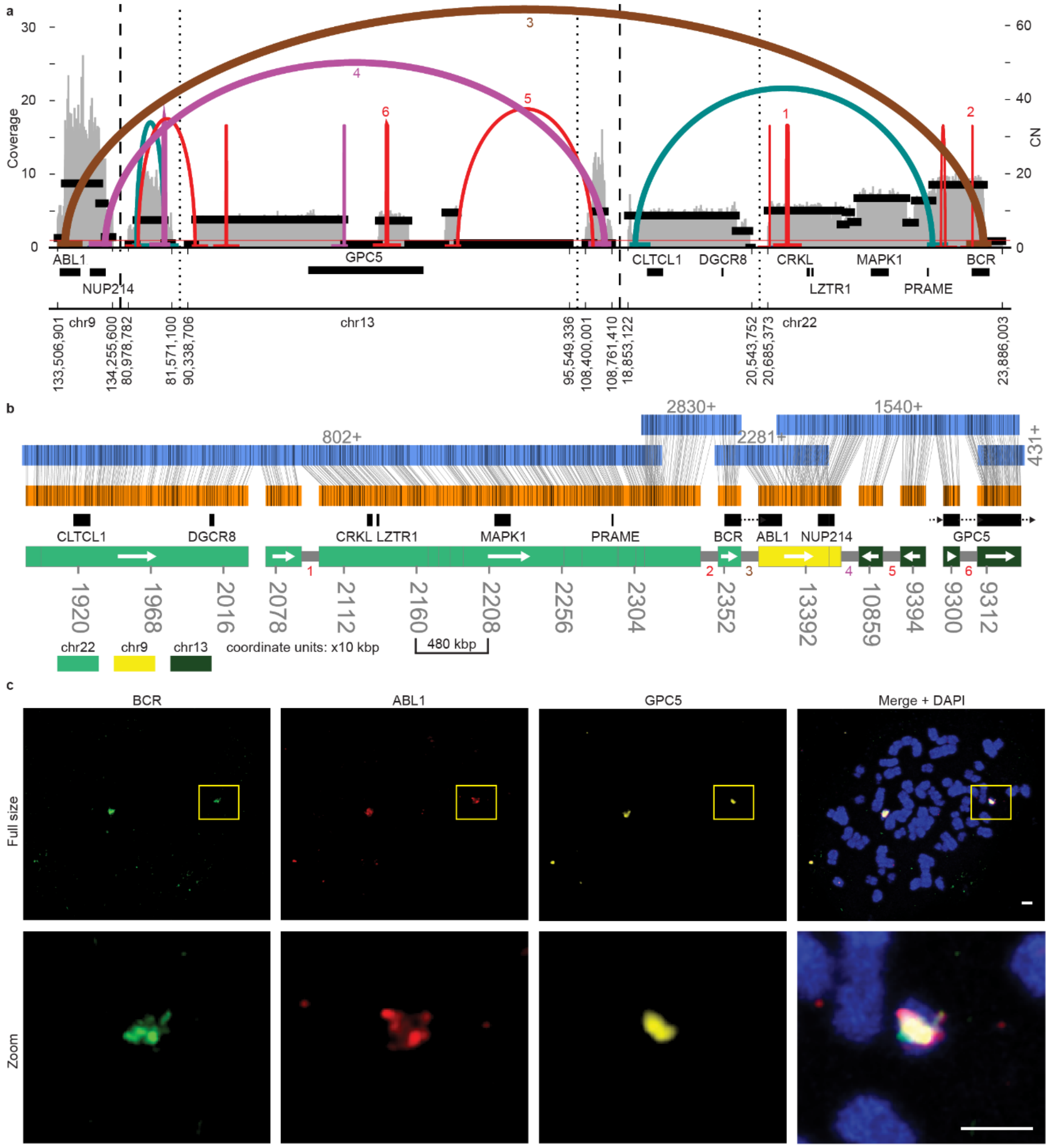


Figure 1.10: Reconstruction of a complex Philadelphia chromosome. **a**, AA-generated breakpoint graph for K562. Estimated copy number (CN), coverage, discordant reads forming breakpoint graph edges, and a subset of the genes in these regions are shown. **b**, AR reconstruction of an 8.5 Mbp focal amplification which was supported by both Irys and Saphyr reconstructions. The tracks from top to bottom are: OM contigs (with contig ID and direction indicated above), graph segments (alignments shown with vertical grey lines), gene subset and color-coded reference genome bar with genomic coordinates (scaled as 10 kbp units). Grey half-height bars between individual segments on the reference genome bar indicate support from edges in the AA breakpoint graph. White arrows inside the chromosome color bar indicate direction of genomic segment(s). Colored numbers correspond to numbered breakpoint graph edges in panel **a**. **c**, Multi-FISH using probes against *BCR*, *ABL1* and *GPC5* with DAPI-stained metaphase chromosomes. Scale bars indicate 2 μm in both “Full size” and “Zoom” rows.

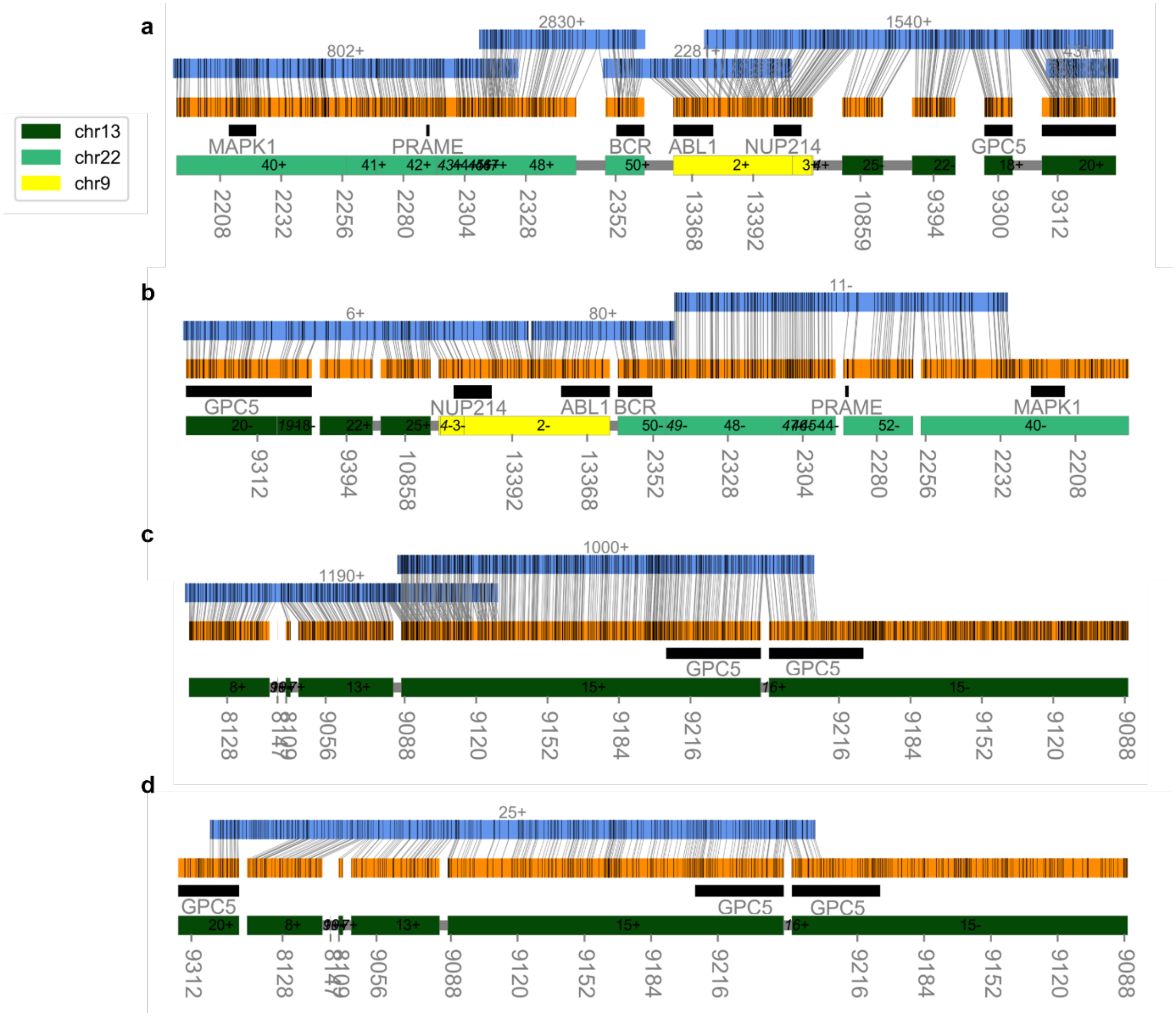


Figure 1.11: AR reconstructions of K562 amplicon with Bionano Saphyr and Bionano Irys data. **a**, AR reconstruction of *BCR-ABL1* amplicon in K562 using Bionano Saphyr data. A legend for reading CycleViz AR figures can be found in Figure 1.3a,b. Coordinate units on the labeled figure are scaled by 10 kbp. **b,** AR reconstruction of BCR-ABL1 amplicon in K562 using Bionano Irys data. Amplicon was reconstructed in the reverse direction from panel **a** and includes a graph segment (19) which was not included in the Saphyr reconstruction. **c,** Additional AR reconstruction (using Bionano Saphyr data) from AA-detected amplified segments outside the reconstructions in panels **a** and **b**. **d**, Additional AR reconstruction (using Bionano Irys data) from AA-detected amplified segments outside the reconstructions in **a** and **b**, with the exception of the leftmost segment, 20, which appears at the leftmost end of the Irys reconstruction in panel **b.**

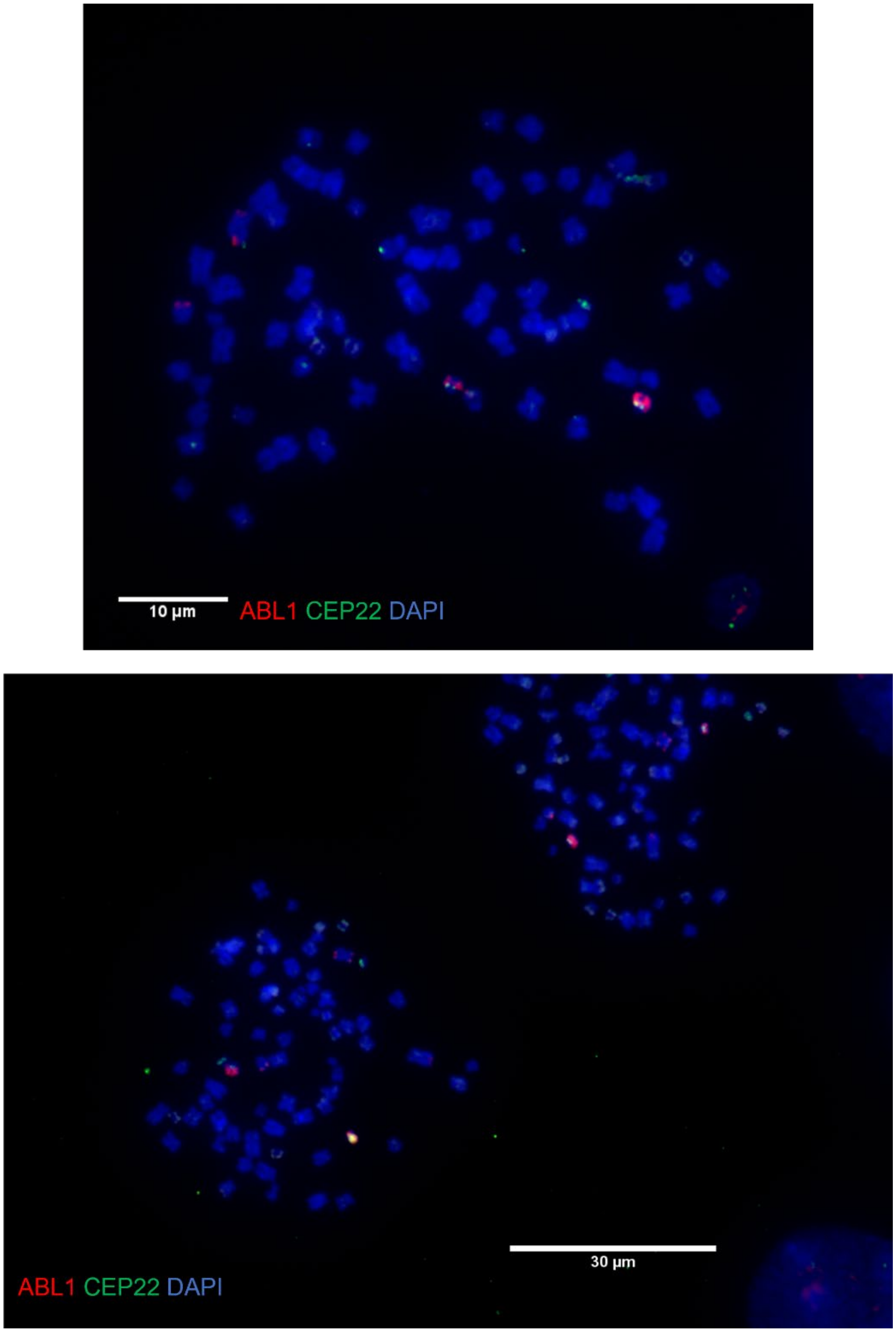


Figure 1.12: Multi-FISH for K562. Multi-FISH using confocal microscopy on DAPI-stained metaphase chromosomes in K562 cells. Centromeric repeat probe for chr22 (CEP22) and *ABL1* are stained in green and red, respectively. Two separate plate regions are shown.

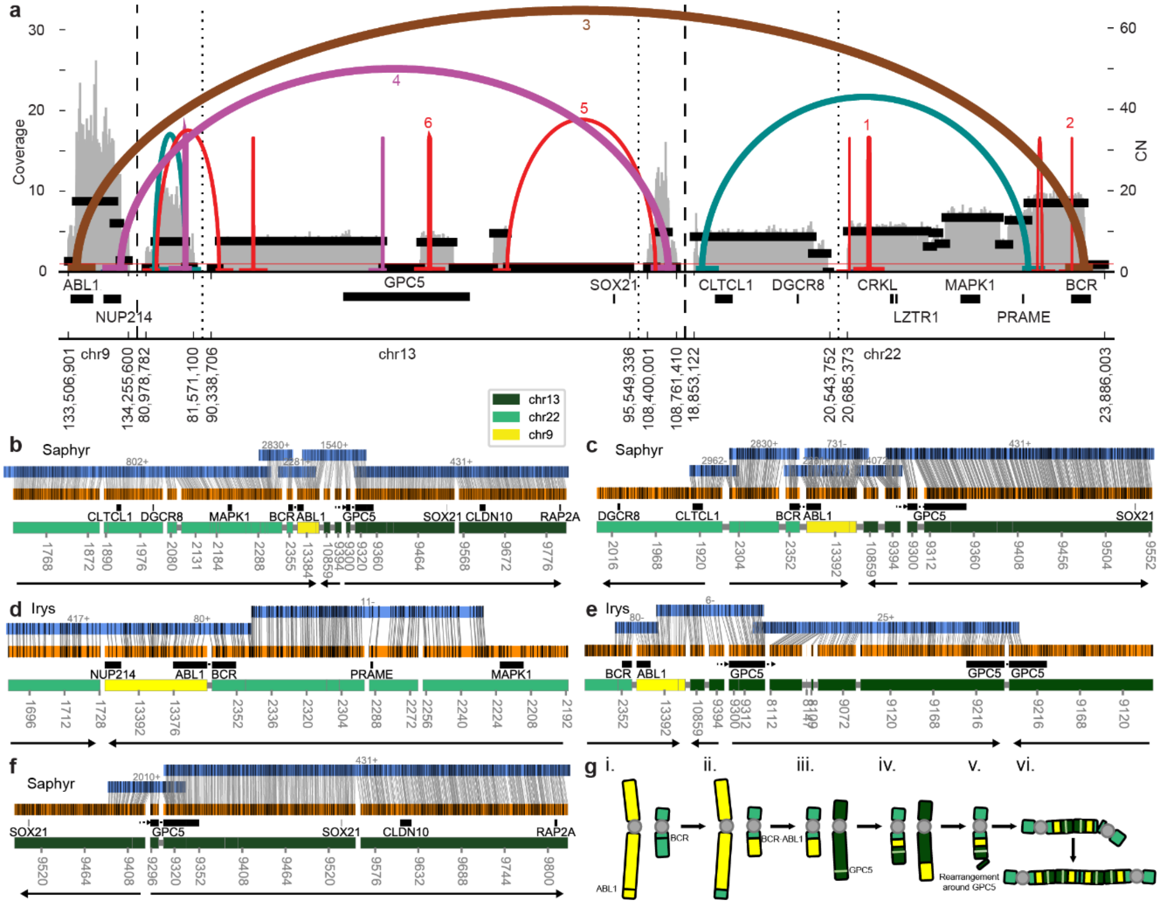


Figure 1.13: Annotated AR reconstructions of K562 amplicon. **a**, Reproduction of Fig. 3a with *SOX21* additionally labeled. **b**, Extended AR reconstruction of K562 *BCR-ABL1* amplicon using Saphyr data. **c**, AR reconstruction with Saphyr data supporting an alternate orientation for the genomic segment containing *CLTCL* as opposed to panel **b**. **d**, AR reconstruction with Irys data containing *BCR-ABL1* flanked by chr22 on both sides. **e**, AR reconstruction with Irys data containing *BCR-ABL1* with an inverted repeat of the 3' end of *GPC5*. **f**, AR reconstruction with Saphyr data shows inverted repeat of genomic segment containing *SOX21*. **g**, Hypothetical model for *BCR-ABL1* amplification in K562, involving the reciprocal translocation of q arms of chromosomes 9 and 22 to form the *BCR-ABL1* fusion (i-ii), then translocation of q arm of chromosome 13 with the *BCR-ABL1* chromosome (iii-iv) with rearrangement of the translocated regions from chr13 on the *BCR-ABL1* chromosome (v). Finally, inverted repeats observed in AR scaffolds can be explained by amplifications of *BCR-ABL1* through a dicentric chromosome model of amplification (vi).

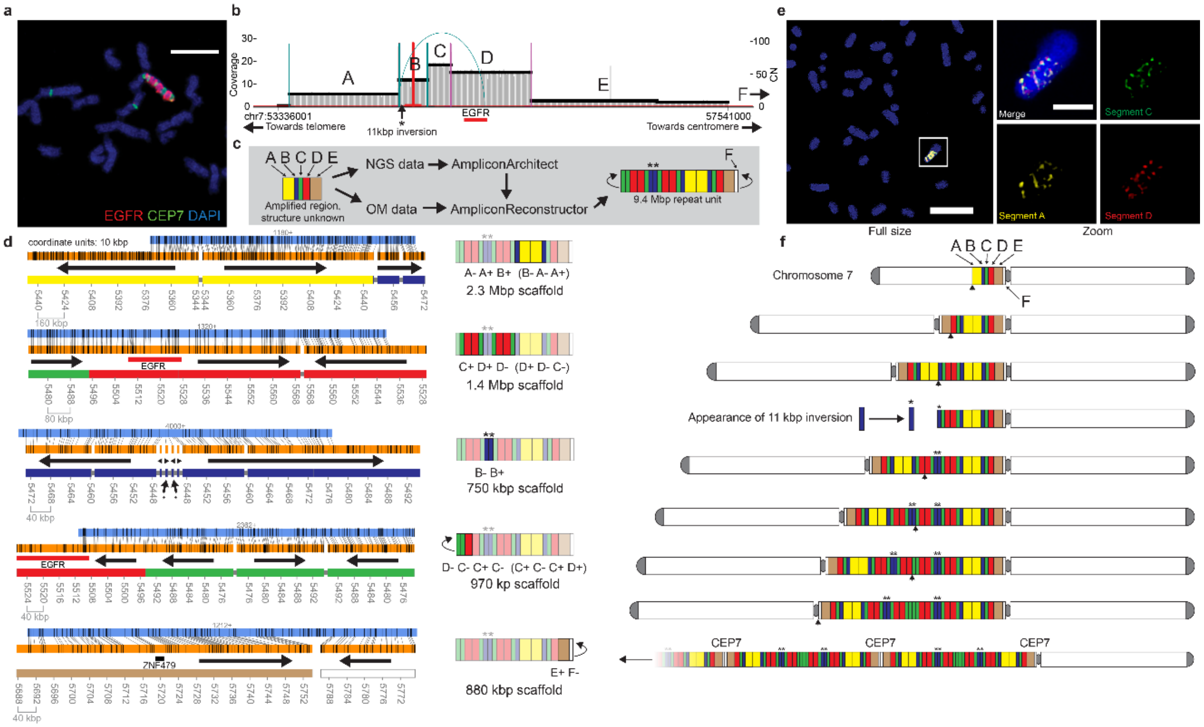
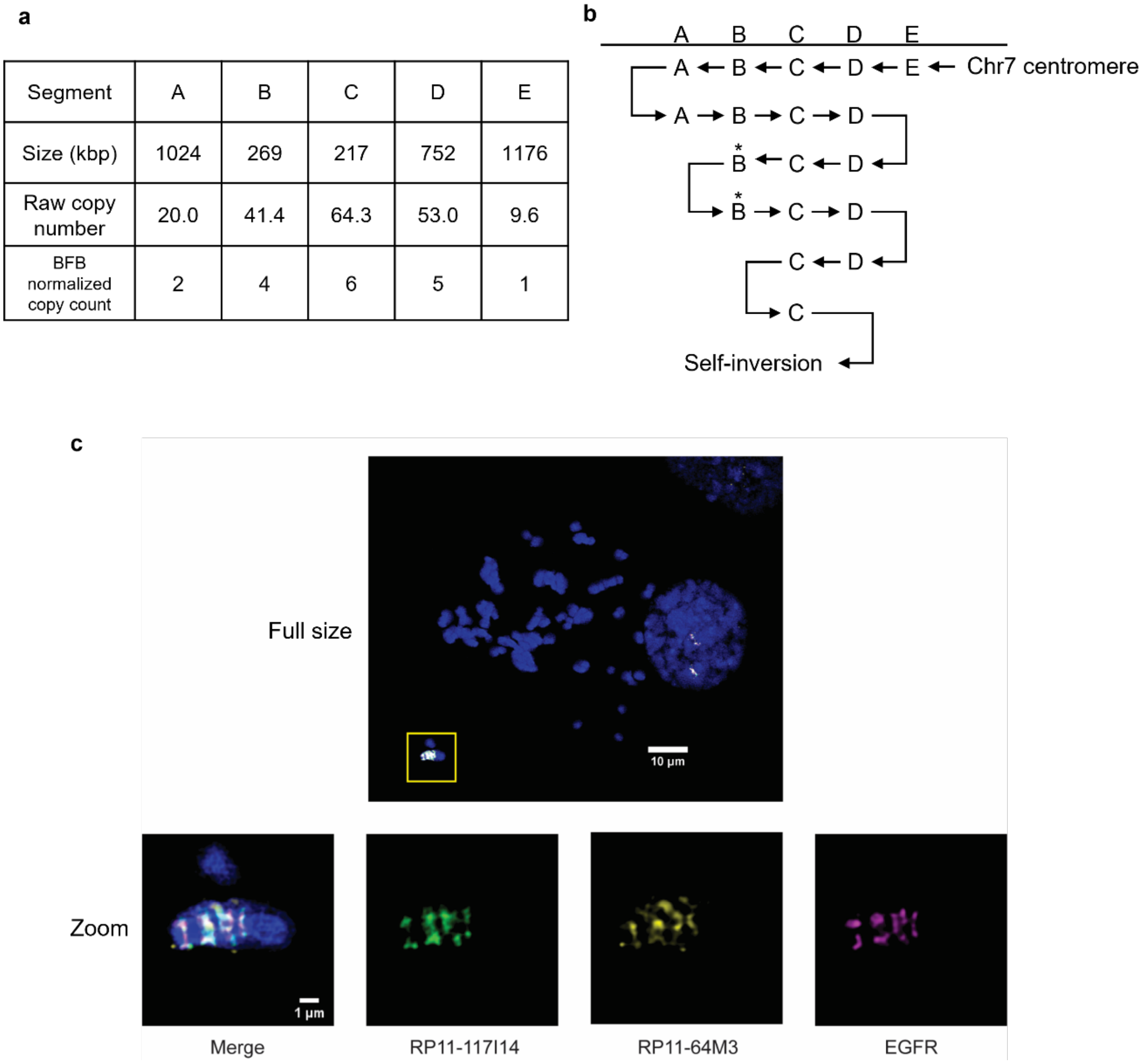


Figure 1.14: Reconstruction of Breakage-Fusion-Bridge. **a**, FISH confocal microscopy of DAPI-stained metaphase chromosomes in HCC827 showing multiple distinct bands of *EGFR* and CEP7 (chr7 centromeric repeat probe). Scale bar indicates 6µm. **b,** AA-generated breakpoint graph for amplified *EGFR* region in HCC827. Asterisk ('*') symbol indicates presence of 11 kbp inversion at 5' end of segment B. **c**, Workflow for analysis of amplified *EGFR* region in HCC827 to reveal BFB repeat unit structure. Amplified intervals detected by AA are labeled A-E and are colored yellow, blue, green, red and brown, respectively. Segment F indicates a region identified by AR but not AA. **d**, Visualization of the AR-generated scaffolds (left column) and cartoon illustration of reconstructed region(s) of the BFB (right column), including segment sequence. Black arrows in the scaffold column indicate segment directionality. **e**, Multi-FISH for BFB segments using super-resolution confocal microscopy on DAPI-stained metaphase chromosomes in HCC827. FISH probes used for segments A, C, and D were RP11-64M3, RP11-117I14, and *EGFR*, respectively. Scale for full size image indicates 11 µm. Scale bar for zoomed images indicates 3 µm. Brightness was decreased using ImageJ between full size and zoomed images. **f**, Theoretical model of formation for HCC827 *EGFR* BFB. Each row indicates a prefix inversion and duplication characteristic of BFB, alongside other SVs. Black arrowheads beneath the intermediate step in each row indicates the breakpoint of the BFB chromosome. The bottom row shows multiple duplications of the BFB unit along with a pericentromeric region of chromosome 7.



Segment	A	B	C	D	E
Size (kbp)	1024	269	217	752	1176
Raw copy number	20.0	41.4	64.3	53.0	9.6
BFB normalized copy count	2	4	6	5	1

Figure 1.15: BFB structure in HCC827. **a**, Listing of BFB amplicon segments showing size, raw copy number as estimated by AA, and normalized BFB count. **b**, A valid BFB string supported by both AA and AR, which contains interior structural variation (marked with an asterisk symbol, “*”) **c**, Multi-FISH using super-resolution confocal microscopy on DAPI-stained metaphase chromosomes in HCC827. Brightness was decreased using ImageJ between full size and zoomed images. Probe RP11-64M3 corresponds to segment A, RP11-117I14 corresponds to segment C, and *EGFR* corresponds to segment D.

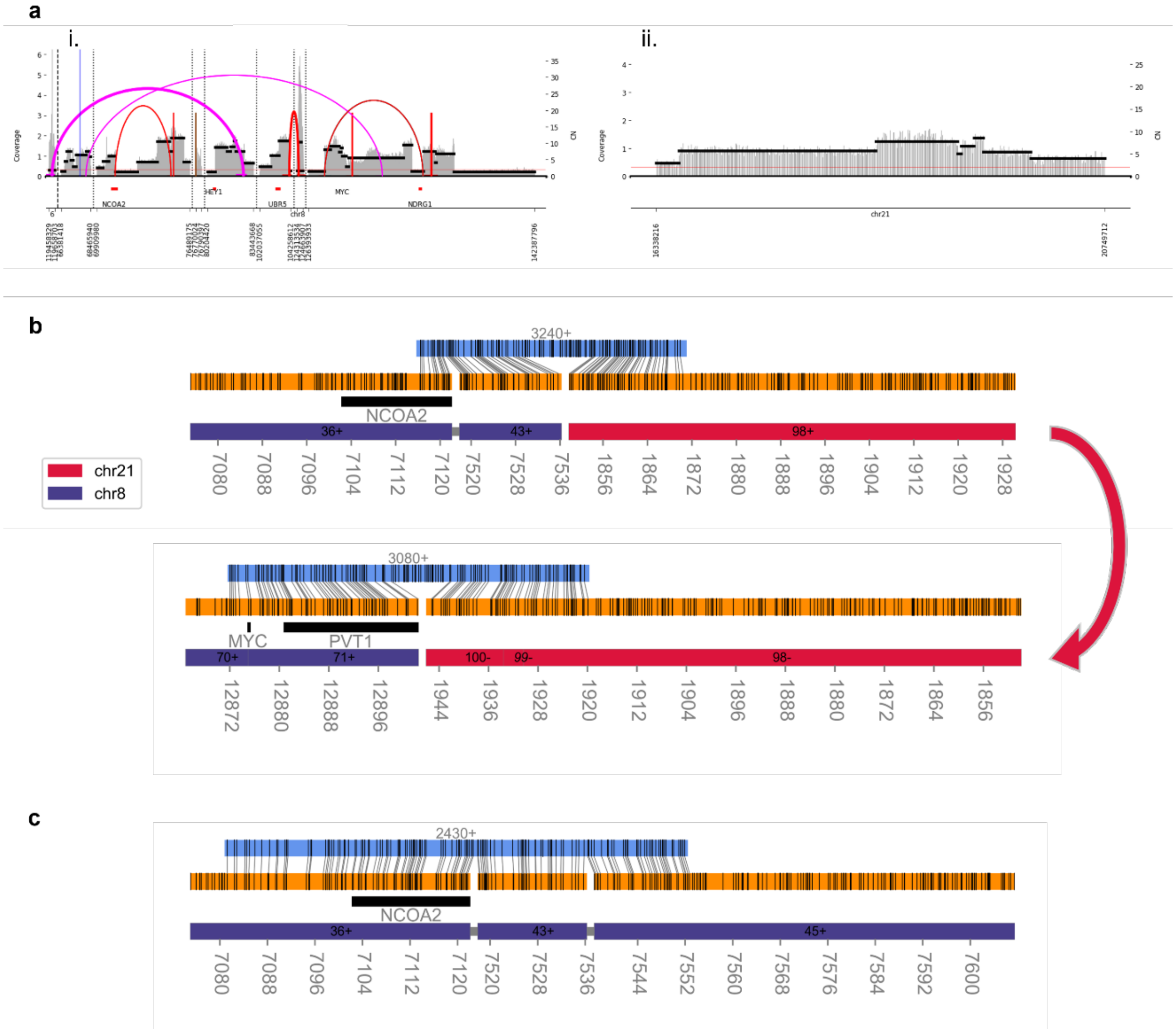


Figure 1.16: Heterogeneous AR scaffolds identified in HCC827. **a**, Two of five additional AA-generated breakpoint graphs for HCC827. The graph on the left (i) shows a complex structure of rearrangements, while the graph on the right (ii) shows distinct copy number changes without breakpoint graph edges. **b**, AR reconstruction of Bionano Saphyr scaffold joining segments from the graph labeled i in panel **a**, including *NCOA2* and *MYC*, to a segment of chr21 from the graph labeled ii in panel **a.** The scaffold on top in panel **b** can be joined to the scaffold on the bottom as the segment labeled 98 overlaps, as indicated with the curved red arrow. **c**, An alternate scaffold reconstructed by AR showing *NCOA2* joined to additional segments from chromosome 8.

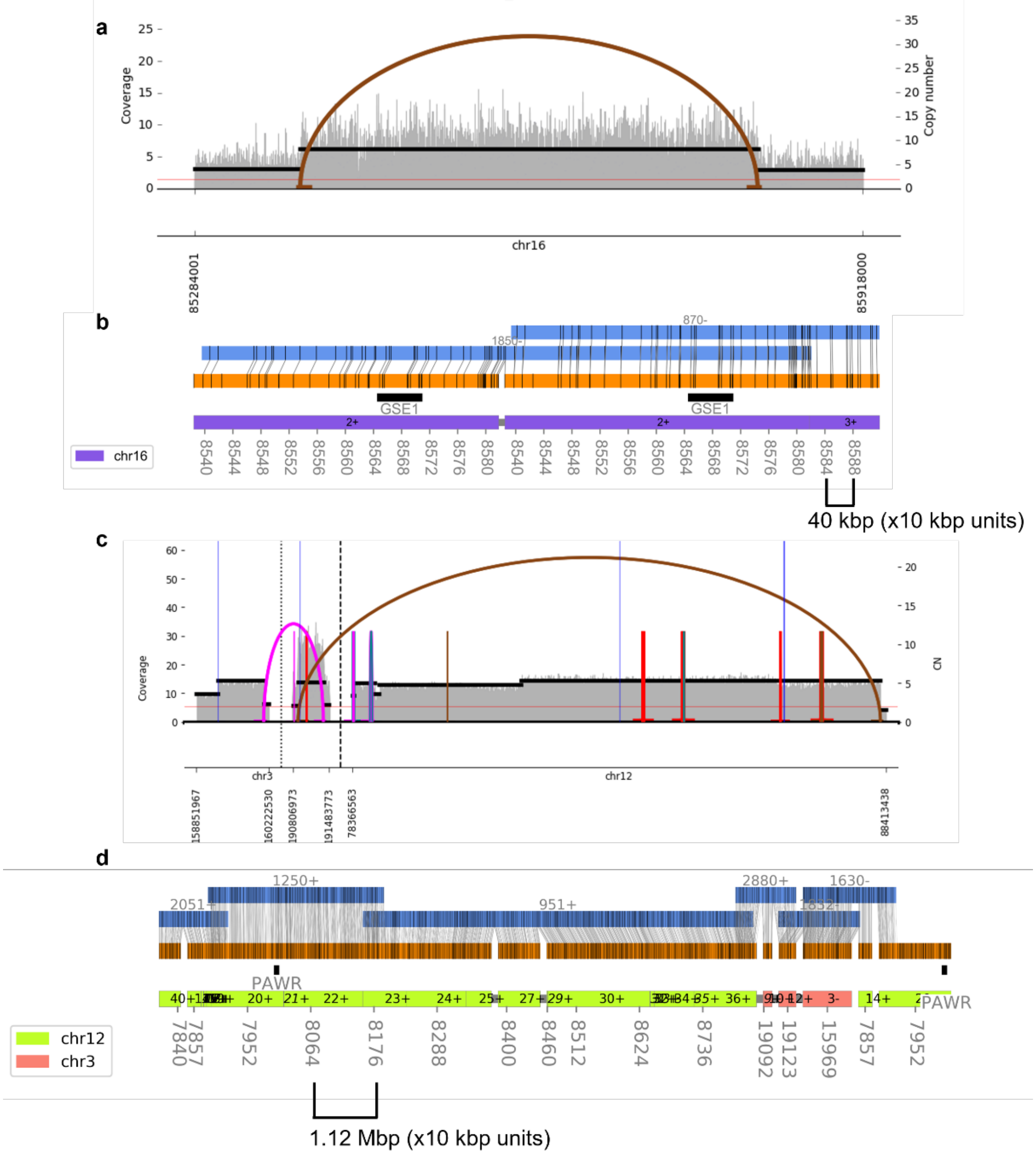
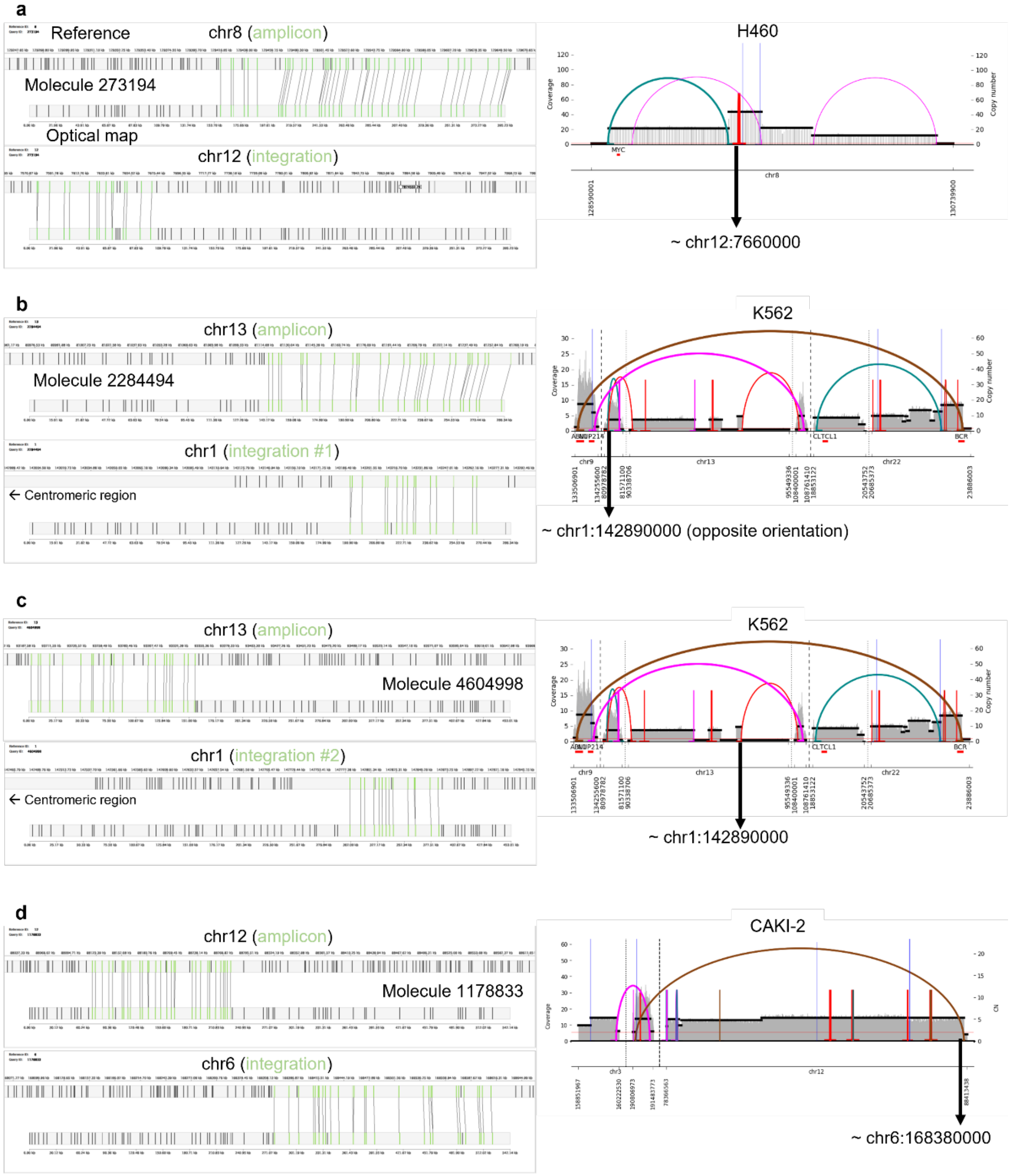


Figure 1.17: AR reconstructions in T47D and CAKI-2. **a**, In T47D, an AA-generated breakpoint graph visualization showing a 430 kbp amplified region with a single breakpoint edge connecting the ends of the amplicon. **b**, In T47D, AR reconstruction of the amplicon shows a 967 kbp region capturing the segmental tandem duplication involving *GSE1*. **c**, In CAKI-2, an AA-generated breakpoint graph visualization showing a complex amplified rearrangement joining segments (12.0 Mbp in total) from chr12 and chr3 in CAKI-2. **d**, In CAKI-2, AR-reconstructed 13.1 Mbp amplicon containing segments from chr12 and 3, including interior deletions and multiple copies of *PAWR*.

Figure 1.18: Integration points of focal amplifications. **a**, (Left) An example of a supporting alignment between an amplified region of chr8 (top) and a non-amplicon region on chr12 (bottom) in H460. (Right) The approximate integration point on the amplicon is indicated with a black arrow pointing at the approximate integration location. **b**, (Left) An example of a supporting alignment between an amplified region of chr13 (top) and a non-amplicon region on chr1 in K562 (bottom). An arrow pointing towards the centromere is also present due to the proximity of the integration point to the centromeric region (~ 80kbp). (Right) The approximate integration point on the amplicon is indicated with a black arrow pointing at the approximate integration location. **c**, (Left) An example of a supporting alignment between an amplified region of chr13 (top) and a non-amplicon region on chr1 in K562 (bottom). An arrow pointing towards the centromere is also present due to the proximity of the integration point to the centromeric region (~ 240kbp). (Right) The approximate integration point on the amplicon is indicated with a black arrow pointing at the approximate integration location. **d**, (Left) An example of a supporting alignment between an amplified region of chr12 (top) and a non-amplicon region on chr6 (bottom) in CAKI-2. (Right) The approximate integration point on the amplicon is indicated with a black arrow pointing at the approximate integration location.


a
Reference
chr8 (amplicon)
Molecule 273194
Optical map
chr12 (integration)
H460
Coverage
Copy number
0
20
40
60
80
100
120
MYC
chr8
128590001
130739900
~ chr12:7660000
b
chr13 (amplicon)
Molecule 2284494
chr1 (integration #1)
← Centromeric region
K562
Coverage
Copy number
ABL1
NUP214
CLTCL1
BCR
chr9
chr13
chr22
133506901
134255600
80978782
81571100
90338706
95549336
108400001
108761410
18853122
20543752
20685373
23886003
~ chr1:142890000 (opposite orientation)
c
chr13 (amplicon)
Molecule 4604998
chr1 (integration #2)
← Centromeric region
K562
~ chr1:142890000
d
chr12 (amplicon)
Molecule 1178833
chr6 (integration)
CAKI-2
Coverage
CN
chr3
chr12
158851967
160222530
190806973
191483773
78366563
88413438
~ chr6:168380000

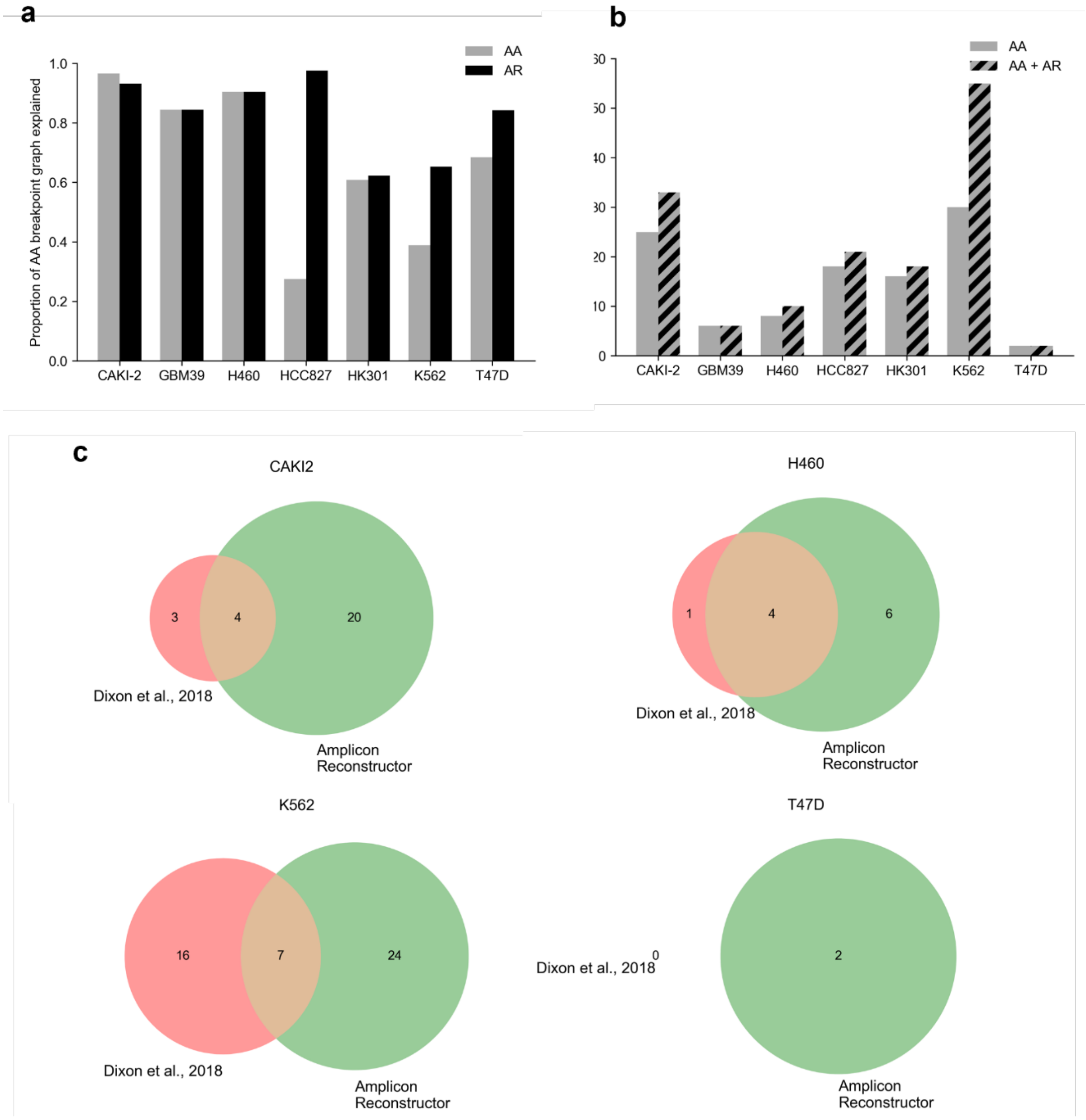
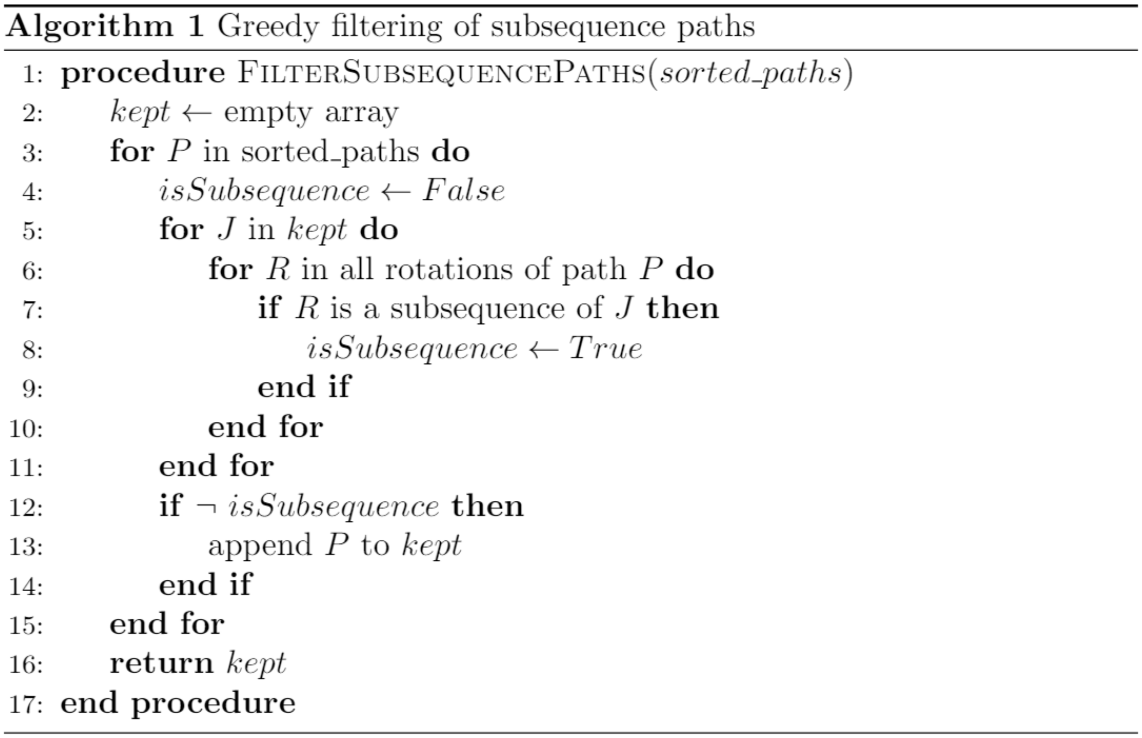


Figure 1.19: Comparing AR's results with AA output and Dixon et al. results. **a**, Comparison between the heaviest path/cycle generated by AA (grey) and the heaviest path/cycle generated by AR showing the proportion of amplified content in the breakpoint graph which is explained in the path/cycle for all samples studied by AA and AR. **b**, The number of breakpoint graph edges suggested by AA (grey) and the total number of edges inferred by AA and AR (black and grey stripes) as measured by the union of the junctions inferred by AR and the edges in the AA graph. **c**, Venn diagrams of the overlap between breakpoints detected by the Dixon et al. study (high confidence integrated breakpoints) and the breakpoints detected by AR for the amplicon regions analyzed by AR for CAKI-2, H460, K562 and T47D.



Algorithm 1 Greedy filtering of subsequence paths

```
procedure FilterSubsequencePaths($sorted\_paths$)
    $kept$ ← empty array
    for $P$ in sorted_paths do
        $isSubsequence \leftarrow False$
        for $J$ in $kept$ do
            for $R$ in all rotations of path $P$ do
                if $R$ is a subsequence of $J$ then
                    $isSubsequence \leftarrow True$
                end if
            end for
        end for
        if $\neg$ $isSubsequence$ then
            append $P$ to $kept$
        end if
    end for
    return $kept$
end procedure
```

Figure 1.20: Algorithm 1 - Greedy filtering of subsequence paths.

1.7 References

1. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell*. **144** (2011), pp. 646–674.

2. G. R. Bignell, C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, S. Widaa, J. Hinton, C. Fahey, B. Fu, S. Swamy, G. L. Dalgliesh, B. T. Teh, P. Deloukas, F. Yang, P. J. Campbell, P. A. Futreal, M. R. Stratton, Signatures of mutation and selection in the cancer genome. *Nature*. **463**, 893–898 (2010).

3. D. Stuart, W. R. Sellers, Linking somatic genetic alterations in cancer to therapeutics. *Curr. Opin. Cell Biol.* **21** (2009), pp. 304–310.

4. T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhang, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, R. Beroukhim, Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

5. 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, P. S. Mischel, Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*. **543**, 122–125 (2017).

6. V. Deshpande, J. Luebeck, N.-P. D. Nguyen, M. Bakhtiari, K. M. Turner, R. Schwab, H. Carter, P. S. Mischel, V. Bafna, Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10**, 1–14 (2019).

7. S. M. Carroll, M. L. DeRose, P. Gaudray, C. M. Moore, D. R. Needham-Vandevanter, D. D. Von Hoff, G. M. Wahl, Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Mol. Cell. Biol.* **8**, 1525–1533 (1988).

8. Y. Oobatake, N. Shimizu, Double-strand breakage in the extrachromosomal double minutes triggers their aggregation in the nucleus, micronucleation, and morphological transformation. Genes. Chromosomes Cancer. 59, 133–143 (2020).

9. P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, P. J. Campbell, Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*. **144**, 27–40 (2011).

10. D. T. W. Jones, S. Kocialkowski, L. Liu, D. M. Pearson, L. M. Bäcklund, K. Ichimura, V. P. Collins, Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res.* **68**, 8673–8677 (2008).

11. 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, E. T. Liu, The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell*. **34**, 197-210.e5 (2018).

12. B. McClintock, The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics*. **26**, 234–82 (1941).

13. D. Soler, A. Genescà, G. Arnedo, J. Egozcue, L. Tusell, Telomere dysfunction drives chromosomal instability in human mammary epithelial cells. *Genes Chromosom. Cancer*. **44**, 339–350 (2005).

14. K. Kitada, 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 Genet. Cytogenet.* **185**, 11–9 (2008).

15. 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, P. S. Mischel, Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. *Science (80-.)*. **343**, 72–76 (2014).

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

17. 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**, 699–703 (2019).

18. H. Cao, A. R. Hastie, D. Cao, E. T. Lam, Y. Sun, H. Huang, X. Liu, L. Lin, W. Andrews, S. Chan, S. Huang, X. Tong, M. Requa, T. Anantharaman, A. Krogh, H. Yang, H. Cao, X. Xu, Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience*. **3**, 34 (2014).

19. M. Li, A. C. Y. Mak, E. T. Lam, P. Y. Kwok, M. Xiao, K. Y. Yip, T. F. Chan, S. M. Yiu, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Springer Verlag, 2016), vol. 9683, pp. 67–79.

20. P. Chen, X. Jing, J. Ren, H. Cao, P. Hao, X. Li, Modelling BioNano optical data and simulation study of genome map assembly. *Bioinformatics*. **34**, 3966–3974 (2018).

21. A. R. Morton, N. Dogan-Artun, Z. J. Faber, G. MacLeod, C. F. Bartels, M. S. Piazza, K. C. Allan, S. C. Mack, X. Wang, R. C. Gimple, Q. Wu, B. P. Rubin, S. Shetty, S. Angers, P. B. Dirks, R. C. Sallari, M. Lupien, J. N. Rich, P. C. Scacheri, Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell* (2019),

doi:10.1016/j.cell.2019.10.039.

22. S. H. Mitsuda, N. Shimizu, Epigenetic Repeat-Induced Gene Silencing in the Chromosomal and Extrachromosomal Contexts in Human Cells. *PLoS One*. **11** (2016), doi:10.1371/journal.pone.0161288.

23. S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, Y. Kamatani, Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20** (2019), doi:10.1186/s13059-019-1720-5.

24. J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardımcı, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert, F. Yue, Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).

25. R. M. Layer, C. Chiang, A. R. Quinlan, I. M. Hall, LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15** (2014), doi:10.1186/gb-2014-15-6-r84.

26. F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. Von Haeseler, M. C. Schatz, Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*. **15**, 461–468 (2018).

27. M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. L. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. R. Hastie, D. Antaki, T. Anantharaman, P. A. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C. S. Chin, Z. Chong, N. T. Chuang, C. C. Lambert, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, D. U. Gorkin, M. Gujral, V. Guryev, W. H. Heaton, J. Korlach, S. Kumar, J. Y. Kwon, E. T. Lam, J. E. Lee, J. Lee, W. P. Lee, S. P. Lee, S. Li, P. Marks, K. Viaud-Martinez, S. Meiers, K. M. Munson, F. C. P. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. W. C. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stütz, D. C. J. Spierings, A. Ward, A. M. E. Welch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, E. Lowy, S. Yakneen, S. McCarroll, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P. Y. Kwok, P. M. Lansdorp, G. T. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. E. Talkowski, R. E. Mills, T. Marschall, J. O. Korbel, E. E. Eichler, C. Lee, Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1–16 (2019).

28. M. Dzamba, A. K. Ramani, P. Buczkowicz, Y. Jiang, M. Yu, C. Hawkins, M. Brudno, Identification of complex genomic rearrangements in cancers using CouGaR. *Genome Res.* **27**, 107–117 (2017).

29. Software Downloads - Bionano Genomics, (available at https://bionanogenomics.com/support/software-downloads/).

30. T. Anantharaman, B. Mishra, D. Schwartz, Genomics via optical mapping. III: Contiging

genomic DNA. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.*, 18–27 (1999).

31. A. Virgili, E. P. Nacheva, Genomic amplification of BCR/ABL1 and a region downstream of ABL1 in chronic myeloid leukaemia: A FISH mapping study of CML patients and cell lines. *Mol. Cytogenet.* **3** (2010), doi:10.1186/1755-8166-3-15.

32. R. K. Chandran, N. Geetha, K. M. Sakthivel, C. G. Aswathy, P. Gopinath, T. V. A. Raj, G. Priya, J. K. K. M. Nair, H. Sreedharan, Genomic amplification of BCR-ABL1 fusion gene and its impact on the disease progression mechanism in patients with chronic myelogenous leukemia. *Gene*. **686**, 85–91 (2019).

33. A. K.-Y. Leung, T.-P. Kwok, R. Wan, M. Xiao, P.-Y. Kwok, K. Y. Yip, T.-F. Chan, OMBlast: alignment tool for optical mapping using a seed-and-extend approach. *Bioinformatics*, btw620 (2016).

34. T. S. Anantharaman, B. Mishra, D. C. Schwartz, Genomics via Optical Mapping II: Ordered Restriction Maps. *J. Comput. Biol.* **4**, 91–118 (1997).

35. M. Qin, B. Liu, J. M. Conroy, C. D. Morrison, Q. Hu, Y. Cheng, M. Murakami, A. O. Odunsi, C. S. Johnson, L. Wei, S. Liu, J. Wang, SCNVSim: Somatic copy number variation and structure variation simulator. *BMC Bioinformatics*. **16**, 66 (2015).

36. L. F. Barr, S. E. Campbell, G. B. Diette, E. W. Gabrielson, S. Kim, H. Shim, C. V Dang, "c-Myc Suppresses the Tumorigenicity of Lung Cancer Cells and Down-Regulates Vascular Endothelial Growth Factor Expression 1" (2000).

37. S. W. Cho, J. Xu, R. Sun, M. R. Mumbach, A. C. Carter, Y. G. Chen, K. E. Yost, J. Kim, J. He, S. A. Nevins, S. F. Chin, C. Caldas, S. J. Liu, M. A. Horlbeck, D. A. Lim, J. S. Weissman, C. Curtis, H. Y. Chang, Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell*. **173**, 1398-1412.e22 (2018).

38. N. Vogt, A. Gibaud, F. Lemoine, P. De La Grange, M. Debatisse, B. Malfoy, Amplicon rearrangements during the extrachromosomal and intrachromosomal amplification process in a glioma. *Nucleic Acids Res.* **42**, 13194–13205 (2014).

39. C. T. Storlazzi, A. Lonoce, M. C. Guastadisegni, D. Trombetta, P. D'Addabbo, G. Daniele, A. L'Abbate, G. Macchia, C. Surace, K. Kok, R. Ullmann, S. Purgato, O. Palumbo, M. Carella, P. F. Ambros, M. Rocchi, Gene amplification as doubleminutes or homogeneously staining regions in solid tumors: Origin and structure. *Genome Res.* **20**, 1198–1206 (2010).

40. J. D. Rowley, A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*. **243**, 290–293 (1973).

41. G. Grosveld, T. Verwoerd, T. van Agthoven, A. de Klein, K. L. Ramachandran, N. Heisterkamp, K. Stam, J. Groffen, The chronic myelocytic cell line K562 contains a breakpoint in bcr and produces a chimeric bcr/c-abl transcript. *Mol. Cell. Biol.* **6**, 607–616 (1986).

42. S. Zakov, M. Kinsella, V. Bafna, An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5546–51 (2013).

43. M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, M. Loose, Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

44. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods.* **9** (2012), pp. 671–675.

45. M. Linkert, C. T. Rueden, C. Allan, J. M. Burel, W. Moore, A. Patterson, B. Loranger, J. Moore, C. Neves, D. MacDonald, A. Tarkowska, C. Sticco, E. Hill, M. Rossner, K. W. Eliceiri, J. R. Swedlow, Metadata matters: Access to image data in the real world. *J. Cell Biol.* **189** (2010), pp. 777–782.

46. S. Zakov, V. Bafna, Reconstructing Breakage Fusion Bridge Architectures Using Noisy Copy Numbers. *J. Comput. Biol.* **22**, 577–594 (2015).

47. M. Kinsella, V. Bafna, Combinatorics of the breakage-fusion-bridge mechanism. *J. Comput. Biol.* **19**, 662–678 (2012).

48. J. Burgin, C. Molitor, F. Mohareb, MapOptics: a light-weight, cross-platform visualization tool for optical mapping alignment. *Bioinformatics*. **35**, 2671–2673 (2019).

49. 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. *Nat. Genet.* **52**, 891–897 (2020).

50. R. P. Koche, E. Rodriguez-Fos, K. Helmsauer, M. Burkert, I. C. MacArthur, J. Maag, R. Chamorro, N. Munoz-Perez, M. Puiggròs, H. Dorado Garcia, Y. Bei, C. Röefzaad, V. Bardinet, A. Szymansky, A. Winkler, T. Thole, N. Timme, K. Kasack, S. Fuchs, F. Klironomos, N. Thiessen, E. Blanc, K. Schmelz, A. Künkele, P. Hundsdörfer, C. Rosswog, J. Theissen, D. Beule, H. Deubzer, S. Sauer, J. Toedling, M. Fischer, F. Hertwig, R. F. Schwarz, A. Eggert, D. Torrents, J. H. Schulte, A. G. Henssen, Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat. Genet.* (2019), doi:10.1038/s41588-019-0547-z.

51. E. K. F. Chan, D. L. Cameron, D. C. Petersen, R. J. Lyons, B. F. Baldi, A. T. Papenfuss, D. M. Thomas, V. M. Hayes, Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. *Genome Res.* **28**, 726–738 (2018).

52. A. C. Decarvalho, H. Kim, L. M. Poisson, M. E. Winn, C. Mueller, D. Cherba, J. Koeman, S. Seth, A. Protopopov, M. Felicella, S. Zheng, A. Multani, Y. Jiang, J. Zhang, D. H. Nam, E. F. Petricoin, L. Chin, T. Mikkelsen, R. G. W. Verhaak, Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease

evolution in glioblastoma. *Nat. Genet.* **50**, 708–717 (2018).

53. F. Morel, M.-J. Le Bris, A. Herry, G. Le Calvez, V. Marion, J.-F. Abgrall, C. Berthou, M. De Braekeleer, Double minutes containing amplified bcr-abl fusion gene in a case of chronic myeloid leukemia treated by imatinib. *Eur. J. Haematol.* **70**, 235–9 (2003).

54. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013) (available at http://arxiv.org/abs/1303.3997).

55. H. Li, R. E. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

56. M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer, G. D. Bader, Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*. **32**, 309–311 (2016).

57. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing (2012) (available at http://arxiv.org/abs/1207.3907).

58. E. Roller, S. Ivakhno, S. Lee, T. Royce, S. Tanner, Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics*. **32**, 2375–2377 (2016).

59. C. A. Miller, O. Hampton, C. Coarfa, A. Milosavljevic, ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*. **6** (2011), doi:10.1371/journal.pone.0016327.

60. X. Huang, M. S. Waterman, Dynamic programming algorithms for restriction map comparison. **8**, 1–520 (1992).

61. A. Valouev, L. Li, Y.-C. Liu, D. C. Schwartz, Y. Yang, Y. Zhang, M. S. Waterman, Alignment of Optical Maps. *J. Comput. Biol.* **13**, 442–462 (2006).

62. S. K. Das, M. D. Austin, M. C. Akana, P. Deshpande, H. Cao, M. Xiao, Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res.* **38** (2010), doi:10.1093/nar/gkq673.

63. S. Karlin, S. F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 2264–8 (1990).

64. G. Miclotte, S. Plaisance, S. Rombauts, Y. Van de Peer, P. Audenaert, J. Fostier, OMSim: a simulator for optical map data. *Bioinformatics*. **33**, 2740–2742 (2017).

CHAPTER 2: FaNDOM: Fast nested distance-based seeding of optical maps

2.1 Introduction

Optical Mapping (OM) is a rapidly maturing genome mapping technology whose historical antecedents are at least a few decades old (33). In the much older restriction mapping technique, the use of sequence-specific restriction sites in a genome enabled unique ‘fingerprints’ of the DNA. The initial restriction site maps were used to compare and position clones (genetic linkage maps) prior to sequencing (3). Now, optical mapping provides single-molecule readouts of the locations of fluorescently-labeled sequence motifs on long fragments of DNA, resolved to nucleotide-level coordinates (15). Despite the development of competing capillary sequencing and next generation sequencing methods, optical maps continue to play an important role in scaffolding and assembly. With the advent of microfluidic technologies for high-throughput of individual molecules and fluorescence-based visualization of covalently marked sites (labels), it is possible to generate high coverage (>100x of the human genome) with long OM molecules (>150 kbp) for $500-$1,000. For instance, the OM data-sets analyzed in this paper had a median length of 191 kbp.

As the optical mapping technology evolves, the error profiles found in OM data also change. Bionano optical mapping (Bionano Genomics, Inc., San Diego, CA) uses direct covalent labeling of fluorescent molecules onto DNA fragments, as opposed to previous generations of OM which used nickases. Its sources of error are orthogonal to DNA sequencing technologies (7), and currently include incomplete labeling of donor sequences, false-positive labels, and imprecise resolution about exact locations of imaged labels. Other technology-specific phenomena such as possible molecular chimerism or molecular stretching also contribute to error. Computational methods which handle OM data must capture these various errors.

Given its uses for scaffold construction in de novo assembly projects (46,37,30), optical mapping has matured to becoming a routine part of assembly pipelines for complex and/or large genomes. As a first step of this process, the OM fragments themselves are assembled into much larger (and error-corrected) OM contigs. The samples considered by our study had a median OM contig N50 of 38.4 Mbp. To achieve this, a computationally challenging problem of identifying overlapping OM fragments must be addressed. Much of the previous work about that problem uses dynamic programming algorithms to compare and align restriction maps (12), and now extends to optical maps (1,39). Newer methods such as Kohdista (27) and MalignerIX (25) tackle the overlapping fragment identification problems. Indexing and alignment-based methods have also been developed to map a sequence contig to a reference optical map genome, a requirement for scaffolding (26,16).

Here, we consider the slightly different problem of mapping optical maps to a reference human genome for the purposes of identifying structural variants (9,5). Such methods have been effective in identifying genomic abnormalities in Mendelian disease (2,8) as well as cancer (6,22,29). Due to similar algorithmics, general methods for pairwise alignment or scaffolding including Valouev (40), SOMA (28), TWIN (26), MalignerDP (25) could be used in principle for mapping optical maps to an in silico digested reference genomic sequence. However, most of these methods do not repurpose well in practice, especially on data from the latest Bionano platform. Moreover, they do not call structural variants. In contrast OMBlast (18), and RefAligner (35) have previously demonstrated superior performance on Bionano data (18,44). RefAligner specifically has been configured to call SVs. A new software, OMSV (19) now combines RefAligner and OMBlast output to call SVs. Notably, RefAligner is a closed-source proprietary method, available only as pre-compiled binaries for specific hardware, and is very resource intensive, as described in the Results.

We introduce FaNDOM (Fast Nested Distance seeding of Optical Maps) - an optical map alignment tool which introduces a novel method for seeding optical map alignments, greatly reducing the search space of the alignment process. FaNDOM is specifically optimized to handle data from the Bionano Saphyr optical mapping technology. The algorithmic and technology specific improvements allow us to be significantly (4-14X) faster than competing tools while maintaining sensitivity and specificity. We used FaNDOM to map variants in 3 cancer cell-lines and identified many structural variations, including deletion of tumor-suppressor genes, duplications, gene fusions and gene-disrupting rearrangements. FaNDOM is publicly available at https://github.com/jluebeck/FaNDOM.

2.2 FaNDOM results

As OMBlast (18,17) and RefAligner (35) were the best performing pre-existing methods for mapping Bionano optical maps to a reference genome, we compared performance of FaNDOM against Bionano RefAligner (Solve 3.5.1) and OMBlast (OMTools Version 1.4a). We also attempted to benchmark TWIN and Kohdista, but they are not specifically designed for this problem and did not perform as well. TWIN and Kohdista did not support the Bionano Saphyr technology file formats, which is the dominant platform currently, and required for us to write custom file format converters to convert the modern .bnx, and .cmap files into older file formats accepted by TWIN and Kohdista.

We took 10,000 OM molecules from NA12878 and used them as queries to align to the in silico digested human reference genome. However, it did not return any mappings. We did test that by using an identical sub-molecule from the *in silico* digestion, we were able to match, suggesting that TWIN showed poor tolerance for missing/false OM labels. TWIN results were previously demonstrated on simulated optical maps and optical map data with error profiles very different from Bionano Saphyr, so it is possible that a change of parameters could have changed the results. However, in personal communication, the authors did not recommend specific parameter settings appropriate for Bionano Saphyr.

We also attempted to use Kohdista for reference based-mapping. Its RAM usage was not optimized for the large human genome reference. Therefore, we tested a small group of 243 optical map molecules derived from the genomic reference chr10:0-46,272K (46 Mbp). Kohdista used more than 130GB of RAM and did not return alignments after 12 hours. Also, when we tried to align this group of molecules to the entire chromosome 10, the RAM usage surpassed 200 Gb. Since optical map data sets frequently contain > 1 million molecules, we concluded that Kohdista was not an appropriate tool for our problem.

Saphyr optical map data is publicly available for samples NA12878, GM09888, GM08331, and GM24143. We collected 270,000 raw molecules from each sample, where more than 85% of each molecule aligned to reference, as reported by Bionano, using their own RefAligner tool. We then ran FaNDOM, OMBlast and RefAligner on this testing set.

2.2.1 FaNDOM running time

We note that RefAligner is already highly optimized for the Saphyr technology and is only provided as precompiled binary code that runs on specific machine architectures. All experiments were conducted on an Intel(R) Core(TM) i9-9900 CPU @3.10GHz with 32 GB of main memory running Ubuntu 18.04.3 LTS (Bionic Beaver), using 10 threads. The results (Figure 2.1a) showed that FaNDOM was 4-6 X faster than OMBlast and 13-14 X faster than RefAligner on all data-sets, highlighting the speedups created by our filtering methods. FaNDOM required approximately 2-2.5 GB of RAM for each thread. While OMBlast required less memory, the memory usage increased with increase in molecule size, and did not scale well for Saphyr assembled contigs. The OMBlast documentation suggests 200Gb RAM for mapping assembled contigs.

2.2.2 Mapping accuracy

We compared the accuracy FaNDOM, RefAligner, and OMBlast reported mappings on simulated and real-data. Unlike DNA-sequencing read mapping, which has discrete character matches & mismatches, it is not trivial to designate an OM molecule alignment as correct or incorrect on real-data. Instead, we treated a mapping as correct if it was supported by at least two of the three methods.

We simulated data-sets with ‘high’ and ‘low’ error, where high (H) corresponded to a false positive label rate of 4 per 100 kbp, and stretch factor with standard deviation 0.02, which matched the Saphyr technology. Low (L) corresponded to a false positive label rate of 1 per 100 kbp and stretch factor with standard deviation 0.01. All tools performed well on low-error. On high-error data, the 3 methods had very similar recall with FaNDOM marginally higher, while FaNDOM precision lay in between RefAligner and OMBlast (Figure 2.1b). On the cell-lines, RefAligner had the highest precision and recall followed by FaNDOM and OMBlast. We note that RefAligner is better positioned to incorporate specifics of the Saphyr technology. The lower recall for FaNDOM relative to RefAligner, can be partially attributed to the occasional removal of true maps during the filtering step. The precision can be improved by post-alignment filtering and will be part of future release of FaNDOM after more data-sets have been analyzed.

FaNDOM was 5X and 15X faster than OMBlast and RefAligner on cell-lines (Fig. 2.1a) as well as simulations. As expected, simulations show that the running time increases with higher error rate for all methods.

2.2.3 SV detection

Structural variant analysis continues to be a challenging problem requiring consensus from different methods and technologies. We compared the three methods using a benchmark of SV deletion calls of length > 2000bp on the genome NA12878. The benchmark was created previously using a multitude of technologies (31). Figure 2.2a compares the performance of FaNDOM and RefAligner using assembled OM contigs. FaNDOM and RefAligner had comparable recall identifying 77% and 79% of the high-confidence calls, respectively, despite FaNDOM using filtering strategies to make the runtime faster by an order of magnitude. FaNDOM was much more aggressive in calling deletions compared to RefAligner. Spot-checking, many of the FaNDOM specific deletion calls appeared to be accurate (e.g., see Fig. 2.2f).

OMSV (19) is another recent method for detecting SVs with OM data. It is an integrative tool that combines the output of RefAligner and OMBlast together and is therefore even more compute-intensive. As we could not run OMBlast on Saphyr contig data, we compared FaNDOM calls against pre-computed OMSV calls on NA12878 mapped to the hg38 reference and compared the calls to a benchmark deletion call-set (9) on the hg38 reference (Fig. 2.2b). The FaNDOM recall was 84% compared to the 70% recall of OMSV.

Detecting genomic insertions is one of the advantages of long-read technologies. FaNDOM predicted 719 insertions (Fig. 2.2c). While there is no established call-set of insertions for NA12878, 73% of the FaNDOM calls were previously reported as insertion polymorphisms in the Database of (human) Genomic Variants (DGV) (23). FaNDOM additionally identified a few ultra-long insertions in OM contigs (Fig. 2.2d) that would be challenging with any competing technology due to the insertion size.

We investigated the FaNDOM specific SV calls for possible error. The high-confidence data set (31) has been collected by integrating a number of technologies and is likely to be accurate. Nevertheless, many of its calls were discovered using short-reads, while many of the FaNDOM specific calls were > 15 kbp (e.g., see Fig. 2.2e). Additionally, some of the FaNDOM specific calls are in regions of low mappability (typically low complexity or repetitive sequence). Those breakpoints typically cannot be captured by short reads, but can be captured by long OM contigs (e.g. chr19:37,760K-37,795K; Fig 2.2f), demonstrating the complementarity of OM data to sequencing technologies. Moreover, assembled optical map contigs enable the detection of multiple breakpoints in one contig. As an example, Fig. 2.2g represents an assembled OM contig from K562 cell line that cover translocation from chr9 to chr13 and multiple breakpoints in chr13 spanning 500 kbp.

2.2.4 SVs in cancer cell-lines

We ran FaNDOM on assembled OM contigs as well as OM molecules for cancer cell-lines K562, CAKI-2 and H460 - all which are known to carry extensive rearrangements. Table 2.1 summarizes some of the rearrangements identified by FaNDOM on assembled OM contigs. The rearrangements identified by FaNDOM which included 1,800 large (> 2 kbp) indels, 133 inter-chromosomal translocations, 28 foldback reads, and 223 breakpoints that disrupted an existing gene, among other rearrangements. In this study, we focused specifically on genes that were deleted, and on translocations that disrupted or fused two genes.

The lung cancer cell line NCI-H460 has previously been documented to bear a focal amplification of the MYC/PVT1 region due to extrachromosomal DNA (ecDNA) and also it has been found to have evidence for intrachromosomal amplification in a homogeneously staining region (HSR) (38). Previous reconstruction of the MYC amplified region revealed a complex duplicated structure which suggested that the ecDNA element containing MYC/PVT1 had reintegrated as an HSR in a non-native location (22). The FaNDOM analysis identified a translocation from within the amplified ecDNA structure (chr8:128,745 kbp) to a non-native location (chr12:7,665K; Fig. 2.3a) revealing chr12 to be the site of the HSR. Figure 2.3a also supports an inverted duplication at Chromosome 8 as part of the amplified structure. In addition to recapitulating the breakpoints of the ecDNA, the FaNDOM analysis identified many partial or complete deletions of tumor suppressor genes, including LRP1B (21) (chr2:141,735K-142,155K), TUSC7A (20) (non-coding; chr3:116,295K-116,775K), FHIT (41) (chr3:60,405K-60,735K), LSAMP (14) (chr3:115,545K-116,145K). Notably, many of these deletions were on chr3. Many other rearrangements were identified providing a scenario of complex rearrangements in the cell-line.

In the renal cancer cell-line CAKI-2, we observed deletions or disruptions involving tumor suppressor genes including CFHR1 (42) (chr1:196,665K-197,295K), RNF217 (chr6:125,265K-125,505K) (10), RBFOX1 (chr16:6,585K-7,155K) (34), FBXL7 (chr5:15,825K-15,945K) (11). We also observed two fusions: TECRL1/GRIP1 (chr4:65,205K, −, chr12:66,975K, −) and RACGAP1/AKAP6 (Fig. 2.3b, chr12:50,385K, −, chr14:33,255K, +). RACGAP1 displays tumor malignancy potential (13) and is known to fuse with other genes such as CERS5 and RAB34 (43).

K562 is a chronic myelogenous leukemia cell-line with the Philadelphia Chromosome. It was comprehensively analyzed recently using a multitude of technologies including whole genome sequencing and Hi-C (45). FaNDOM confirmed some of the rearrangements of the previous study such as the BCR-ABL1 fusion (Fig. 2.3c), between chr22 and chr9. Among other rearrangements, we also observed an atypical microdeletion in 22q11, almost identical to a deletion previously associated with a congenital syndrome (36), and a subset of a larger deletion reported for DiGeorge syndrome. The deletion encompasses the genes GSTT1, GSTT2 and GSTT2B and deletions in these genes have previously been associated with esophageal cancer (24).

While our results often matched the previously reported SVs (45), there were a few notable differences. For example, in contrast with the previous finding of an inversion involving ORC6, MYLK3 on chr16, we observed a deletion (16:46,725K-46,845K, Fig. 2.3d) that partially removed ORC6 as well as a microinversion involving MYLK3. In a second example, the Zhou et al. study also identified a fusion of CDC25A/GRID1 (45). While we observe the same translocation, the directionality provided by the long reads suggests the disruption of the two genes, but not a fusion product (Fig. 2.3e). We could confirm other chromosome 16 rearrangements, including an inverted duplication (88,605K-88,785K), and another inverted duplication at chr13:92,475K (Fig. 2.3f).

2.3 Discussion

Improvements to the optical mapping technology in terms of accuracy and cost has made it competitive for structural variant detection. At the same time, the raw data is harder to interpret and motivates the development of public domain tools for interpretation. In this paper, we focus on speeding up the mapping by relying on a novel filtering strategy that greatly improved speed without a significant loss of accuracy. The filtering relies on two ideas: (a) for most high-quality optical maps, it is relatively easy to find seeds that locate the reference target region for a query, and (b) by merging distances, thousands of queries can identify their target seeds in a single search-and-merge strategy. The results demonstrate the viability of this trade-off, leading to high speedups over other tools with only a small loss of sensitivity.

We recognize that our proposed method uses many parameters and for the most part, the parameters are empirically determined to work for Saphyr. The optimal parameter values will be determined only after a large number of data-sets have been analyzed and will need to be retrained for newer technologies. Additionally, non-human genomes such as plants may also require some significant recalibration of parameters and low-complexity annotations, which we have not yet explored. Nevertheless, because we have used FaNDOM to analyze many tens of thousands of molecules, the current choice of parameters appears to be robust for the current technology. Taken together, our results point to the value of using optical mapping as a complementary technology for structural variation identification.

The detection of structural variants is a key benefit of the OM technology, but it is harder to benchmark given the lack of large-scale, robust truth data sets. Our results suggest that FaNDOM can identify discordant alignments and breakpoints with high sensitivity. As many of the calls are based on cut-offs that can be adjusted, the results do not reveal any fundamental limitation of the filtering but indicate a lack of additional calibration against a true gold-standard. Additional analysis will be needed to identify systemic sources of false positive calls.

We note that calling the structural variation mechanism itself is a secondary process that will require integration with other information including copy number changes, and this will be a topic of ongoing research. For example, one possible improvement includes pruning deletion calls by limiting results to the regions with a decrease in copy number consistent with heterozygous or homozygous deletion. With further improvements and methods development, OM technologies could be used to replace cytogenetics as a method of choice for revealing large-scale genetic abnormalities in Mendelian diseases and cancer (2,29,22).

2.4 Method details

Conceptually, we define an optical map as a sorted list of numeric values, representing the relative positions of labels on a fragment of DNA (Figure 2.4a). These numeric lists can be generated for any collection of individual OM molecules, assembled OM molecules, or from in silico predicted label positions on the reference genome. FaNDOM utilizes standard optical map data formats (.bnx or .cmap) where each imaged DNA fragment has been pre-converted to label position lists specified in base pair coordinates. An overview of the structure of the FaNDOM software is available in Figure 2.4b.

2.4.1 Preprocessing

Query fragments with length < 25 kbp or containing less than 10 labels were filtered out from mapping. Similarly, queries containing consecutive labels with distance > 250 kbp were removed. Scaling refers to a systematic translation of physical inter-label distances into nucleotide distances. We define ‘scaling’ and ‘stretch’ as two independent error modalities which we examined when benchmarking FaNDOM. Scaling refers to the calibration of measured base pairs per pixel in the imaging of optical map molecules by the instrument. If this calibration is not completely accurate, we observed that this error can lead to global lengthening or shortening of all molecules derived from the instrument. To ensure that optical map data has been properly scaled following the image processing performed by the Bionano instrument, we apply a grid-search method to try a range of re-scaling factors. Stretch on the other hand refers to the physical lengthening or shortening of individual DNA fragments traveling through the nanochannel array. It is accounted for after ’scaling’ has been resolved. The Bionano Saphyr instrument performs a calibration to scale distances, estimating the number of base pairs present per image pixel. The process can on occasion be erroneous (32). To recalibrate, FaNDOM randomly selects 250 molecules and estimates a corrected scaling factor using a grid search in a range of values between 0.96 to 1.2. The range was determined by experimenting from a set of 38 human samples. The rescaled molecules in each iteration are aligned to the reference. The scaling factor that achieves the highest total alignment score is selected for rescaling molecules prior to alignment.

Assembled OM contigs can be very large often exceeding thousands of labels. As the alignment time grows quadratically with length, FaNDOM pre-processes assembled OM contigs by splitting them into smaller fragments, each containing 75 labels, with an overlap of 50 labels between endpoints of consecutive fragments. When alignment is completed, FaNDOM merges the alignments from overlapping fragments from assembled OM contigs to produce a complete alignment for the OM contig. In the case of conflicting alignments between overlapping contig fragments, FaNDOM maintains both partial alignments.

We converted the reference genome into a collection of expected label locations based on in silico the presence of the labeling motif throughout the reference. If the distance between two consecutive reference labels is less than 800bp, they are replaced with the average of the two locations to account for the potential inability of resolving nearby OM labels. We also adapted a Bionano method (35) to identify and mask low-complexity regions in the human genome. Formally, denote a low-complexity region as containing at least 5 consecutive labels where the distance between adjacent labels is identical within 10% tolerance. Those could result in spurious alignments and are masked out. Specifically, in reference genome build hg19, 1.5 Mbp which (0.04% of total reference genome) was masked out, while in hg38, 2.8 Mbp (0.09% of the total reference genome) was masked out.

2.4.2 Optical map alignment with FaNDOM

The crux of a mapping procedure is an alignment of an optical map query to an in silico optical map of a reference sequence interval. The alignment maps query labels to the reference labels so that the inter-label distances between the query and reference are preserved (Figure 2.4a). The alignment of optical maps is a well-studied problem (33,40). FaNDOM's scoring function follows previous methodologies but diverges slightly. Consider reference R of length m and reference *Q* of length n labels. For $j \leq m$ and $q \leq n$, define $S[j][q]$ as the optimum score of aligning a subsequence (local alignment) ending at label *j* on *R* with a subsequence ending at label *q* on query *Q*. *S* can be computed using the following banded dynamic programming recurrence, where the band size is *d*:

$$S[j][q] = \max S[i][p] + \text{Score-region}(R, i, jQ, p, q)$$

$$\max\{0, j - d\} \leq i < j,$$

$$\max\{0, q - d\} \leq p < q.$$

Score-region scores a match after penalizing for discrepancies in the match. Specifically, for $i < j, p < q$, let $fn = (q - p - 1), fp = (j - i - 1)$ denote the number of unmatched labels in the query and reference, respectively. Then,

$$\text{Score-region}(R, i, j, Q, p, q) = L - c(fn + fp) - |(R[j] - R[i]) - (Q[q] - Q[p])|^k$$

We set *L* = 10,000 to represent a perfect match score. Empirical tests indicated that a wide range of *k, c* showed identical performance. Increasing *k, c* resulted in the same alignments but with tighter boundaries. We chose the distance scale parameter *k* = 1.15 and false-label parameter *c* = 3000. After computing initial alignments for molecules, FaNDOM then identifies molecules which are candidates for local/partial alignment discovery, as a prelude to structural variant analysis. In this partial alignment mode (See "Computing partial alignments for SV detection" below), where split-molecule alignments are allowed, FaNDOM computes more stringent partial alignments (*c* = 7500, *k* = 1.4).

2.4.3 Alignment running time suggests the necessity of filtering

The ungapped alignment algorithm has complexity $O(mnd^2)$. Despite algorithmic improvements and optimizations, our empirical results suggested that aligning a collection of two million OM fragments representing (100X) whole genome coverage against every position on the human genome would take ~700,000 CPU-hours. While assembly of OM fragments into contigs reduces the number of query sequences, the OM contigs are longer and the estimated time remains ~15,000 CPU-hours. Therefore, similarly to the Bionano RefAligner (35), and OMBlast (18,17), we deploy a filtering strategy, where for each query molecule. The goal is to identify a small collection of reference intervals to align the query with. The filter must be fast, sensitive (defined by the probability of the true reference location being included in the filtered reference intervals), and efficient (defined by the number of filtered regions per query–smaller being better). The filtered regions, or seeds are used to compute alignments and return the full or partial mappings of each query OM fragment or contig.

2.4.4 Search-and-Merge filtering for optical maps

The key idea of filtering is that in a correct alignment there are some parts of query and reference which are highly similar to each other, or that all inter-label distances in those regions are practically equivalent. Let $R[i,j]$ (respectively $Q[i,j]$) denote the genomic distance between labels *i*, *j* in *R* (respectively, *Q*). Denote a window W_a in the reference as a collection of distances $R[i,j]$ for all $a \leq i < j < a + 3$. Windows W_b, in the query OMs are defined similarly. Let

$$W_b \underset{\text{match}}{\longrightarrow} W_a \iff \forall x \in W_b, \exists y \in W_a : |x - y| \leq T$$

A default value of T = 350 was chosen empirically. In the Search-and-Merge procedure, we sort all genomic distances from every window of the reference (typically a chromosome) to a list L_m (Fig. 2.5a). Similarly, for a collection of query OMs, we merge all sorted distances from all windows of each query in the collection into list L_n. Each distance $x \in L_m$ (respectively, $y \in L_n$) is associated with all reference windows (respectively, query windows) containing distance x (respectively, y).

Next, the sorted lists L_m, L_n are 'search-merged' (Fig. 2.5a). For each element $x \in L_n$ we perform two binary searches to identify the smallest and largest distances $y_1, y_2 \in L_m \ni x - y_1 \leq T, y_2 - x \leq T$. For all 'matches' (*x, y*) where $y_1 \leq y \leq y_2$, we increment the match score of all window pairs associated with *x* and *y*. Finally, for all reference labels *a*, query labels *b* such that $W_b \underset{\text{match}}{\longrightarrow} W_a$, a seed (*a, b, o*) is generated with $o \in \{+, -\}$ representing direction of match.

2.4.5 Packing seeds into bands

For each reference label a, and each query OM, FaNDOM explores a diagonal band B_a around a of width B_w (default value $B_w = 12{,}000$). Label *a* is filtered out if B_a contains fewer than $T_h = 4$ seeds. For retained bands, an edge weighted directed acyclic graph (DAG) *G* is constructed as follows: Each node *u* in *G* corresponds to a pair of (query, reference) labels (u_q, u_r), where u_q (respectively, u_r) represents the nucleotide distance of the query label (respectively, reference label) from the first query (reference) label. Also, add nodes $s = (0, 0)$

and $t = (l, l)$ corresponding to the start and end of band B_a. For each seed u in the band, designate nodes u_1, u_2, u_3 corresponding to start, middle, and end of the seed. With few exceptions, we use Euclidean distances for edge weights so that

$$w(u, v) = \|v - u\| = \sqrt{(v_r - u_r)^2 + (v_q - u_q)^2}$$

Specifically,

1. For each seed u, add edges (u_1, u_2) and (u_2, u_3) with weights 0 each; edge (s, u_1) with weight $\sqrt{2}(l - u_{3q})$.
2. For each pair of seeds u, v such that $u_{3q} \leq v_{1q}$ and $u_{3r} \leq v_{1r}$, add edge (u_3, v_1) with weight $\|v_1 - u_3\|$.
3. For each pair of seeds u, v such that $u_{2q} = v_{1q}$ and $u_{2r} \leq v_{1r}$, add edge (u_2, v_1) with weight $\|v_1 - u_2\|$.

We use dynamic programming to compute the weight w_{st} of the shortest (least-weight) paths from *s* to *t* in *G*. The score of B_a is given by

$$\text{Score}(B_a) = 1 - \frac{w_{st}}{\|t - s\|}$$

A similar process is used for seeds in the reverse direction, with $s = (0, l), t = (l, 0)$. For each query OM, we save the highest scoring 150 bands.

As a first idea, we could align the query map with the reference region for each of the 150 bands, and still achieve high speed and sensitivity. However, we observed that in some cases, the top scoring bands were significantly more likely to yield true alignments than other high-scoring bands, and the correct region was near the tail of the band score distribution could be identified without aligning every candidate. We empirically fit the band-scores to an exponential distribution with parameter λ and used the following empirical guidelines for scoring. For each query

$$\text{max score}(B_a) = \begin{cases} > 2.2\lambda, & \text{Align top 10 bands} \\ > 1.7\lambda, & \text{Align top 50 bands} \\ > 1.5\lambda, & \text{Align top 100 bands} \\ \text{else}, & \text{Align top 150 bands} \end{cases}$$

A band that is selected for alignment is converted to reference alignment boundaries by using the reference coordinate s_r of the source node s, and the query molecule *Q* of length |Q|. Specifically, for a padding factor *p* (default $p = 1000$), the region $s_r - p$ to $sr + |Q| + p$ on the reference is used to align to the query molecule.

2.4.6 Computing partial alignments for SV detection

We identify structural variants in two steps. First, queries that are either a) unaligned, b) have a mean alignment score less than 5000/label, c) the alignment does not cover 80% of the query length, or d) has a total alignment length 25 kbp, are targeted for partial alignments. The banding procedure is identical. For partial alignments, we compute local shortest paths between all pairs of seeds u, v as long as $\|v_3 - u_1\| \geq 20$ kbp and the path contains at least 4 labels. If the corresponding band-score

$$\left(1 - \frac{w_{(u1,v3)}}{\|v_3 - u_1\|}\right) \geq 0.4$$

Then the region gets a score of $v_{3q} - u_{1q})\left(1 - \frac{w_{(u1,v3)}}{\|v_3 - u_1\|}\right)$, and the top 300 candidate regions, each designated by a pair of nodes, are selected for alignment and re-ranking. A gapped-alignment module is used and if the score exceeds a threshold, the partial or gapped alignment is reported.

FaNDOM identifies discordant alignments (defined below) and breakpoints, which form the core of any SV discovery strategy, and defers the calling of actual SVs to a subsequent script that can be customized by the user. Recall that an alignment is a chain of matches $(q_0, r_0), (q_1, r_1), \ldots, (q_t, r_t)$. For alignments below a threshold score, if there exists $0 \leq i < t$

such that (a) $|(q_{i+1} - q_i) - (r_{i+1} - r_i)| > 2000$, (b) $|q_{i+1} - q_0| \geq \max\{10000, 0.25 \times |Q|\}$, and (c) $q_t - q_{i+1}| \geq \max\{10000,\ 0.25 \times |Q|\}$, then a discordant alignment is called. Discordant alignments typically represent insertions/deletions but may also represent small inversions flanked by high quality alignments on both sides, or other structural variants.

Breakpoints refer to a pair of coordinates that are non-adjacent on the reference but are together on the query. Consider two partial alignments that involve the same query molecule, described by $A_1: (q_0, r_0), \dots, (q_i, r_i)$ and $A_2: (q_j, r_j), \dots, (q_t, r_t)$. Note that $r_0, \dots, r_i$ could potentially be on a different chromosome than $r_j, \dots, r_t$. Define $o_i, o_j \in \{+, -\}$ using $o_i = \mathrm{sgn}(r_i - r_o)$ and $o_j = \mathrm{sgn}(r_t - r_j)$. FaNDOM calls a breakpoint (r_i, o_i, r_j, o_j) if there is no partial alignment involving the labels between q_i and q_j. Breakpoints are clustered if their endpoints are within 30 kbp, and each breakpoint is listed along with its “support”, or the number of alignments consistent with the breakpoint. Subsequent scripts are used to describe the rearrangement that creates the breakpoint. For example, $(r_i, +, r_j, +)$ describes a homozygous (respectively heterozygous) deletion if r_i and r_j are on the same chromosome and the fragment coverage in the interval $[r_i, r_j]$ is 0 (respectively, half of normal coverage).

2.5 Acknowledgements

Chapter 2, in full, is a reprint of the material “FaNDOM: Fast nested distance-based seeding of optical maps” by Siavash Raeisi Dehkordi, Jens Luebeck and Vineet Bafna as it appears in Cell Pattens 2021. The dissertation author was one of the primary investigators and authors of this paper. We would like to thank Andy Pang of Bionano Genomics, Inc. for his feedback and assistance with data interpretation and explanations of the Bionano pipelines.

2.6 Appendix

Table 2.1: Rearrangements in cancer cell lines.

Cell line	Indels	Interchromosomal translocations	Foldback maps	Gene-disrupting breakpoints
CAKI-2	626	56	7	95
H460	571	26	4	62
K562	603	21	17	66

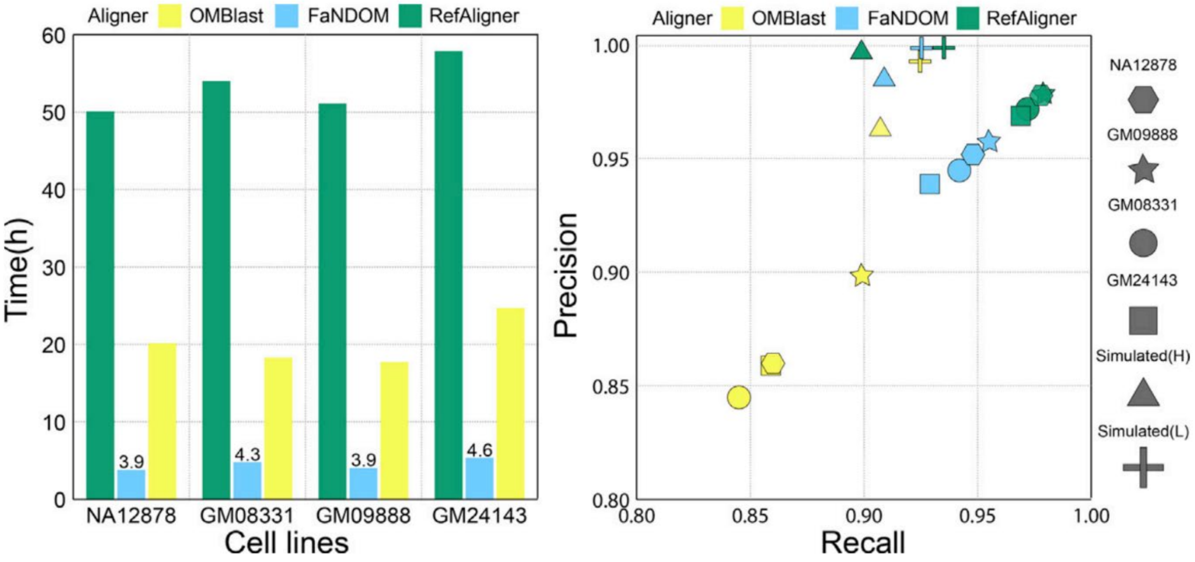
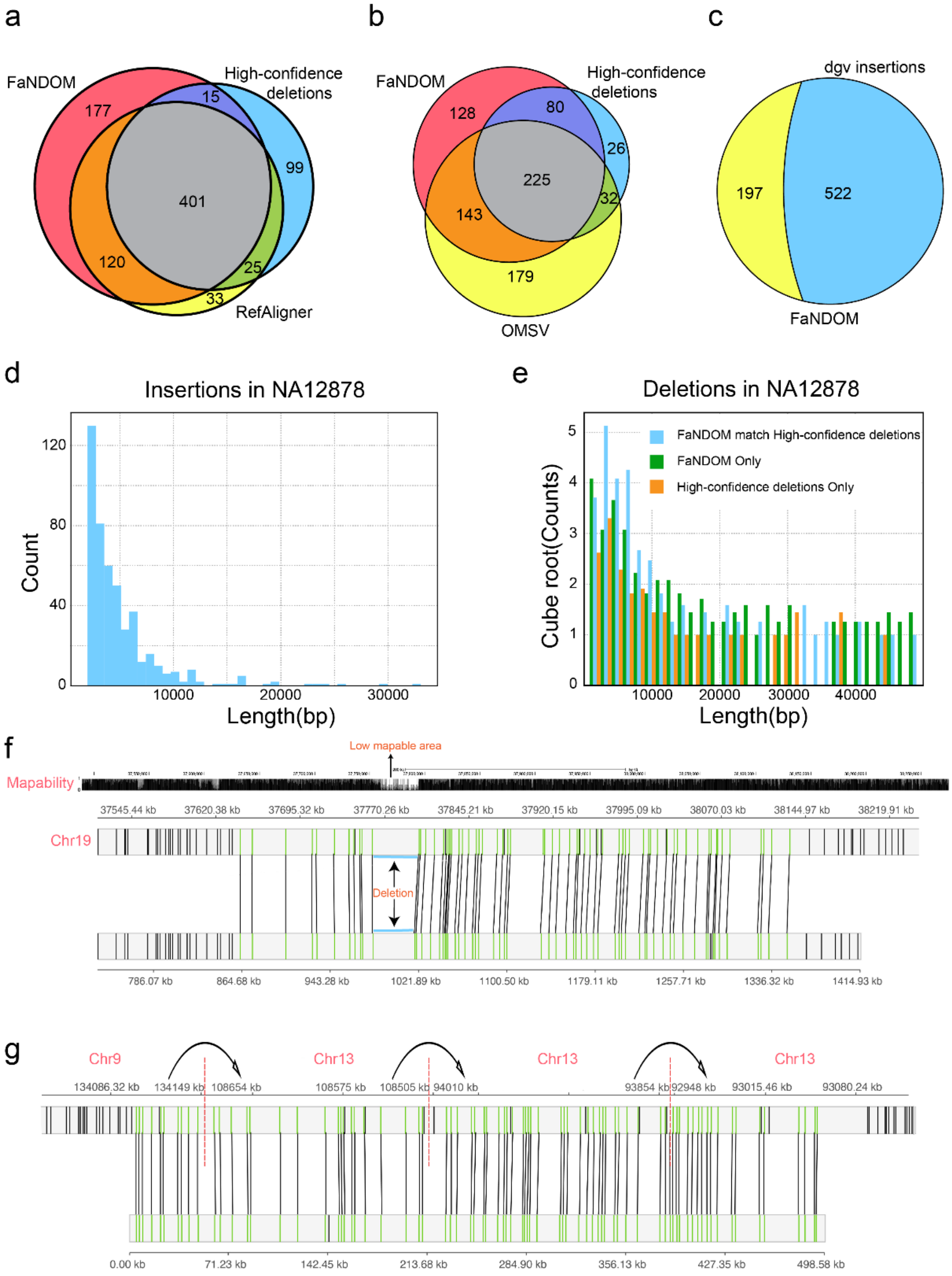


Figure 2.1: FaNDOM performance. (left) Running time, (right) Accuracy. The set of true-positives (TP) were all mappings identified by at least 2 of the 3 methods. Recall=TP/(TP+FN), Precision=TP/(TP+FP).

Figure 2.2: SV calling performance. **a,** Comparison of FaNDOM and RefAligner deletion calls on NA12878 against a benchmark data set from Parikh et al., using the hg19 reference. **b,** comparison of FaNDOM and OMSV deletion calls on NA12878 against a benchmark created using multiple sequencing technologies published in Dixon et al. using the hg38 reference. **c,** Insertions identified by FaNDOM for NA12878. The blue region signifies insertion polymorphisms identified by FaNDOM also in the Database of Genomic Variants. **d,** Length distribution of FaNDOM insertion calls for NA12878. **e,** Length distribution of FaNDOM and benchmark deletion calls (Parikh et al.). **f,** A FaNDOM deletion not in the Parikh et al. benchmark data-set likely due to its presence in a low mappability region. **g,** A FaNDOM alignment using assembled OM contigs that chains multiple breakpoints across 400 kbp on the K562 cell-line. OM alignment visualizations were generated with MapOptics.



a
FaNDOM
High-confidence deletions
RefAligner
177
15
99
401
120
25
33
b
FaNDOM
High-confidence deletions
OMSV
128
80
26
225
32
143
179
c
dgv insertions
FaNDOM
197
522
d
Insertions in NA12878
Count
0
40
80
120
10000
20000
30000
Length(bp)
e
Deletions in NA12878
Cube root(Counts)
0
1
2
3
4
5
FaNDOM match High-confidence deletions
FaNDOM Only
High-confidence deletions Only
10000
20000
30000
40000
Length(bp)
f
Low mapable area
Mapability
37545.44 kb
37620.38 kb
37695.32 kb
37770.26 kb
37845.21 kb
37920.15 kb
37995.09 kb
38070.03 kb
38144.97 kb
38219.91 kb
Chr19
Deletion
786.07 kb
864.68 kb
943.28 kb
1021.89 kb
1100.50 kb
1179.11 kb
1257.71 kb
1336.32 kb
1414.93 kb
g
Chr9
Chr13
Chr13
Chr13
134086.32 kb
134149 kb
108654 kb
108575 kb
108505 kb
94010 kb
93854 kb
92948 kb
93015.46 kb
93080.24 kb
0.00 kb
71.23 kb
142.45 kb
213.68 kb
284.90 kb
356.13 kb
427.35 kb
498.58 kb

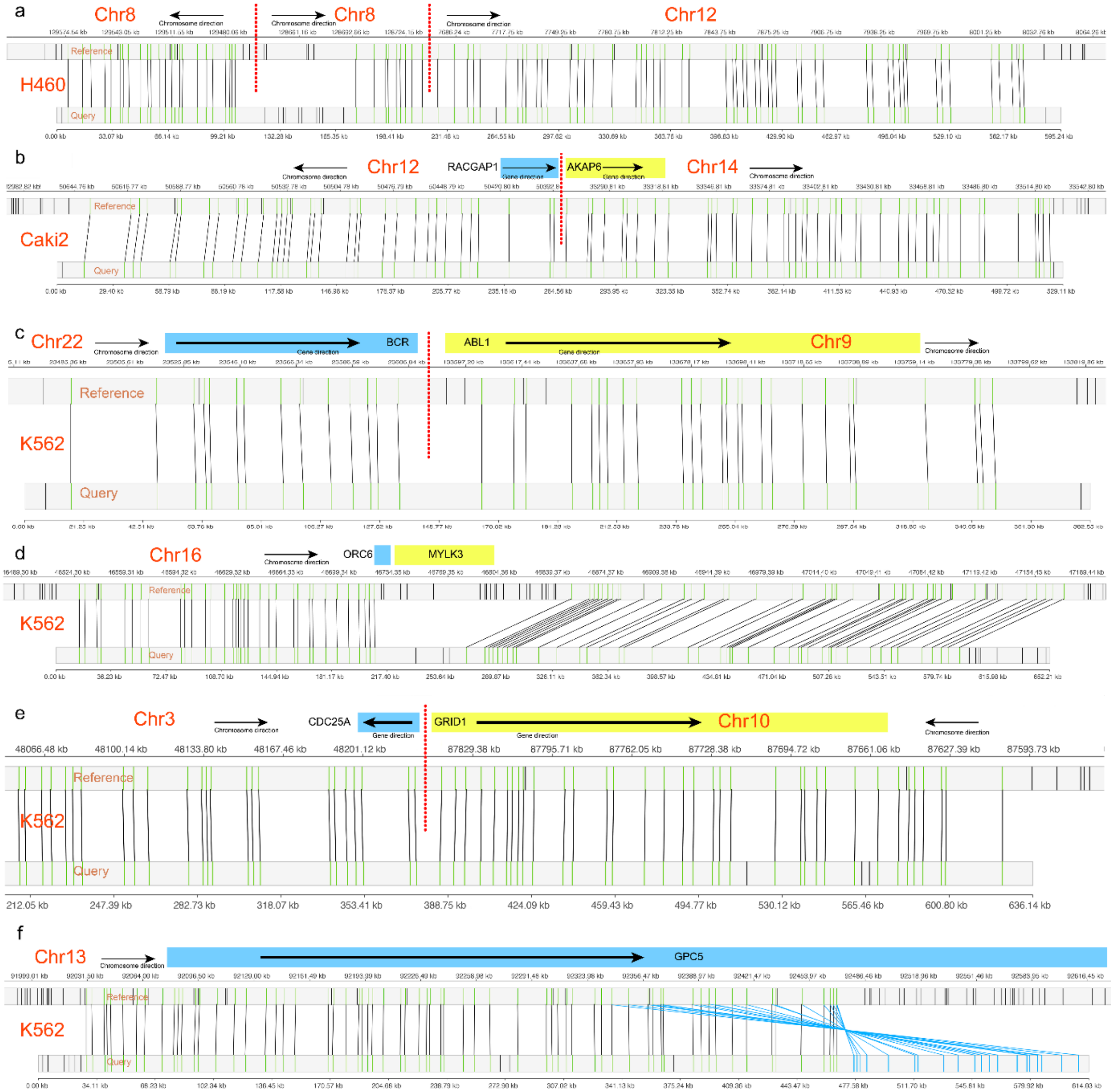


Figure 2.3: Examples of detected structural variants in cancer cell-lines. **a,** A chr8-chr12 translocation shows the integration of a MYC-carrying ecDNA molecule onto chr12 in H460. **b,** A RACGAP1-AKAP6 fusion on CAKI-2. **c,** the BCR-ABL1 fusion on K562. **d,** Deletion of the genes ORC6 and MYLK3 with a partial inversion. **e,** A translocation that disrupts CDC25A and GRID1 but the direction is inconsistent with a fusion event. **f,** A foldback inversion that duplicates and inverts GPC5 in K562.

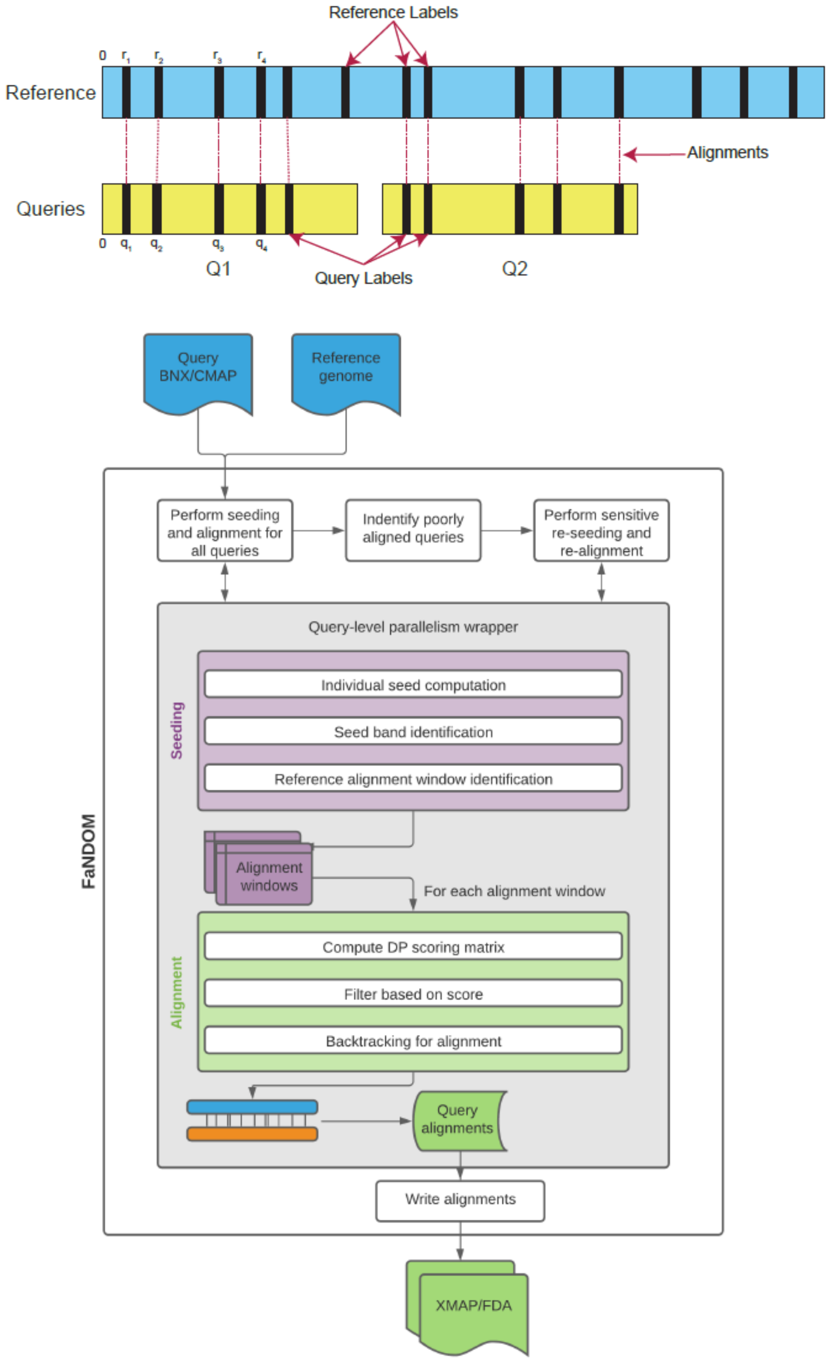

Reference Labels
0
r_1
r_2
r_3
r_4
Reference
Alignments
Queries
0
q_1
q_2
q_3
q_4
Q1
Query Labels
Q2
Query BNX/CMAP
Reference genome
FaNDOM
Perform seeding and alignment for all queries
Indentify poorly aligned queries
Perform sensitive re-seeding and re-alignment
Query-level parallelism wrapper
Seeding
Individual seed computation
Seed band identification
Reference alignment window identification
Alignment windows
For each alignment window
Alignment
Compute DP scoring matrix
Filter based on score
Backtracking for alignment
Query alignments
Write alignments
XMAP/FDA

Figure 2.4: Overview of OM alignments and FaNDOM software. (top) Cartoon diagram of optical map queries aligned to *in silico* reference map. $r_1, r_2, ...$ and $q_1, q_2, ...$ all represent numeric values in base-pair units of the expected or measured locations of labels in the reference and query, respectively. (b) Overview of FaNDOM software. The seeding and alignment modules are called by each parallel thread to produce and store alignments of query to reference.

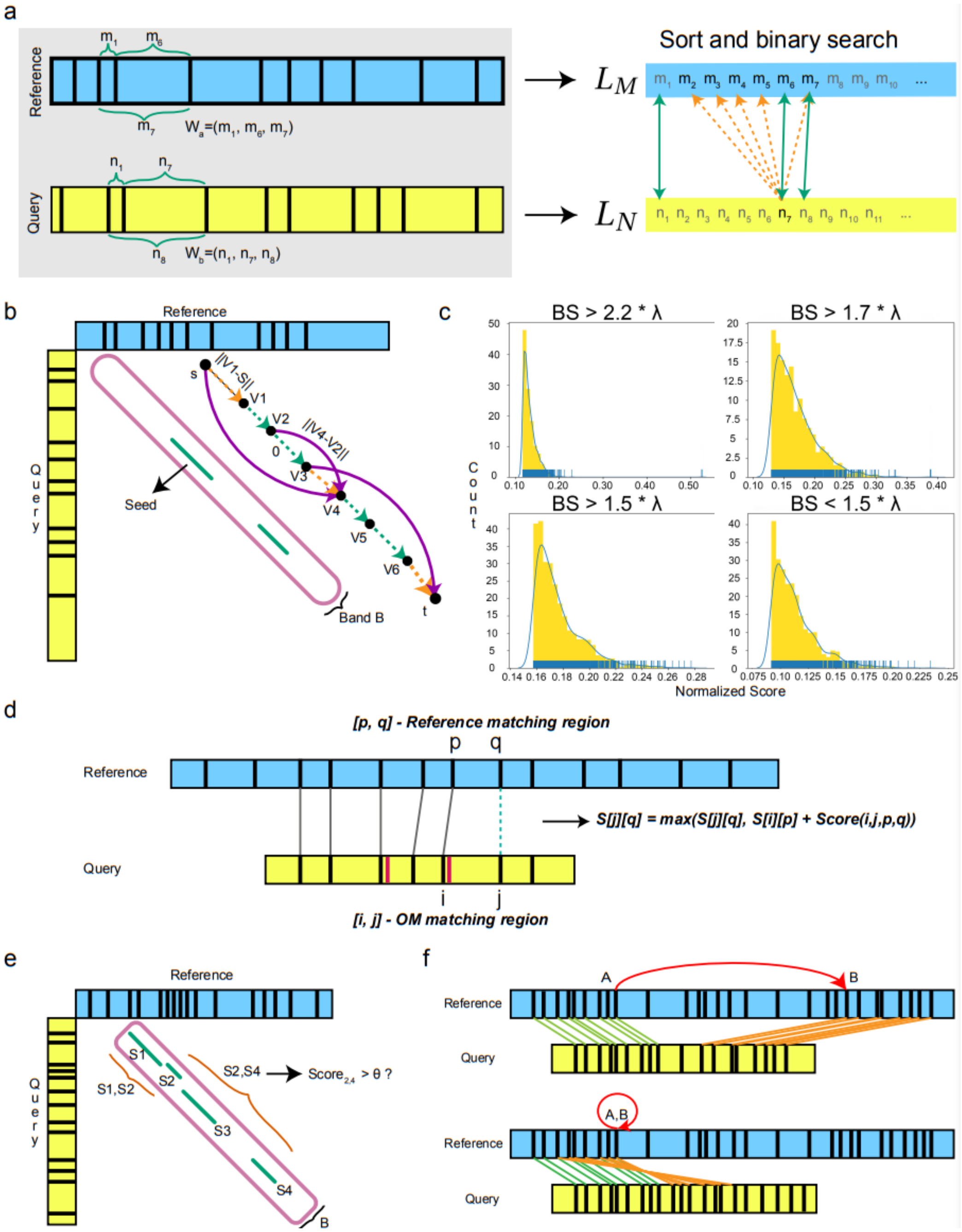


Figure 2.5: The FaNDOM workflow. **a,** Search-and-merge filtering step in which genomic distances extracted from windows (W_a, W_b) and added to lists L_M and L_N. The lists L_M and L_N are merged and seeds are identified. **b,** Packing of seeds into bands, in which for each band *B* seeds inside it are formed into a DAG, *G* and the band is scored by finding the shortest path from *s* to *t*. **c,** Different score threshold possibilities for the band score distribution of bands for a single query. The best score is denoted as "BS". **d,** Dynamic programming for the alignment module in FaNDOM. **e,** Seed selection for partial alignment, which scores bands based on the shortest path between each pair of seeds inside the band *B*. **f,** SV detection module which finds breakpoints based on multiple partial alignments. The alignment on top shows a breakpoint from *A* to *B*, the lower alignment visualizes an inversion, or "foldback".

2.7 References

1. T. S. Anantharaman, B. Mishra, D. C. Schwartz. Genomics via optical mapping. II: Ordered restriction maps. J Comput Biol 4, 91–118. (1997).

2. H. Barseghyan, W. Tang, R. T. Wang, M. Almalvez, E. Segura, M. S. Bramble, A. Lipson, E. D. Douine, H. Lee, E. C. Délot, S. F. Nelson, E. Vilain, Next-generation mapping: a novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. Genome Med. 9 (2017), doi:10.1186/s13073-017-0479-0.

3. D. Botstein, R. L. White, M. Skolnick, R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32, 314–331. (1980).

4. J. Burgin, C. Molitor, F. Mohareb. MapOptics: a light-weight, cross-platform visualization tool for optical mapping alignment. Bioinformatics 35, 2671–2673. (2019).

5. M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. L. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. R. Hastie, D. Antaki, T. Anantharaman, P. A. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C. S. Chin, Z. Chong, N. T. Chuang, C. C. Lambert, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, D. U. Gorkin, M. Gujral, V. Guryev, W. H. Heaton, J. Korlach, S. Kumar, J. Y. Kwon, E. T. Lam, J. E. Lee, J. Lee, W. P. Lee, S. P. Lee, S. Li, P. Marks, K. Viaud-Martinez, S. Meiers, K. M. Munson, F. C. P. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. W. C. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stütz, D. C. J. Spierings, A. Ward, A. M. E. Welch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, E. Lowy, S. Yakneen, S. McCarroll, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P. Y. Kwok, P. M. Lansdorp, G. T. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. E. Talkowski, R. E. Mills, T. Marschall, J. O. Korbel, E. E. Eichler, C. Lee, Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1–16 (2019).

6. E. K. F. Chan, D. L. Cameron, D. C. Petersen, R. J. Lyons, B. F. Baldi, A. T. Papenfuss, D. M. Thomas, V. M. Hayes. Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. Genome Res 28, 726–738. (2018).

7. P. Chen, X. Jing, J. Ren, H. Cao, P. Hao, X. Li. Modelling bionano optical data and simulation study of genome map assembly. Bioinformatics 34, 3966–3974. (2018).

8. Y. Dai, P. Li, Z. Wang, F. Liang, F. Yang, L. Fang, Y. Huang, S. Huang, J. Zhou, D. Wang, L. Cui, K. Wang, Single-molecule optical mapping enables quantitative measurement of D4Z4 repeats in facioscapulohumeral muscular dystrophy (FSHD). J. Med. Genet. 57, 109–120 (2020).

9. J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardımcı, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert, F. Yue, Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).

10. L. M. F. Krause, A. S. Japp, A. Krause, J. Mooster, M. Chopra, M. Muschen, S. K. Bohlander. Identification and characterization of OSTL (RNF217) encoding a RING-IBR-RING protein adjacent to a translocation breakpoint involving ETV6 in childhood ALL. Sci Rep 4, 6565. (2014).

11. J. Gong, Y. Zhou, D. Liu, J. Huo. F-box proteins involved in cancer-associated drug resistance. Oncol Lett 15, 8891–8900. (2018).

12. X. Huang, M. S. Waterman. Dynamic programming algorithms for restriction map comparison. Comput Appl Biosci 8, 511–520. (1992).

13. H. Imaoka, Y. Toiyama, S. Saigusa, M. Kawamura, A. Kawamoto, Y. Okugawa, J. Hiro, K. Tanaka, Y. Inoue, Y. Mohri, M. Kusunoki. RacGAP1 expression, increasing tumor malignant potential, as a predictive biomarker for lymph node metastasis and poor prognosis in colorectal cancer. Carcinogenesis 36, 346–354. (2015).

14. S. H. Kresse, H. O. Ohnstad, E. B. Paulsen, B. Bjerkehagen, K. Szuhai, M. Serra, K. L. Schaefer, O. Myklebost, L. A. Meza-Zepeda. LSAMP, a novel candidate tumor suppressor gene in human osteosarcomas, identified by array comparative genomic hybridization. Genes Chromosomes Cancer 48, 679–693. (2009).

15. E. T. Lam, A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, P. Y. Kwok. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 30, 771–776. (2012).

16. M. Leinonen, L. Salmela. Optical map guided genome assembly. BMC Bioinformatics 21, 285. (2020).

17. A. K. Y. Leung, N. Jin. K. Y. Yip, T. F. Chan. OMTools: a software package for visualizing and processing optical mapping data. Bioinformatics 33, 2933–2935. doi:10.1093/bioinformatics/btx317. (2017).

18. A. K. Y. Leung, T. P. Kwok, R. Wan, M. Xiao, P. Y. Kwok, K. Y. Yip, T. F. Chan. OMBlast: alignment tool for optical mapping using a seed-and-extend approach. Bioinformatics 33, 311–319. doi:10.1093/bioinformatics/btw620. (2016).

19. L. Li, A. K.-Y. Leung, T.-P. Kwok, Y. Y. Y. Lai, I. K. Pang, G. T.-Y. Chung, A. C. Y. Mak, A. Poon, C. Chu, M. Li, J. J. K. Wu, E. T. Lam, H. Cao, C. Lin, J. Sibert, S.-M. Yiu, M.

Xiao, K.-W. Lo, P.-Y. Kwok, T.-F. Chan, K. Y. Yip, OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. Genome Biol. 18, 230 (2017).

20. N. Li, K. Shi, W. Li. TUSC7: A novel tumor suppressor long non-coding RNA in human cancers. J Cell Physiol 233, 6401–6407. (2018).

21. C. X. Liu, Y. Li, L. M. Obermoeller-McCormick, A. L. Schwartz, G. Bu. The putative tumor suppressor LRP1B, a novel member of the low density lipoprotein (LDL) receptor family, exhibits both overlapping and distinct properties with the LDL receptor-related protein. J Biol Chem 276, 28889–28896. (2001).

22. 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. Nat. Commun. 11, 1–14 (2020).

23. J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, S. W. Scherer. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42, D986–992. (2014).

24. M. Matejcic, D. Li, N. J. Prescott, C. M. Lewis, C. G. Mathew, M. I. Parker. Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study. PLoS One 6, e29366. (2011).

25. L. M. Mendelowitz, D. C. Schwartz, M. Pop. Maligner: a fast ordered restriction map aligner. Bioinformatics 32, 1016–1022. (2016).

26. M. Muggli, S. J. Puglisi, C. Boucher. Efficient Indexed Alignment of Contigs to Optical Maps , 68–81. (2014).

27. M. D. Muggli, S. J. Puglisi, C. Boucher. Kohdista: an efficient method to index and query possible Rmap alignments. Algorithms Mol Biol 14, 25. (2019).

28. N. Nagarajan. T. D. Read, M. Pop. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics 24, 1229–1235. (2008).

29. K. Neveling, T. Mantere, S. Vermeulen, M. Oorsprong, R. van Beek, E. Kater-Baats, M. Pauper, G. van der Zande, D. Smeets, D. O. Weghuis, M. J. P. L. Stevens-Kroef, A. Hoischen, Next-generation cytogenetics: Comprehensive assessment of 52 hematological malignancy genomes by optical genome mapping. Am. J. Hum. Genet. 108, 1423–1435 (2021).

30. W. Pan, T. Jiang, S. Lonardi. OMGS: Optical Map-Based Genome Scaffolding. J Comput Biol 27, 519–533. (2020).

31. H. Parikh, M. Mohiyuddin, H. Y. Lam, H. Iyer, D. Chen, M. Pratt, G. Bartha, N. Spies, W. Losert, J. M. Zook, M. Salit, M. svclassify: a method to establish benchmark structural variant calls. BMC Genomics 17, 64. (2016).

32. W. F. Reinhart, J. G. Reifenberger, D. Gupta, A. Muralidhar, J. Sheats, H. Cao, K. D. Dorfman. Distribution of distances between dna barcode labels in nanochannels close to the persistence length. The Journal of chemical physics 142, 064902. (2015).

33. D. C. Schwartz, X. Li, L. I. Hernandez, S. P. Ramnarain, E. J. Huff, Y. K. Wang. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science 262, 110–114. (1993).

34. N. Sengupta, C. Yau, A. Sakthianandeswaren, D. Mouradov, P. Gibbs, N. Suraweera, J. B. Cazier, G. Polanco-Echeverry, A. Ghosh, M. Thaha, S. Ahmed, R. Feakins, D. Propper, S. Dorudi, O. Sieber, A. Silver, C. Lai, Analysis of colorectal cancers in British Bangladeshi identifies early onset, frequent mucinous histotype and a high prevalence of RBFOX1 deletion. Mol. Cancer. 12 (2013), doi:10.1186/1476-4598-12-1.

35. J. M. Shelton, M. C. Coleman, N. Herndon, N. Lu, E. T. Lam, T. Anantharaman, P. Sheth, S. J. Brown. Tools and pipelines for bionano data: molecule assembly pipeline and fasta super scaffolding tool. BMC genomics 16, 734. (2015).

36. H. Shi, Z. Wang. Atypical microdeletion in 22q11 deletion syndrome reveals new candidate causative genes: A case report and literature review. Medicine (Baltimore) 97, e9936. (2018).

37. B. Teague, M. S. Waterman, S. Goldstein, K. Potamousis, S. Zhou, S. Reslewic, D. Sarkar, A. Valouev, C. Churas, J. M. Kidd, S. Kohn, R. Runnheim, C. Lamers, D. Forrest, M. A. Newton, E. E. Eichler, M. Kent-First, U. Surti, M. Livny, D. C. Schwartz, High-resolution human genome structure by single-molecule analysis. Proc. Natl. Acad. Sci. U. S. A. 107, 10848–10853 (2010).

38. 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, P. S. Mischel. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 543(7643):122-5 doi 10.1038/nature21356. (2017).

39. A. Valouev, L. Li, Y.-C. Liu, D. C. Schwartz, Y. Yang, Y. Zhang, M. S. Waterman, Alignment of Optical Maps. J. Comput. Biol. 13, 442–462 (2006).

40. A. Valouev, D. C. Schwartz, S. Zhou, M. S. Waterman. An algorithm for assembly of ordered restriction maps from single DNA molecules. Proc Natl Acad Sci USA 103, 15770–15775. (2006).

41. C. E. Waters, J. C. Saldivar, S. A. Hosseini, K. Huebner. The FHIT gene product: tumor suppressor and genome “caretaker”. Cell Mol Life Sci 71, 4577–4587. (2014).

42. G. Wu, Y. Yan, X. Wang, X. Ren, X. Chen, S. Zeng, J. Wei, L. Qian, X. Yang, C. Ou, W. Lin, Z. Gong, J. Zhou, Z. Xu, CFHR1 is a potentially downregulated gene in lung adenocarcinoma. Mol. Med. Rep. 20, 3642–3648 (2019).

43. K. Yoshihara, Q. Wang, W. Torres-Garcia, S. Zheng, R. Vegesna, H. Kim, R. G. Verhaak. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene 34, 4845–4854. (2015).

44. Y. Yuan, C. Y. Chung, T. F. Chan. Advances in optical mapping for genomic research. Comput Struct Biotechnol J 18, 2051–2062. (2020).

45. B. Zhou, S. S. Ho, S. U. Greer, X. Zhu, J. M. Bell, J. G. Arthur, N. Spies, X. Zhang, S. Byeon, R. Pattni, N. Ben-Efraim, M. S. Haney, R. R. Haraksingh, G. Song, H. P. Ji, D. Perrin, W. H. Wong, A. Abyzov, A. E. Urban, Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. Genome Res. 29, 472–484 (2019).

46. S. Zhou, F. Wei, J. Nguyen, M. Bechner, K. Potamousis, S. Goldstein, L. Pape, M. R. Mehan, C. Churas, S. Pasternak, D. K. Forrest, R. Wise, D. Ware, R. A. Wing, M. S. Waterman, M. Livny, D. C. Schwartz, A single molecule scaffold for the maize genome. PLoS Genet. 5 (2009), doi:10.1371/JOURNAL.PGEN.1000711.

CHAPTER 3: Extrachromosomal DNA in HPV mediated oropharyngeal cancer drives diverse oncogene transcription

3.1 Introduction

Oropharynx cancer has become the second-fastest growing cause of cancer death and the third-fastest growing in frequency among solid organ cancers in the U.S (1,2). The main histology is oropharynx squamous cell carcinoma (OPSCC) which is driven by high-risk human papillomavirus (HPV) type 16 (3,4). The annual number of HPV-related oropharynx carcinoma (HPVOPC) cases has already surpassed the number of cervical cancer cases in the US in 2009, and by 2030 approximately half of all head and neck cancers in the US are predicted to be HPV-related (3). Although HPVOPC exhibits an improved clinical prognosis compared to HPV-negative OPSCC, 20-35% of tumors exhibit an aggressive course despite multimodality therapy (5). A major hurdle to understanding HPV-mediated oncogenesis is an incomplete understanding of the role of viral and viral-human hybrid transcripts and viral-human DNA integration.

Extrachromosomal DNA (ecDNA) has recently been shown to play a critical role in human cancer (6-9). Because of its non-chromosomal mechanism of inheritance, ecDNA can drive high copy number while promoting intratumoral heterogeneity, promoting accelerated tumor evolution and drug resistance (6,10). Moreover, chromatin rewiring on ecDNA allows for higher accessibility and increased expression of oncogenes (6,7,11). More recent reports have conjectured that hybrid human-virus ecDNA formation could be a possible mechanism for increased copy number of the HPV oncogenes E6, E7 (12-16). Our previous studies have demonstrated that the HPVOPC cell line UPCI:SCC090 features hybrid human-viral circular ecDNA containing FOXE1 and HPV-16 through conventional and long-read whole genome sequencing, and which we verified *in vitro* using fluorescent in situ hybridization (FISH) (11).

Given these data, we hypothesized that the genetic structure and viral gene expression in primary HPVOPC as well as the expression of human viral hybrid transcripts may be related to ecDNA. We combined whole genome sequencing, conventional RNA-seq and long-read RNA-seq to analyze HPV and human viral hybrid genomic and transcriptomic structure in the context of HPVOPC (17). Analysis of ecDNA and associated transcript structure clarified HPV transcript structure and the role of viral, human, and hybrid ecDNA in enhancing expression of diverse and oncogenic viral, human, and hybrid transcripts with functional validation.

3.2 Methods

3.2.1 Patient samples

Forty-four primary tumor tissue samples were obtained from a cohort of HPV-positive OPSCC patients from the Johns Hopkins Tissue Core (institutional review board protocol #NA_00-36235) and Moores Cancer Center Biorepository and Tissue technology shared at University of California, San Diego Human Research Protections Program (institutional review board approved protocol HRPP# 181755). Pathology of the primary tumors confirmed by two independent pathologists and tumor tissue was microdissected to yield at least 80% tumor purity. HPV tumor status was determined by *in situ* hybridization or p16 immunohistochemistry. In equivocal cases, HPV-16 E6 and E7 viral oncoproteins were detected via PCR for confirmation. Whole genome sequencing using paired-end Illumina sequencing along with conventional RNA-seq was acquired for 37 samples.

3.2.2 Whole genome sequencing

DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen) for high-quality extraction per the manufacturer instructions. DNA samples from tumor were quantified using a Qubit (ThermoFisher Scientific). Greater than 1ug of each sample was prepared using a sonication-based library construction and enrichment method per the Beijing Genomics Institute (BGI) as previously described (18).

DNA was isolated from 0.35 mm thick frozen tissue cuts digested in 1% SDS (Sigma-Aldrich, St. Louis, MO) and 50 µg/ml proteinase K (Invitrogen, Carlsbad, CA) solution at 48°C for 48 hours. The DNA was purified by phenol-chloroform extraction and ethanol precipitation. DNA was resuspended in LoTE buffer, and the DNA concentration was quantified using the NanoDrop spectrophotometer. Sequencing was performed with the Illumina Hiseq Xten 151PE strategy with 350bp insert library. The pipeline steps included preparation of HPV reference genome file, performance of quality control on BAM files, extraction of unmapped read pairs, conversion of unmapped read pairs to FASTQ format, alignment of unmapped read pairs to the HPV reference genomes (accession number: AY686584.1).

3.2.3 RNA preparation

Frozen tissue specimens were cut into 0.35-mm thick sections and RNA was extracted according to the Qiagen RNeasy Plus Mini Kit (Qiagen, Hilden, Germany). RNA concentration was verified using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA). The absorbance ratio of 260 nm to 280 nm was used to verify adequate quality, defined as > 1.8. An RNA Integrity Number (RIN) of 7.0 or greater was required for quality assessment. RNA from all eleven tumors passed quality assessment.

3.2.4 cDNA library preparation, long read RNAseq, and alignment to HPV16 genome

Briefly, the RNA was extracted from 0.35 mm thick frozen tissue sections and a stranded RNA library was prepared using the Illumina TruSeq stranded total RNA seq poly A+ Gold kit (San Diego, CA) following the manufacturer’s recommendations. Long-read RNAseq of full-length transcripts was performed on 2 non-integrated and 3 integrated tumors according to the PacBio Iso-Seq pipeline (Menlo Park, CA). Briefly, 500 ng of purified RNA was used to prepare cDNA was using the Clontech SMARTer PCR cDNA synthesis kit (Mountain View, CA) and cDNA was then repaired. Large-scale PCR was performed using the Blue-Pippin size selection system for three sized cDNA libraries (<1.5kb, 1.5 – 2.5 kb, >2.5 kb). SMRTbell templates were then purified and sequenced on the PacBio SMRT Sequencing platform. The general SMARTer IIA oligonucleotide was used to anneal to the polyA tail of transcripts during cDNA sample preparation. Junction-spanning reads covered by fewer than 5 reads were dropped from analysis.

The Spliced Transcripts Alignment to a Reference (STAR) software was used to align long-read RNA seq reads to the HPV16 reference genome (GenBank: AY686584.1) (19). Full length transcripts were visualized with IGV for confirmation, and erroneously mapping transcripts were removed from analysis.

3.2.5 Short read RNA-seq alignment and analysis

Standard (short read) RNA-seq was performed as previously described.(20) A ribosomal RNA reduction was performed and the purified RNA was fragmented, then converted to double stranded cDNA, and the cDNA was 3′ adenylated and ligated with barcode adapters. The library was then enriched using PCR and AMPure XP bead purification. Sequencing was then performed using the HiSeq 2500 platform sequencer (Illumina), and the TruSeq Cluster Kit for 2×100 bp sequencing. The reads were trimmed to remove adapter sequences and low-quality reads using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

The RNA sequences were aligned to the HPV16 genome (GenBank: AY686584.1) and hg19 assembly using MapSplice2 version 2.0.1.9. Integration and expression of HPV genes were identified by taking reads in RNA-seq data aligned to a combined database of human reference genome and high-risk HPV16, HPV33, HPV35 reference genomes using MapSplice (https://github.com/favorov/viruses-in-sequencing) (21). MapSplice was run with the default command line arguments. RNA-seq reads were aligned to the HPV16 genome and reads spanning canonical HPV16 splice junctions were extracted. For splicing analysis, RNA-seq reads were normalized by dividing by the total number of junction-spanning reads in the sample. Junction-spanning reads were discarded if the junction constituted < 1% of all junctions in the sample.

Quantification of RNA-seq expression was performed following the HISAT, StringTie, and Ballgown pipeline (22). Briefly, HISAT2 was used to align the RNA-seq reads to the hg19+HPV reference genomes. StringTie was run on each individual alignment to identify the assembled transcripts within the sample. All identified transcripts across all samples were merged using StringTie to create a consistent set of reference transcripts across the entire dataset. Abundances for each transcript for each sample was re-estimated using StringTie, and Ballgown was run on the resulting output to obtain read counts, coverage, and expression data across all samples.

3.2.6 Detection of focal amplifications with AmpliconArchitect (AA)

WGS reads were aligned to hg19 and 337 viral genomes using BWA-MEM (23). AA seed detection was performed with CNVKit (24). Copy number amplification regions matching low complexity, repetitive or poorly mappable genomic regions were filtered using the AA database. Some unfiltered regions corresponding to repetitive genomic regions still existed after this step and were shared across multiple samples. We removed any such regions existing in 10% or more of samples. Remaining copy number seed intervals larger than 10 kbp and with estimated CN > 4.3 were used as input to AA. Resulting amplicons generated by AA were examined for the presence of breakpoint graph cycles containing solely amplified human DNA (human ecDNA) and viral breakpoint edges linking HPV to an amplified cyclic human DNA structure (human-viral, ‘hybrid’, ecDNA).

3.2.7 Integrative Genome Viewer confirmation

Putative full-length transcripts and splice junctions were visualized using Integrative Genome Viewer (IGV/ Broad Institute, version 2.4.5) (25). From long-read and short-read RNA sequencing data, BAM files were loaded into IGV and visualized at the start and end of each junction. Long-read isoforms from IGV were individually verified and isoforms with mapping error were removed from analysis.

3.2.8 Integration, fusion, and splicing analysis

WGS and RNA-seq reads were aligned to hg19 and viral genomes using BWA-MEM. ViFi was run on each aligned BAM file to detect viral integration and transcription fusion location. Several transcription fusion events did not have a proximal viral integration event. Closer inspection of these fusion transcripts revealed that they had much lower support (mean 74 RNA-seq reads supporting the fusion event) compared to fusion transcripts that were proximal to a viral integration event (mean 955 RNA-seq reads supporting the fusion event). As such, we removed all fusion transcripts without supporting genomic integration.

Samples were classified according to the presence of viral integration and transcript fusion events. Samples containing a viral integration were classified as Hybrid-DNA. Samples that also contained a fusion transcript were classified as Hybrid-RNA.

In-depth splicing analysis was performed by taking the reads aligned to the HPV16 reference genome and counting the number of splice events detected by the alignment denoted by HPV16 donor and acceptor site pair (i.e., SD_x-SA_y). For the cases in which the splice junction started with an HPV16 donor site and ended in the human genome, it was denoted by the HPV16 donor site to human splice event (i.e., SD_x-Human). From this data, we generated a splicing matrix where each row is a sample and each column is a splicing event, and the entries in the matrix are the total number of times that splicing event was observed within the sample. We performed principal component analysis on the splicing matrix in order to examine how the samples clustered.

3.2.9 Unsupervised hierarchical clustering and splicing analysis

Conventional RNA-seq reads spanning canonical HPV16 splice donor and acceptor sites (SD226, SA409, SA526, SA742, SD880, SD1302, SA2582, SA2709, SA3358, SD3632, SA5639) from the Institutional and TCGA cohorts were extracted and normalized to the number of reads specific to that tumor. Tumors with fewer than 500 mapped RNA-seq reads were excluded from inclusion in the heatmap (T14 – 241, T12 – 4, T30 – 2). RNAseq reads spanning HPV16 splice donor to human acceptor sites were also extracted and normalized to the number of reads specific to that tumor. To calculate the proportion of E6 protein reads with truncation to E6*I in each tumor, the following formula was applied to each tumor:

$$proportion\ E6 * I = \frac{nRNAseq\ reads\ SD226\_SA409}{nRNAseq\ reads\ SD226\_SA409 + nRNAseq\ reads\ SD226\_nt227}$$

We also quantified the relative frequency of HPV16 SD880 to human splicing events with the following:

$$proportion\, SD880human = \frac{nRNAseq\ reads\ SD880\_human}{nRNAseq\ reads\ in\ sample}$$

3.2.10 Insertion of HPV16 into Human Genome Analysis

Integration breakpoints and intragenomic viral breakpoints were identified with ViFi and AA. For DNA-based breakpoints both ViFi and AA were used to identify integration sites. To detect intragenomic viral breakpoints, AA alone was used. For RNA-seq data ViFi and MapSplice2 were used to identify splicing and human-viral chimeric sequences.

3.2.11 Analysis of non-canonical HPV16 structures

Canonical and non-canonical circular viral genome structure status was determined by AmpliconArchitect analysis. Tumor samples which did not have hybrid ecDNA were classified as non-canonical if they contained a cyclic AA graph decomposition and > 100 bp of rearranged genomic content (including indels), while canonical circular status was assigned if no such large rearrangements were present and cyclic AA graph decomposition of virus was present.

3.2.12 Quantification of splice acceptor cluster ranges for hybrid splicing events

For measurement of splice cluster ranges, splice clusters were defined via k-means clustering. Any group of donor, sample, and chromosome with fewer than 10 samples was excluded from consideration. Clustering was repeated 30 times at a given threshold to account for random seeding, with the optimal clustering at a given k threshold determined via silhouette score. The number of clusters was increased until a loss in performance was observed, and the number of clusters was confirmed by visual inspection. Clusters containing fewer than five observations were then filtered out, and the range of splicing sites was then computed on each remaining cluster.

Hybrid RNAseq reads from both Institutional and TCGA cohorts were extracted and mapped to specific viral splice donor and human acceptor sites for each tumor. The location of the human acceptor (chromosome and nucleotide) for each read was then determined using split reads, which were identified as having a primary alignment to HPV and a secondary alignment to the human genome. Histograms were created to map the distribution of splice acceptor sites for a given HPV16 donor (i.e. SD226, SD880, etc.) across the human genome for each tumor. Samples with viral read counts < 10 were removed from analysis.

3.2.13 Functional Studies

Proliferation of HCT116 and NOKSI cells was investigated in the presence of empty vector (negative control), as well as E6E7 (positive controls), and parent/daughter constructs. Cells were seeded in 96 well plates at a density of 3,000 cells/well for NOKSI and 4,000 cells/well for HCT116. Individual vectors composed of daughter constructs were then transfected by X-tremeGENE9 (Roche). Proliferation was measured as a ratio of relative absorbance two days after transfection vs. day of transfection.

The effect of FOXE1 siRNA was investigated on cell line SCC090. Cells were seeded at a density of 2,000 cells/well. SiFOXE1 (Santa Cruz Biotechnology) was added at a concentration of 10nM. Percent viability was measured using % viability = (absorbance of siRNA)/(absorbance of vehicle) x 100. For proliferation experiments each datapoint is the average of five replicates with standard error represented by error bars, and all experiments were repeated at least three times demonstrating consistent results.

HCT116 and SCC090 cells were obtained from ATCC, NOKSI was provided as a gift by the Silvio Gutkind Lab (University of California, San Diego, Department of Pharmacology. Cell lines were used for between 4 to 20 passages after thawing from frozen stock. Mycoplasma testing was conducted monthly using the MycoAlert-Plus Mycoplasma Detection Kit (Lonza).

3.3 Results

Forty-four primary tumors were acquired from a cohort of HPV-positive OPSCC patients. HPV tumor status was determined by *in situ* hybridization or p16 immunohistochemistry. In equivocal cases, HPV-16 E6 and E7 viral DNA were detected via PCR for confirmation. Forty of 44 samples were HPV16-positive, three were HPV33-positive, and one was HPV35-positive. Whole genome sequencing (WGS) using paired-end Illumina reads at mean coverage of 30x along with RNA-seq was acquired for 38 HPV16 samples (20), and long-read RNA-seq of full-length transcripts was generated for 5 samples using PacBio Iso-Seq technology (26). Twenty-eight additional HPVOPC tumors from the Cancer Genome Atlas (TCGA) project were analyzed as a separate validation cohort. WGS and RNA-seq data were mapped to the hg19 reference and analyzed using ViFi (27).

3.3.1 EcDNA that carry oncogenes are common in HPVOPC

As we had previously demonstrated the presence of ecDNA in an HPVOPC cell line (11), we hypothesized that ecDNA may be present in primary HPVOPC. A recently developed method, Amplicon Architect (AA) analyzes whole genome sequences to predict ecDNA with

85% precision and 83% sensitivity, as well as reconstruct the fine structure of the amplicons (8). We applied AA to the 28 samples from the HPVOPC cohort (Figure 3.1) (11). Remarkably, we found six hybrid viral-human ecDNA, and another six with human-only ecDNA, (one tumor exhibited both hybrid and human ecDNA; T14 - Figure 3.2a) in our institutional cohort. Additionally, eighteen tumors contained only HPV viral circular DNA (vcDNA). One tumor was not classified due to low viral copy number (T26). HPV vcDNA was present in an intact form including a complete, non-rearranged (canonical) form in 16 tumors. Interestingly, another 15 tumors contained a non-canonical truncated vcDNA with deletions mostly in the L1 and L2 region (Figure 3.3) suggesting that a significant fraction of HPVOPC tumors contain HPV genomes which have undergone substantial genomic rearrangement prior to enrichment in copy number.

To test that the prediction of ecDNA in HNSC samples was not specific to the institutional cohort, we analyzed WGS samples from the HNSC data in the Cancer Genome Atlas (TCGA). Ten samples contained viral-human hybrid ecDNA and 8 contained human-only ecDNA (four tumors exhibited both; Figure 3.2b). VcDNA was also prevalent, with two tumors containing canonical vcDNA while 13 other tumors contained non-canonical truncated HPV vcDNA. One tumor was not classified due to low viral copy number (CV-7406).

EcDNA do not carry centromeres and therefore segregate independently, allowing tumors to rapidly modulate copy numbers of genes on ecDNA, specifically when the genes provide a growth or proliferative advantage (28). Consistent with this hypothesis, seven of the ecDNA+ tumors (both hybrid and human-only) in the institutional cohort carried oncogenic protein-coding genes or ncRNA, including many known oncogenes: EGFR, SEC61G, VOPP1, VSTM2A on chr7 in T1; DUSP4 and KIF138 on chr8 as also CD93 in T29; and, CST1 and

THBD on chr20 in T14. The TCGA cohort revealed similar findings, with protein-coding genes found in five hybrid ecDNA-carrying samples, and three human only ecDNA samples. The structures were largely sample specific. However, we did observe two samples with ecDNA segments on chromosome 11 carrying six oncogenes (ANO1, CTTN, FADD, MIR548K, PPFIA1, SHANK2). Tumor TCGA-CQ-5323 contained an ecDNA with 13 cancer-associated genes, including ANO1, CCND1, CPT1A, CTTN, TRPC4AP, FADD, IGHMP2, MIR548K, MRPL21, ORAOV1, PPFIA1, and SHANK2. In TCGA-CV-5443, a hybrid ecDNA amplicon containing and amplifying the immune regulating ligand PDL1 was identified (Figure 3.2d). PDL1 amplification occurs in a subset of lung, kidney, bladder, and head and neck cancers. The PD1 checkpoint is the most common immunotherapeutic target in solid tumors currently and PD1/PDL1 directed immunotherapy has gained FDA approval for first-line treatment of unresectable or metastatic head and neck cancer (29).

Overlaying RNA -seq data on to the hybrid ecDNA structure in the institutional cohort showed hybrid RNA combining HPV-16 E6, E7, E1, E4, L1, and L2 with a multitude of human sequences containing genes EGFL7, TBCD (chr17; T1), SOX2-OT (chr3; T19), TTC33 (chr5; T41 see Figure 3.2c, PVT1 chr8; T14), LINC01363 (chr1; T47), and TBC1D16 (chr17; T49), As one interesting example, we identified an ecDNA in T41 amplicon that connected viral promoter to multiple exons in TTC33 (Figure 3.2c). Hybrid splicing was additionally confirmed using long-read Iso-Seq data (Figure 3.2c) consistent with the circular hybrid ecDNA structure. The TTC33 gene (tetraticopeptide repeat domain 33) has been implicated as an mRNA chimera in breast, ovarian, stomach, colon, kidney, and uterine cancer (30). These interesting patterns suggested a possible rewiring of the regulatory circuitry in hybrid ecDNA.

3.3.2 Hybrid ecDNA is associated with increased human gene expression.

Prior data have shown that ecDNA mediate increased expression of oncogenic human transcripts contained within ecDNA structure. To examine the effect of viral DNA genomic integration and viral integration in ecDNA, we examined expression of both viral and human transcripts in these contexts. For each gene on an ecDNA, we computed the ratio of its expression (in FPKM units) in the target sample to its mean expression value in all samples where the gene was not on an ecDNA (Methods) and called it the 'FPKM-ratio'. Genes associated with the 38 ecDNA amplicons in the institutional cohort and the TCGA cohort were upregulated nearly 150X, with a mean FPKM-ratio of 149.8 (SD 1,015); median 4.26 (IQR 1.67– 8.92) (Figure 3.4a-b. Thirty-three of 38 (86.8%) of these genes were oncogenes or associated with oncogenic phenotypes. For example, oncogene TNFSF4 on chromosome 1 in TCGA-CR-6473 was upregulated to FPKM-ratio of 116.58. TNFSF4 has been reported to be upregulated in brain metastases (31). EGFR on chromosome 7 in T48 was also upregulated with FPKM-ratio 30.7. PVT1 transcripts are found at a high level in lung cancer and contribute to VEGFC expression (32). CD274/PDL1, associated with immune checkpoint activation, was one of the most upregulated genes by tumor TCGA-CV-5443 in the TCGA cohort with FPKM-ratio 24.4.

Importantly, human genes associated with hybrid transcriptomes showed increased expression for both the institutional and TCGA cohorts (Figure 3.4c). The increased expression was most pronounced in tumors with hybrid transcripts, but also increased expression was noted for all genes located on hybrid ecDNA. To further explore this relationship, we assembled and annotated human reads overlapping with human viral splice junctions, and spatially defined expression along genomic fusions. We noted that strand-specific expression of human genes downstream of viral sequences in fusion DNA structures could exceed 30-fold compared to surrounding genes. For example, T41 shows dramatic increase in TTC33 expression downstream of HPV-human hybrid sequence (Figure 3.4d). A similar phenomenon was noted in the context of other hybrid ecDNA structures (Figure 3.5).

3.3.3 Viral transcripts show diverse isoforms in hybrid ecDNA and hybrid transcript expression

Unsupervised hierarchical clustering based on frequency of HPV16 splicing junctions in mapped RNA reads showed a number of distinct patterns. First, we observed that SD226-SA409 represent a significant portion of junction-spanning reads in the HPV16 cohort, occurring in every sample with high frequency 36.3%, range 18.3% to 63.5%, SD 9.7% of junction-spanning reads. The use of this junction creates a shortened form of E6, called E6*I, which results in a premature stop codon (33). The mean fraction of reads demonstrating truncation of E6 to E6*I was 81.1% (SD 12.0%; min 44%.5 max 94.7%) across the cohort of HPV16-positive tumors (see Methods), demonstrating that E6*I is more commonly expressed than full-length E6 across all HPVOPC. Our results are consistent with previous results identifying E6*I in cervical dysplasia, cervical cancer, and HPVOPC (33-37). We calculated the proportion of E6 transcripts that were E6*I in the institutional cohort and found that tumors with either form of ecDNA had reduced E6*I production compared to the non-ecDNA tumors [0.72 (0.15) vs. 0.82 (0.09); p=0.0197 by T-test; mean (SD)]. Human ecDNA tumors had reduced E6*I production compared to the non-human ecDNA tumors [0.69 (0.13) vs. 0.81 (0.11); p=0.0288 by T-test; mean (SD)]. Hybrid ecDNA tumors did not have reduced E6*I production compared to the non-human ecDNA tumors [0.75 (0.18) vs. 0.80 (0.11); p=0.42 by T-test; mean (SD)]. To validate these findings, we examined the TCGA cohort of 28 HPVOPSCC tumors. Similar to the institutional cohort, the majority of E6 was truncated to E6*I (mean 0.89, SD 0.05, median 0.89 IQR 0.87-0.92), although we did not detect a difference in proportion of

E6*I based on ecDNA status. This does confirm that, contrary to the classic model of HPV carcinogenesis, E6*I, rather than E6, is the most common viral transcript in HPVOPC.

Second, although splicing of the 5’ SD880 splice donor site to the 3’ SA3358 site had been described as the most frequent splicing event in HPV-16 cervical cancers and in cell lines (38-40), we found that 10 of 37 tumors (27%) preferentially spliced from SD880 to a human splice acceptor site instead of the canonical SA3358 (Figure 3.2a). In these 10 tumors, 47% (mean; SD=11%) of reads spanning the SD880 splice site spliced to a human locus, and only 3% (mean; SD=3%) of SD880 reads spliced to the canonical SA3358 receptor (p<0.001 by T-test). This questions previous findings that efficient usage of SA3358 is necessary for production of E6, E7, E4, E5, L1, and possibly L2 proteins (38). However, we did not detect preferential splicing of SD880 to a human acceptor to be associated with ecDNA status (40% of any ecDNA tumors preferentially spliced SD880 to human vs. 22% non-ecDNA; p=0.28; 40% of hybrid ecDNA tumors preferentially spliced SD880 to human vs. 25% non-hybrid ecDNA; p=0.482; 40% of human ecDNA tumors preferentially spliced SD880 to human vs. 25% non-ecDNA; p=0.482). We also analyzed splicing patterns of SD880 in TCGA and found that 10/28 (35%) preferentially spliced to a human splice acceptor rather than canonical SA3358. Tumors with either form of ecDNA were significantly more likely to splice to a human acceptor from SD880 (8/14; 57% vs. 2/14; 14%, p=0.018).

We also noted a strong association of splicing patterns depending on hybrid DNA or RNA status in the institutional cohort, irrespective of ecDNA (Figure 3.2a-b). Hybrid-DNA tumors (n=18) exhibited a greater fraction of RNA reads covering SD880-human junctions than non-hybrid-DNA tumors [n=19; 0.26 (0.25) vs. 0.02 (0.01); p<0.001; and fewer RNA reads covering the SD880-SA3358 junction [0.21 (0.21) vs. 0.42 (0.12); p<0.001 by T-test]. Similarly, hybrid-RNA tumors (n=12) exhibited a greater fraction of RNA reads covering SD880-human junctions than non-hybrid-RNA tumors [n=25; 0.38 (0.21) vs. 0.01 (0.05); p<0.001 by T-test] and fewer RNA reads covering the SD880-SA3358 junction [0.09 (0.12) vs. 0.43 (0.12); p<0.001 by T-test]. Principal component analysis of splicing patterns in hybrid RNA tumors also demonstrated a distinct subset based on splicing signature, in which HPV splicing pattern most closely relates to the presence of viral-human hybrid transcripts (Figure 3.6a and Figure 3.7). We observed selective enrichment of E6/E7 regions in both WGS and RNA data, which is pronounced in hybrid RNA tumors compared to non-hybrid RNA tumors (Figure 3.6b), as well as depletion of L2 in hybrid samples.

We defined aggregate splice donors in HPV and hybrid transcripts, and noted that SD226, SD880, and other known splice sites in the early region of HPV16 genome were strongly preserved and limited to a single donor canonical nucleotide. However, splice acceptors in hybrid transcripts showed broad variation. In the institutional cohort, the mean variation of the SD880 splice acceptor was 11,060 nucleotides (SD 37,217), but tighter for SD226 (mean 885, SD 2,290) and SD1302 (mean 48, SD 58). In TCGA, the degree of variation was similar (Figure 3.6c). We also noted that splicing patterns varied depending on hybrid DNA or RNA status in TCGA, irrespective of ecDNA. Sixty-four percent (18/28) exhibited a hybrid genome and 53% (15/28) exhibited hybrid transcriptomes, and hybrid DNA was a prerequisite for hybrid RNA (p<0.001). Similarly, SD226-SA409 represented a significant portion 40.0% (range 19.1 – 73.2%, SD 17.1%) of junction-spanning reads, TCGA tumors also preferentially expressed E6*I compared to full length E6 (89.2% E6*I reads (SD 5.3%)), and hybrid genome tumors (n=18) exhibited significantly higher percentage of splicing events from SD880 to a human locus [11.0 (10.0%) vs. 2.9 (9.0%); mean (SD); p=0.0508] and fewer reads covering the SD880-SA3358 junction [22.2 (23.9%) vs. 50.0 (19.2%); p=0.005]. Hybrid transcriptome tumors (n=15) also exhibited significantly higher fraction of splicing events from SD880 to a human locus [13.2 (10.0%) vs. 2.2 (8.0%); mean (SD); p=0.0041] and fewer reads covering the SD880-SA3358 junction [16.5 (21%) vs. 50.0 (17.4%); p=0.001].

3.3.4 Novel HPV transcript structures related to ecDNA are found in HPVOPC

Long read polyA RNA sequencing provides direct sequencing of full-length transcripts and can avoid artifacts introduced by short read transcript assembly. To provide a more precise understanding of HPV transcripts in HPVOPC, we performed long-read RNA whole genome polyA transcript sequencing on a subset of two tumors without hybrid transcripts (T2 and T38), one human ecDNA tumor with hybrid transcript expression (T19), and one hybrid ecDNA tumor with hybrid transcripts (T45) (Figure 3.8A).

Long read sequencing confirmed a divergent transcript structure of hybrid ecDNA and vcDNA HPV transcripts. For example, T2 and T38 vcDNA (non-hybrid DNA/transcriptome tumors) exhibited 0% of conventional RNA-seq reads covering SD880 spliced to a human junction. By contrast, T19 (human ecDNA / hybrid transcripts) exhibited 19.5% and T45 (hybrid ecDNA/hybrid-RNA tumors) 100% of conventional RNA-seq reads of SD880 to be spliced to a human splice acceptor. Of interest, even though long-read RNA-seq is not quantitative in nature, we observed that both T19, a non-hybrid ecDNA tumor that expressed hybrid transcripts, and T45, a hybrid ecDNA/hybrid transcript tumor, essentially displayed no full-length transcripts that mapped to HPV alone; rather the transcripts were all hybrid. In T19, 36/40,474 (0.1%) long-reads mapped to HPV16 and in hybrid ecDNA tumor T45 17/55,814 (0.03%) long-reads mapped to HPV16. Conversely, the two non-hybrid vcDNA tumors T2 and T38 carried more full-length HPV-only transcripts. In T2, 515/19,220 (2.6%) long reads mapped to HPV and in T38 405/38,026 (1.1%) long reads mapped to HPV; (p<0.001 between non-hybrid and hybrid tumors).

Long-read data confirmed the presence of canonical splicing events seen in conventional RNA sequencing. The most common full-length transcript in non-hybrid tumors was 1,476 nt long, beginning at the p97 promoter with splicing at SD226-SA409 and SD880-SA3358 extending to the early polyA tail, with coding potential for the E6 oncoprotein variant E6*I defined by SD226-SA409, full-length E7, full-length E4, and full-length E5 (Fig 3.8b-c). This transcript was observed in 55% of full length reads mapped to HPV16 in T2 and 65% of full length reads mapped to HPV16 in T38, and included splice-junctions commonly observed in RNA-seq data, including SD226-SA409.

The long-read results and the RNA-seq data from the institutional and TCGA cohorts suggested that the predominant form of E6 in full-length transcripts found in HPVOPC was not the full-length E6 isoform, but truncated version E6*I defined by SD226-SA409. Additional isoforms E6*II, and E6*III(41,42), were also noted in long read transcripts. Full-length E7 was also common and present in the majority of HPV coding transcripts (91%). Finally, E1^E4, which is the result of SD880-SA3358 splicing was observed in 11/23 (48%) of distinct full-length isoforms.

3.3.5 EcDNA Hybrid HPV transcripts are functionally active

We selected a tumor (T41) for functional characterization of transcripts due to presence of fusion RNA reads, in addition to the fusion WGS reads and observation of up to 512x increased expression of segments associated on this ecDNA structure. After confirming the presence of AmpliconArchitect-predicted hybrid junctions using RT-PCR and sequencing

(Figure 3.9), we cloned the entire transcript (E6-E7-E1-TTC33*-E5*) as well as daughter constructs into a pcDNA 3.1(+)-myc-His A vector (Genscript, Inc.) (Figure 3.10a-b), as well as the most common full-length HPV16 in the form of component HPV gene transcripts, into the same backbone. To explore the functional effects of these transcripts, we transfected these constructs into HPV null diploid HCT116 (p53 and Rb wt MMR deficient colorectal carcinoma) cells as well as an HPV null normal oral keratinocyte NOKSI (spontaneously immortalized oral keratinocyte) cell lines that respond to HPV E6/E7 gene expression with enhanced proliferation, to provide an assessment of the effects of transcripts from this primary tumor (43,44). In HCT116 cells, E6*I, E6E7, E6*I/E7, E6*I/E7/E4/E5, E7, TTC33, E6E7E1 and the entire hybrid transcript from T41 induced significant growth compared to the empty vector (p=0.02, 0.02, $6x10^{-3}$, 0.03, $3x10^{-4}$, $2x10^{-4}$, $7x10^{-3}$, 0.01, respectively, Student's *t*-test) (Figure 3.10c). Similarly, in NOKSI cells, E6*I, E6E7, E6E7E1 and the entire intact hybrid transcript from T41 increased proliferation (p=0.02, 0.01, 0.03, and 0.04, respectively, Student's *t*-test) (Figure 3.10d).

We have previously reported on the presence of ecDNA in an HPOPC cell line UPCI;SCC090, demonstrating reconstruction of a complex hybrid structure (>100 kbp) containing the oncogene FOXE1 as well as highly expressing HPV16 sequence, and shown the presence of FOXE1 in both chromosomal HSRs as well as in ecDNA using FISH probes in metaphase imaging (11). To examine the functional contribution of FOXE1 in an ecDNA context, we treated SCC090 cells with siRNA for FOXE1 demonstrating significant inhibition of growth, indicating that overexpression of FOXE1 via ecDNA mediated mechanisms is a major driver of growth in SCC090 cells (Figure 3.10e).

To define the potential for hybrid ecDNA in HPVOPC to drive protein expression related to immune evasion, data derived via reverse phase protein arrays (RPPA) corresponding to head and neck squamous cell carcinoma samples included in TCGA were extracted from The Cancer Proteome Atlas (TCPA) (45). Nine tumor samples identified as being HPV+ in the TCGA cohort had RPPA-derived expression data available in TCPA. Mean PDL1 protein expression in the tumor with PDL1 present in a hybrid ecDNA structure (TCGA-CV-5443) was increased 7.6x relative to the mean of the other eight TCGA tumors (one sample T-test p<0.001; Figure 3.10f).

We analyzed the presence of hybrid DNA, hybrid RNA, hybrid ecDNA, and human circular ecDNA as compared to clinical features in the institutional cohort and in TCGA. In the institutional cohort there was no significant correlation with aggressive pathologic features (perineural invasion, lymphovascular invasion, or extranodal extension), poorly differentiated tumors, smoking status. Hybrid RNA status was more likely in patients with a drinking history (11/11 vs. 18/26; p = 0.038), and in patients who had a drinking history and had more than 10 pack years of smoking (6/11 vs. 3/23, p = 0.010). There was no association with the above groups and AJCC v7 staging. We then compared overall and recurrence-free survival based on presence of hybrid DNA, hybrid RNA, hybrid circular ecDNA, and human circular ecDNA. There were ten deaths, with the median time to death being 121 months. There were eleven recurrences and/or deaths, with the median time to event being 104 months. We found no significant difference in overall or recurrence free survival. In a multivariable proportional hazards model adjusting for age, stage, and treatment modality (surgery, radiation, and/or chemotherapy), we found no significant relationship between recurrence-free survival and hybrid DNA (p = 0.105), hybrid RNA (p = 0.540), hybrid ecDNA (p = 0.486), or human circular ecDNA (p = 0.282). Nor did we find a significant relationship between overall survival and hybrid DNA presence (p = 0.150), hybrid RNA presence (p = 0.525), hybrid ecDNA presence (p = 0.667), or human circular ecDNA presence (p=0.195).

In TCGA, there were 7 deaths in the 28 patients for whom survival data was available. Multivariable survival analysis was not possible due to the size of the study population, however univariable analysis showed that neither patients with hybrid DNA (p=0.172) or human circular ecDNA (p = 0.649) were more likely to die. Patients with hybrid ecDNA had decreased likelihood of death (0/7 deaths vs. 7/21 deaths in non-hybrid ecDNA; p = 0.078), as did patients with hybrid RNA (1/15 deaths vs. 6/13 in non-hybrid RNA; p = 0.016).

It is critical to note that neither our cohort nor TCGA was not powered to detect an expected difference in survival. We were unable to perform multivariable analyses to adjust for additional clinical factors due to the small cohort and infrequency of failure events.

3.4 Discussion

Intrachromosomal oncogene transcription and amplification have been the dominant paradigm for oncogene mediated transformation for decades. Similarly, HPV viral integration into human chromosomes and expression of oncoproteins E6 and E7 have been the classic mechanism for HPV-mediated oncogenesis (4). The recent definition of ecDNA as a driver of oncogene amplification and overexpression that is found in nearly of half of all human cancers has altered the paradigm of the role of extrachromosomal circular DNA in carcinogenesis (6,11). Recently non-integrated HPV has been reported in HPVOPC, indicating that HPV may exert oncogenic effects while maintaining status as vcDNA (46). The data we present describes an unexpected interaction of the HPV genome with cancer associated ecDNA. Specifically, our results suggest that HPVOPC frequently employ ecDNA in human as well as human-viral hybrid forms that leverage ecDNA mediated viral gene transcription to drive high expression of hybrid viral-human oncogene transcripts. Furthermore, HPVOPC ecDNA structures express a broad variety of human and hybrid cancer related transcripts including functional oncogenic noncoding RNA and immune evasion checkpoint proteins, in addition to traditional oncogenes. HPVOPC hybrid transcripts drive high expression of E6*I as well as other noncanonical HPV transcripts. This confirms and expands the spectrum of oncogenic HPV gene expression beyond the traditional expression of E6 and E7. In addition, HPV vcDNA is found in HPVOPC in deletion structures that lack coding for viral capsid proteins without evidence of integration into host DNA, as well as in the traditional full length episomal state. Most strikingly, nearly all HPVOPC contain transcriptionally active, circular, oncogenic DNA outside of host chromosomes, including vcDNA, human ecDNA, or viral-human hybrid ecDNA.

As noted, HPVOPC tumors express hybrid human viral sequences with increased human and viral gene expression driven by HPV promoters from integrated and ecDNA structures. These hybrid transcripts often include human genes implicated in carcinogenesis, and ecDNA hybrid transcripts are often expressed at a higher level than hybrid transcripts expressed from genomic HPV integration. Our data show that E6 and E7 expression is slightly enhanced in tumors in the context of HPV integration into chromosomal loci, and E1, E2, E4, and E5 expression is only slightly decreased. However, hybrid genomes incorporated into circular ecDNA show with significant upregulation of E6 and E7 as well as dramatic upregulation of human genes within 10kb up and down-stream of the recombined region, with loss of E2, E4, and E5 expression. Taken together, these data indicate that the mechanism for the classic model of HPV carcinogenesis characterized by E6 and E7 overexpression is often driven by the formation of viral-human hybrid ecDNA structures, facilitating amplification of E6 and E7 and associated human genes.

The near universal presence of oncogenic, circular DNA that expresses oncogenic transcripts, whether viral, human, or hybrid, indicates that HPVOPC create a genomic context that is generally permissive for the stable maintenance of non-chromosomal, transcriptionally active, circular DNA. Traditionally, the E2 protein has been shown to be the regulator of episomal maintenance for high-risk HPV (47,48). However, we describe hybrid ecDNA structures that exclude E2 in tumors that have dominant or exclusive expression of E6 and E7. It is possible that high risk HPV gene products other than E2 may have effects on genomic maintenance and chromosomal structure, allowing for perturbations in DNA homeostasis that facilitate stability, replication, and heritability of circular, non-chromosomal DNA structures. The evolution and progression of benign HPV infection to HPVOPC, therefore, requires ongoing maintenance of stable ecDNA or vcDNA as part of carcinogenesis. In this context, these data show three main viral genomic/transcriptomic HPVOPC pathways: 1) nonintegrated HPV vcDNA in a canonical or non-canonical form with broad expression of HPV gene products, 2) chromosomal integration of HPV with retained expression of HPV gene products and overexpression of human hybrid transcripts, and 3) formation of viral-human hybrid ecDNA as well as integrated viral DNA with dramatic amplification and overexpression of E6 and E7 as part of human viral fusion transcripts, with dramatic reduction in early E2, E4, and E5 HPV gene product expression. These pathways are independently supported by recent data showing HPV integration does not necessarily result in high levels of E6 and E7 transcripts, and that tumors with non-integrated HPV16 actually overexpress E6 compared to tumors with integrated or mixed genomes (49,50).

We identified the most common viral HPVOPC mRNAs as polycistronic transcripts typically > 1,000 nt in length, and that the most common full-length transcript in viral ecDNA HPVOPC is 1,476 nt long, beginning at the p97 promoter with splicing at SD226-SA409 and

SD880-SA3358 extending to the early polyA tail, with coding potential for E6*I, full-length E7, E1^E4, and full-length E5. The function of the E6*I protein (a truncated protein with 43 residues) remains incomplete, although E6*I is found predominantly in high risk HPV strains and not in low risk strains (51). Studies suggest it may have opposing functions to full-length E6 with respect to p53 and, through procaspase 8, the cellular response to the TNF-family of cytokines (52,53). Meanwhile, the E6*I RNA (which contains the E7 ORF) functions as a source of E7 mRNA and increases the efficiency of E7 protein translation.(36,54) E6*I has been observed in E6-expressing keratinocytes to upregulate IL6, a key cytokine in tumorigenesis and inflammation which functions via the JAK-STAT pathway (55). In addition, E6*I expressing cells demonstrated increased p53 levels as well as increased reactive oxygen species levels which has the effect of increasing DNA damage in cells (56). E6*I RNA levels were significantly higher in cervical samples from patients who had higher grades of dysplasia (57). In our cell line systems, we were able to demonstrate that the most common transcript expressing E6*I, full-length E7, E1^E4, and full-length E5 was able to induce proliferation, indicating that the net biologic effect of coordinated expression of these genes can support a malignant phenotype. Conventional and long-read RNA-seq of HPVOPC also suggest that mechanisms of oncogenesis independent of E6 exist, as we found the truncated form E6*I to be more common than full length E6 in mRNAs. This finding is of interest because the crystal structure of the ternary complex that inactivates p53 is comprised of full length E6, not E6*I (58). We found E5 protein was commonly expressed, and E5 is increasingly being recognized as a driver of HPV-mediated tumorigenesis (59) and mediates resistance to checkpoint inhibitor blockade in head and neck SCC and can be targeted (60). Finally, we have recently published data demonstrating a similar proliferative cellular phenotype in systems where E2/E4/E5 are expressed in comparison to E6/E7 expression, indicating that E6/E7 gene products may not be as critical for HPV mediated carcinogenesis (61).

We have noted a striking propensity for canonical splicing junctions to be preserved in HPV viral transcripts, and that these canonical junctions serve as donors for acceptors to diverse human hybrid transcripts, resulting in consistent expression of specific HPV transcripts. Previous studies have identified hybrid viral-human transcripts in HPV anogenital lesions with the HPV splice donor being in the E1 ORF, but these findings have not yet been described in HPVOPC (62). We found that a unifying feature of tumors with hybrid genomes was preferential use of a human splice acceptor site for SD880, rather than the SA3358 HPV site. This is of interest because SA3358 has been documented as the primary splice acceptor involved in transcripts coding for E4, E5, E6, and E7 (38). Together, these data show that alternatively spliced forms of HPV genes, including E6*I, are in fact more highly expressed than conventional transcripts. These transcripts may be key to HPV carcinogenesis with potential value as therapeutic targets. In addition, the proliferation data we provide as well as prior reports of E2, E4, and E5 cooperative ability to support carcinogenesis indicate that coordinated expression of multiple transcripts may be necessary to reproduce phenotypic characteristics of malignancy (61). Similarly, we have found that the overexpressed human transcripts that are associated with human and hybrid ecDNA are quite diverse, comprising traditional oncogenes, e.g. CCND1, EGFR, oncogenic long non-coding RNAs like PVT1, as well as immune modulatory genes, including PDL1. In addition, circular HPV DNA molecules have often been assumed to represent traditional intact, full-length forms found in intact virus. However, we found that non-canonical HPV genome, including deletion in the L1/L2 region, were noted in the majority of tumors that contained HPV vcDNA, and that full-length intact HPV genome was not found in these tumors with non-canonical HPV vcDNA.

These data do have limitations and present opportunities for further investigation. The effect of viral promoters and genes that facilitate high expression of hybrid transcripts and human oncogene expression in hybrid ecDNA structures is presumably facilitated by chromatin modification that is found in human ecDNA (7). However, the additional effect of viral promoters in an ecDNA context may facilitate chromatin structural effects. Although we have previously noted that human ecDNA is associated with poorer prognosis for human cancers in general, we have not seen worse prognosis associated with human ecDNA in HPVOPC, perhaps due to the small sample sizes available for ecDNA characterization, as well as the general favorable prognosis of HPVOPC.

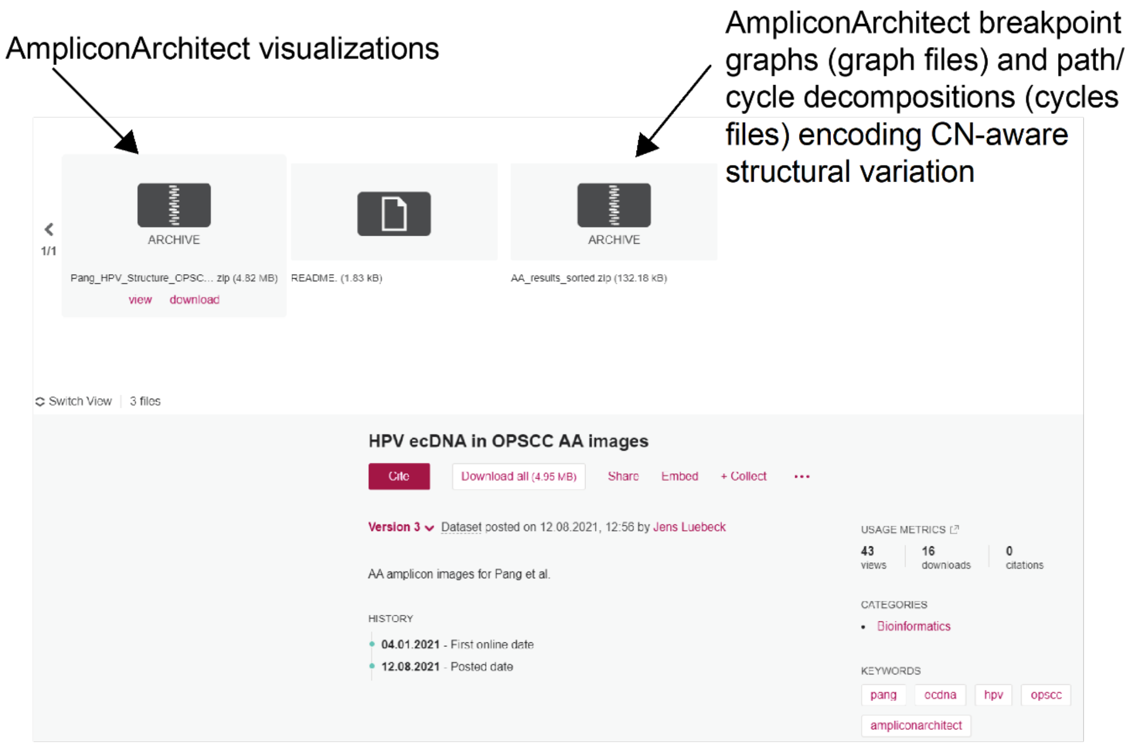
The understanding of the role of ecDNA in HPVOPC has direct implications for development of HPVOPC therapy as ecDNA-mediated amplification has been recognized as a potential mechanism of therapeutic resistance and specific, overexpressed oncogenes on ecDNA may be targeted with precision medicine approaches (6). EcDNA may also have interactions with chromosomes, affecting stability and integration that may also provide therapeutic opportunities for HPVOPC (63), including the presence of ecDNA as a marker for susceptibility to DNA repair inhibition. Indeed, HPVOPC has been shown to be sensitive to Wee-1 and CHK1/2 inhibition, and SCC090 cells that we have identified as ecDNA driven in this manuscript have shown dramatic apoptotic response to ionizing radiation in combination with the Chk1/2 inhibitor prexasertib (64-66). Non-chromosomal circular oncogenic DNA is present and transcriptionally active in nearly all HPVOPC in the form of ecDNA and vcDNA, and the mechanisms of circular DNA formation and maintenance themselves may be potential therapeutic targets. It is possible that the HPV gene products that facilitate general mechanisms of non-chromosomal circular DNA formation may be targeted in all HPVOPC, to target vcDNA as well as human and hybrid ecDNA that transcribe a variety of oncogenic products. Finally, human genes overexpressed via ecDNA in individual tumors have key driver oncogenic roles as well as roles in immune evasion that may serve as targets for personalized therapy.

3.5 Acknowledgements

Chapter 3, in full, is a reprint of the material “Extrachromosomal DNA in HPV mediated oropharyngeal cancer drives diverse oncogene transcription” by John Pang, Nam Nguyen, Jens Luebeck, Laurel Ball, Andrey Finegersh, Shuling Ren, Takuya Nakagawa, Mitchell Flagg, Sayed Sadat, Paul S. Mischel, Guorong Xu, Kathleen Fisch, Theresa Guo, Gabrielle Cahill, Bharat Panuganti, Vineet Bafna and Joseph Califano, as it appears in Clinical Cancer Research 2021. The dissertation author was one of the primary investigators and authors of this paper. In addition, we acknowledge Amanda Birmingham for assistance with data processing within the UCSD Center for Computational Biology.

3.6 Appendix

Screenshot of
dx.doi.org/10.6084/m9.figshare.135200 87


AmpliconArchitect visualizations
AmpliconArchitect breakpoint graphs (graph files) and path/ cycle decompositions (cycles files) encoding CN-aware structural variation
ARCHIVE
ARCHIVE
1/1
view
download
Switch View
Cite
Download all (4.95 MB)
Share
Embed
+ Collect
Version 3

Pang_HPV_Structure_OPSC... zip (4.82 MB)

README. (1.83 kB)

AA_results_sorted.zip (132.18 kB)

3 files

HPV ecDNA in OPSCC AA images

Dataset posted on 12.08.2021, 12:56 by Jens Luebeck

AA amplicon images for Pang et al.

HISTORY

- 04.01.2021 - First online date
- 12.08.2021 - Posted date

USAGE METRICS

43	16	0
views	downloads	citations

CATEGORIES

- Bioinformatics

KEYWORDS

pang ecdna hpv opscc ampliconarchitect

Figure 3.1: Output from Amplicon Architect of the institutional and TCGA cohorts. Output from Amplicon Architect of the institutional and TCGA cohorts, including Amplicon figures and text files detailing breakpoints.

Figure 3.2: HPV oropharynx cancers display distinct patterns of hybrid status including extrachromosomal DNA structures. **a,** Unsupervised hierarchical clustering of splicing patterns for HPV16+ tumors (n=37). Hybrid genome (DNA) and transcriptome (RNA) status were assigned based on ViFi. Human ecDNA (extrachromosomal DNA) and hybrid ecDNA status were assigned based on AmpliconArchitect analysis. The accompanying heatmap displays splice junction coverage normalized as a percent of RNA-seq reads per tumor. Clustering of hybrid genome and hybrid transcriptome depended on splice junction coverage. Tumor T14 had evidence of hybrid ecDNA but had too few reads (241 RNA-seq reads mapping to HPV16) to integrate into the heatmap. Canonical and non-canonical circular viral genome structure status was determined from AmpliconArchitect analysis. Tumor samples which did not have hybrid ecDNA were classified as non-canonical if they contained a cyclic decomposition and > 100 bp of rearranged genomic content (including indels), while canonical circular status was assigned if no such large rearrangements were present. **b,** Unsupervised hierarchical clustering of splicing patterns of HPV16+ tumors in the TCGA cohort (n=28) validates the presence of hybrid, human, or viral ecDNA in the HPV16 cohort tumors using aforementioned ViFi and AmpliconArchitect. Tumor CV-7406 had HPV16 genomic copy number < 1. **c,** Circular genome structure of a 28 kbp human-viral hybrid ecDNA in T41. The circular genome CycleViz plot shows the following properties from outside to inside. Outer: putative genomic structure of the ecDNA, with genes and gene directions. In light blue a long-read transcript mapping to the circular ecDNA is shown, connecting viral regions to TTC33, then back to viral regions again. Middle track: black primary data indicates the RNA-seq positional coverage. Orange secondary data indicates the log2 ratio of T41's exon-level FPKM values to the median FPKM values for those locations in study samples without ecDNA or viral integration. The purple line indicates a log2 positional FPKM ratio of 0. Inner track: black primary data indicates the WGS positional coverage. Light green secondary data indicates the mean WGS coverage for the chromosome from which the ecDNA was derived. Numeric values for the light grey tick lines are shown in the legend. The orange log2 positional FPKM ratio values in the middle track demonstrate increased expression of the regions compared to the median FPKM of study samples without ecDNA at those locations. **d,** Circular genome structure of a 67 kbp human-viral hybrid ecDNA in TCGA-CV-5443. The circular genome CycleViz plot shows the following properties from outside to inside. Outer: putative genomic structure of the ecDNA, with genes and gene directions. Middle track: black primary data indicates the RNA-seq positional coverage. Orange secondary data indicates the log2 value of the ratio of TCGA-CV-5443's exon-level FPKM values to the median FPKM values for those locations in study samples without ecDNA or viral integration. The purple line indicates a log2 positional FPKM ratio of 0. Inner track: black primary data indicates the WGS positional coverage. Light green secondary data indicates the mean WGS coverage for the chromosome from which the ecDNA was derived. Numeric values for the light grey tick lines are shown in the legend. The orange log2 positional FPKM ratio values in the middle track demonstrate increased expression of the regions compared to the median FPKM of study samples without ecDNA at those locations.

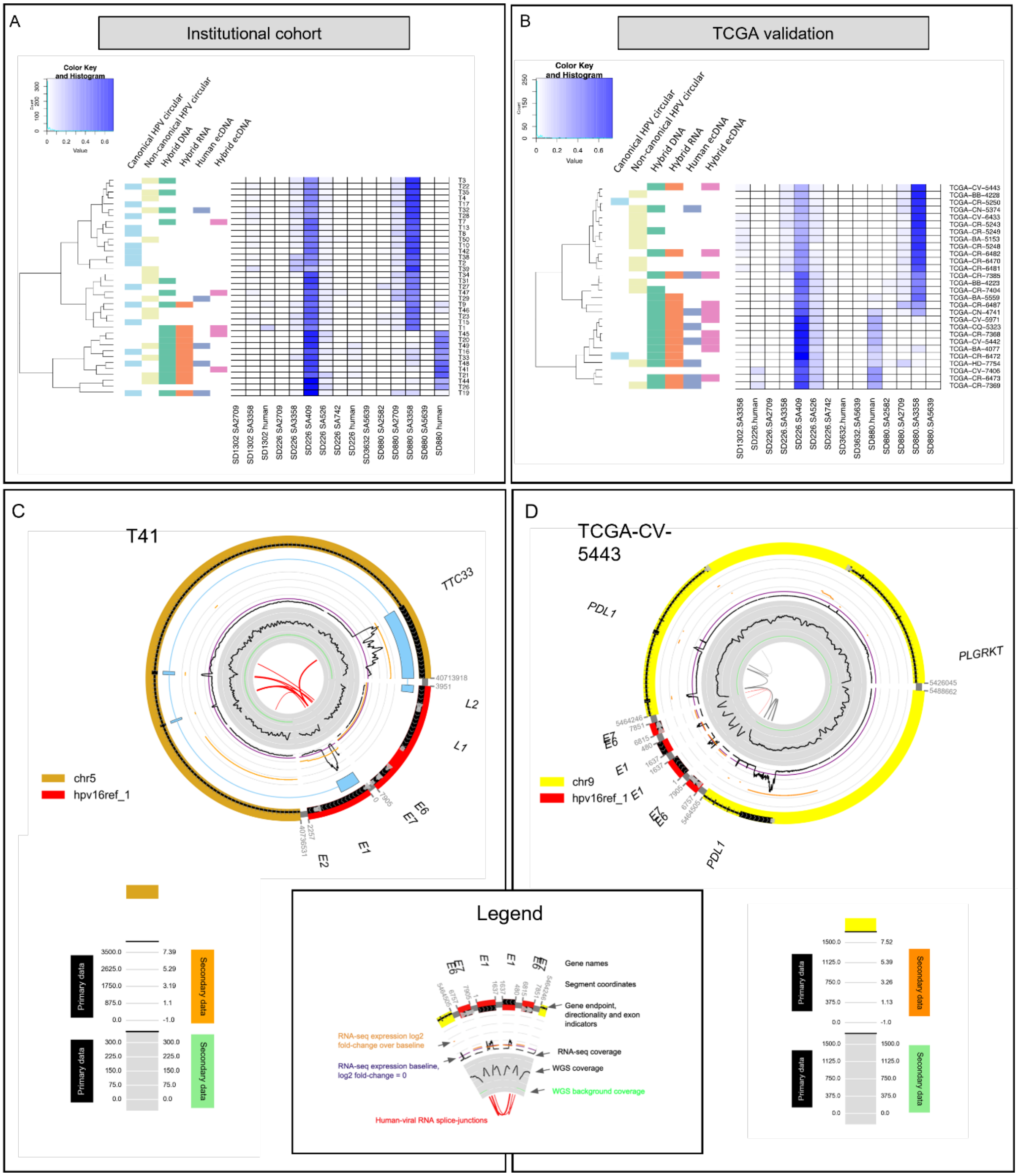

A
Institutional cohort
Color Key and Histogram
Count
Value
Canonical HPV circular
Non-canonical HPV circular
Hybrid DNA
Hybrid RNA
Human ecDNA
Hybrid ecDNA
T3
T22
T35
T4
T17
T32
T28
T7
T13
T8
T50
T10
T42
T38
T2
T39
T34
T31
T27
T47
T29
T9
T46
T23
T15
T1
T45
T20
T49
T16
T33
T48
T41
T21
T44
T26
T19
SD1302 SA2709
SD1302 SA3358
SD1302 human
SD226 SA2709
SD226 SA3358
SD226 SA409
SD226 SA526
SD226 SA742
SD226 human
SD3632 SA5639
SD880 SA2582
SD880 SA2709
SD880 SA3358
SD880 SA5639
SD880 human
B
TCGA validation
Color Key and Histogram
Count
Value
Canonical HPV circular
Non-canonical HPV circular
Hybrid DNA
Hybrid RNA
Human ecDNA
Hybrid ecDNA
TCGA–CV–5443
TCGA–BB–4228
TCGA–CR–5250
TCGA–CN–5374
TCGA–CV–6433
TCGA–CR–5243
TCGA–CR–5249
TCGA–BA–5153
TCGA–CR–5248
TCGA–CR–6482
TCGA–CR–6470
TCGA–CR–6481
TCGA–CR–7385
TCGA–BB–4223
TCGA–CR–7404
TCGA–BA–5559
TCGA–CR–6487
TCGA–CN–4741
TCGA–CV–5971
TCGA–CQ–5323
TCGA–CR–7368
TCGA–CV–5442
TCGA–BA–4077
TCGA–CR–6472
TCGA–HD–7754
TCGA–CV–7406
TCGA–CR–6473
TCGA–CR–7369
SD1302.SA3358
SD226.human
SD226.SA2709
SD226.SA3358
SD226.SA409
SD226.SA526
SD226.SA742
SD3632.human
SD3632.SA5639
SD880.human
SD880.SA2582
SD880.SA2709
SD880.SA3358
SD880.SA5639
C
T41
TTC33
40713918
3951
L2
L1
E6
E7
7905
0
E1
E2
2257
40736531
chr5
hpv16ref_1
Primary data
Secondary data
3500.0
2625.0
1750.0
875.0
0.0
7.39
5.29
3.19
1.1
-1.0
300.0
225.0
150.0
75.0
0.0
D
TCGA-CV-5443
PDL1
PLGRKT
5426045
5488662
5464246
7851
6815
480
1637
E7
E6
E1
chr9
hpv16ref_1
6757
5464565
1500.0
1125.0
750.0
375.0
0.0
7.52
5.39
3.26
1.13
-1.0
Legend
Gene names
Segment coordinates
Gene endpoint, directionality and exon indicators
RNA-seq expression log2 fold-change over baseline
RNA-seq expression baseline, log2 fold-change = 0
RNA-seq coverage
WGS coverage
WGS background coverage
Human-viral RNA splice-junctions

Figure 3.3: Non-canonical virus structures. **a,** Amplicon of T39 non-canonical HPV16 viral ecDNA, demonstrating deletion (red line) of 678bp segment overlapping L1 gene. **b,** Circularized depiction of T39 non-canonical HPV16 viral ecDNA, inner grey arced lines demonstrate deletion of 678bp segment of L1 gene and closure of circular genome. **c,** Top: genome map of HPV16, indicating viral intra-genomic breakpoints and RNA breakpoints. Middle (DNA): Mean proportion of the CN at each position over the maximum HPV16 CN for each sample with non-canonical viral structures (yellow) and those with canonical viral structures (blue), based on AA copy number estimates. Bottom (RNA): RNA-seq coverage for non-canonical viral structures (yellow) and those with canonical viral structures (blue). Coverage is rescaled by the median coverage across the virus, and the log2 value is shown. The lower plot shows the log2 ratio of scaled coverage between orange and grey from the upper plot, representing a log2 foldchange in mean scaled coverage. Non-canonical viral structures demonstrated loss of copy number across the L2 and L1 viral genomes, with decreased transcription of the E6, E7, L1, and L2 genomic regions.

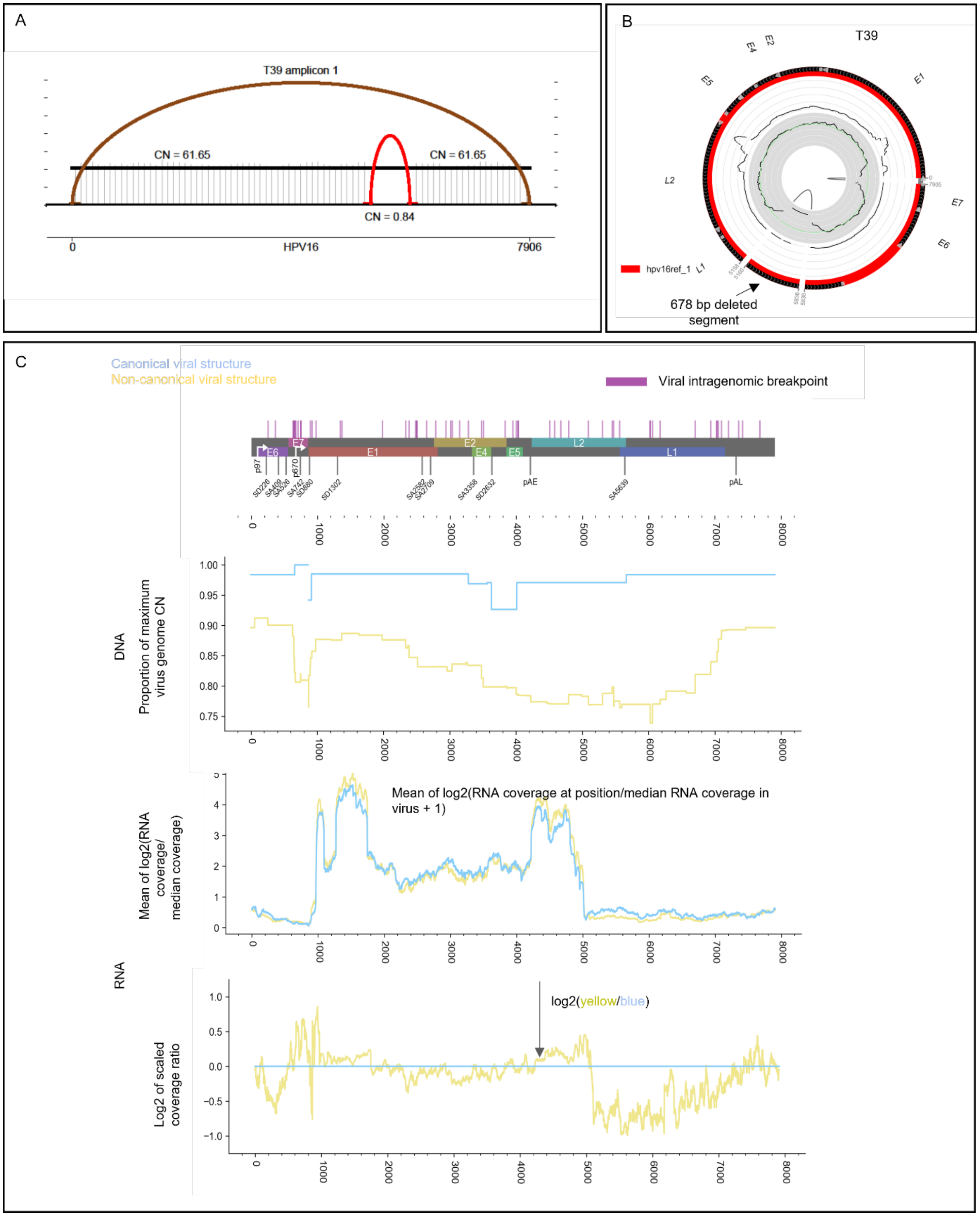
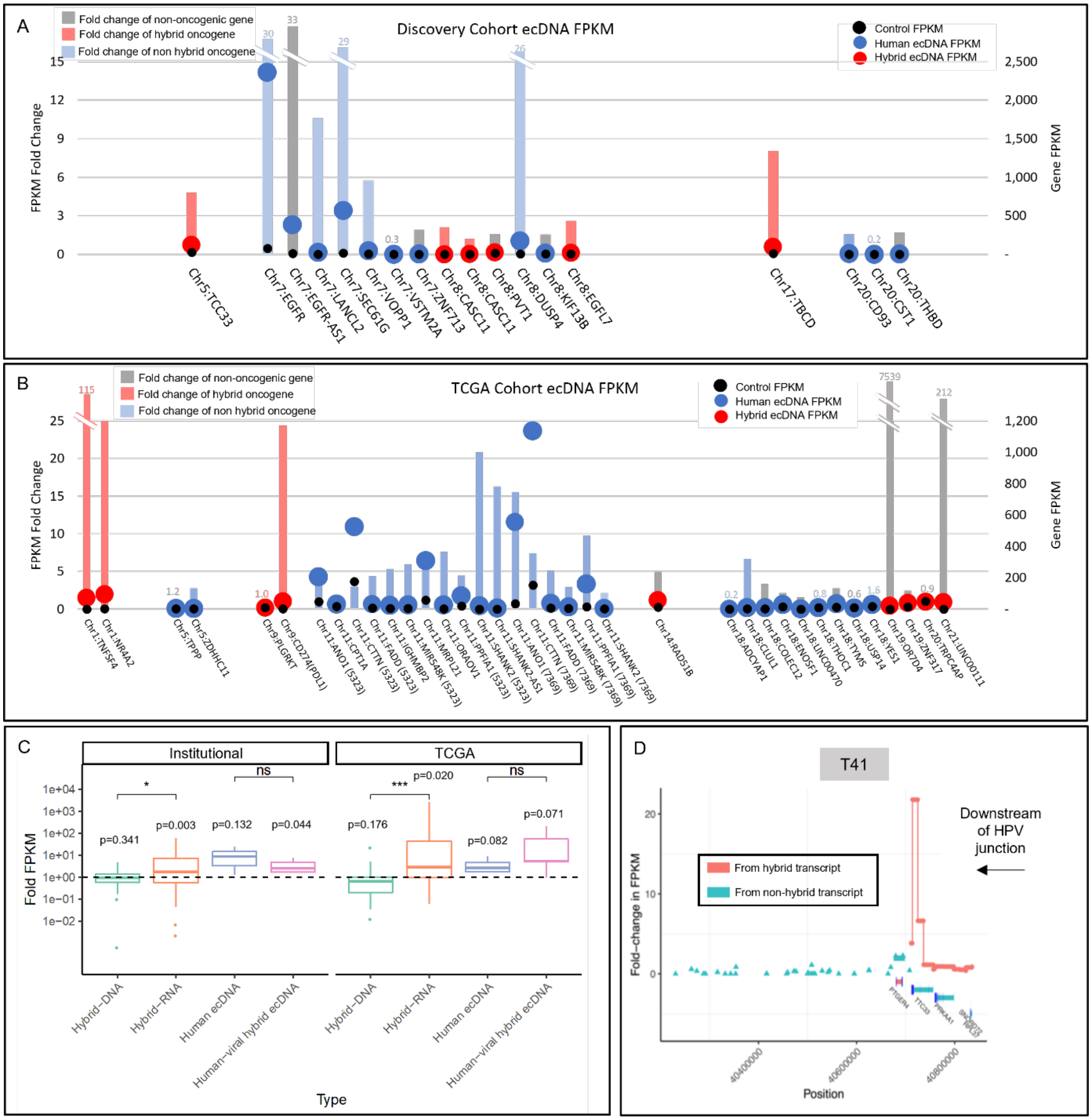

A
T39 amplicon 1
CN = 61.65
CN = 61.65
CN = 0.84
0
HPV16
7906
B
T39
E2
E4
E5
E1
L2
E7
E6
L1
hpv16ref_1
678 bp deleted segment
C
Canonical viral structure
Non-canonical viral structure
Viral intragenomic breakpoint
E7
E6
E1
E2
E4
E5
L2
L1
p97
p670
SD226
SA409
SA526
SA742
SD880
SD1302
SA2582
SA2709
SA3358
SD3632
pAE
SA5639
pAL
0
1000
2000
3000
4000
5000
6000
7000
8000
DNA
Proportion of maximum virus genome CN
1.00
0.95
0.90
0.85
0.80
0.75
RNA
Mean of log2(RNA coverage/ median coverage)
5
4
3
2
1
0
Mean of log2(RNA coverage at position/median RNA coverage in virus + 1)
Log2 of scaled coverage ratio
1.0
0.5
0.0
−0.5
−1.0
log2(yellow/blue)

Figure 3.4: Oncogene expression is upregulated in extrachromosomal DNA structures. For **a** and **b**, x-axis groups ecDNA structures for individual tumor samples and orders by chromosomal location of the gene. The left axis corresponds to the bars, which are colored by the gene's status as hybrid ecDNA oncogene (red), human ecDNA oncogene (blue), or not an established oncogene (grey) and indicates the fold-change in expression (FPKM) of the ecDNA+ tumor's gene (*g*) against the expression of *g* in other tumors where the gene is not found on ecDNA (FPKM of *g* in the ecDNA+ tumor)/(mean FPKM of *g* in tumors where *g* is not on ecDNA). The right axis uses overlaid bubbles defining FPKM of each gene of interest in ecDNA in individual tumors colored by ecDNA type (blue for human or red for hybrid) and also for mean FPKM in tumors that do not contain the gene of interest on ecDNA (black dots). FPKM is corrected for depth and fragment length to provide a representative coverage of the gene, normalized against the read depth of the sequenced sample, and corrected for overrepresentation of shorter reads. **a,** EcDNA associated with human genes from the Institutional cohort exhibited upregulation of associated genes, with a mean FPKM-ratio of 9.59 (SD 12.06); median 2.59 (IQR 1.58– 10.60). **b,** EcDNA associated with human genes from TCGA exhibited upregulation of associated genes, with a mean FPKM sample:control ratio of 212.54 (SD 1,221.3); median 4.32 (IQR 2.14– 7.57). Genes (parentheses) refers to TCGA tumor affiliated with ecDNA gene. **c,** Human genes proximal to the viral human junction showed increased expression in association with both hybrid genomes and hybrid transcriptomes for both the institutional and TCGA cohorts. Human-viral hybrid ecDNA was associated with increased expression of associated genes in the Institutional cohort. Hybrid RNA tumors showed higher upregulation of associated genes than hybrid DNA tumors in both Institutional and TCGA cohorts. **d,** T41 shows dramatic increase in TTC33 expression downstream of the HPV-human hybrid sequence junction.


A
Discovery Cohort ecDNA FPKM
Fold change of non-oncogenic gene
Fold change of hybrid oncogene
Fold change of non hybrid oncogene
Control FPKM
Human ecDNA FPKM
Hybrid ecDNA FPKM
FPKM Fold Change
Gene FPKM
15
12
9
6
3
0
2,500
2,000
1,500
1,000
500
-
30
33
29
26
0.3
0.2
Chr5:TCC33
Chr7:EGFR
Chr7:EGFR-AS1
Chr7:LANCL2
Chr7:SEC61G
Chr7:VOPP1
Chr7:VSTM2A
Chr7:ZNF713
Chr8:CASC11
Chr8:CASC11
Chr8:PVT1
Chr8:DUSP4
Chr8:KIF13B
Chr8:EGFL7
Chr17:TBCD
Chr20:CD93
Chr20:CST1
Chr20:THBD
B
TCGA Cohort ecDNA FPKM
Fold change of non-oncogenic gene
Fold change of hybrid oncogene
Fold change of non hybrid oncogene
Control FPKM
Human ecDNA FPKM
Hybrid ecDNA FPKM
FPKM Fold Change
Gene FPKM
25
20
15
10
5
0
1,200
1,000
800
600
400
200
-
115
1.2
1.0
7539
212
0.2
0.8
0.6
1.6
0.9
Chr1:TNFSF4
Chr1:NR4A2
Chr5:TPPP
Chr5:ZDHHC11
Chr9:PLGRKT
Chr9:CD274(PDL1)
Chr11:ANO1 (5323)
Chr11:CPT1A
Chr11:CTTN (5323)
Chr11:FADD (5323)
Chr11:IGHMBP2
Chr11:MIR548K (5323)
Chr11:MRPL21
Chr11:ORAOV1
Chr11:PPFIA1 (5323)
Chr11:SHANK2 (5323)
Chr11:SHANK2-AS1
Chr11:ANO1 (7369)
Chr11:CTTN (7369)
Chr11:FADD (7369)
Chr11:MIR548K (7369)
Chr11:PPFIA1 (7369)
Chr11:SHANK2 (7369)
Chr14:RAD51B
Chr18:ADCYAP1
Chr18:CLUL1
Chr18:COLEC12
Chr18:ENOSF1
Chr18:LINC00470
Chr18:THOC1
Chr18:TYMS
Chr18:USP14
Chr18:YES1
Chr19:OR7D4
Chr19:ZNF317
Chr20:TRPC4AP
Chr21:LINC00111
C
Institutional
TCGA
Fold FPKM
1e+04
1e+03
1e+02
1e+01
1e+00
1e−01
1e−02
*
ns
***
ns
p=0.341
p=0.003
p=0.132
p=0.044
p=0.176
p=0.020
p=0.082
p=0.071
Hybrid−DNA
Hybrid−RNA
Human ecDNA
Human−viral hybrid ecDNA
Hybrid−DNA
Hybrid−RNA
Human ecDNA
Human−viral hybrid ecDNA
Type
D
T41
Fold−change in FPKM
20
10
0
From hybrid transcript
From non-hybrid transcript
Downstream of HPV junction
PTGER4
TTC33
PRKAA1
SNORD72
RPL37
40400000
40600000
40800000
Position

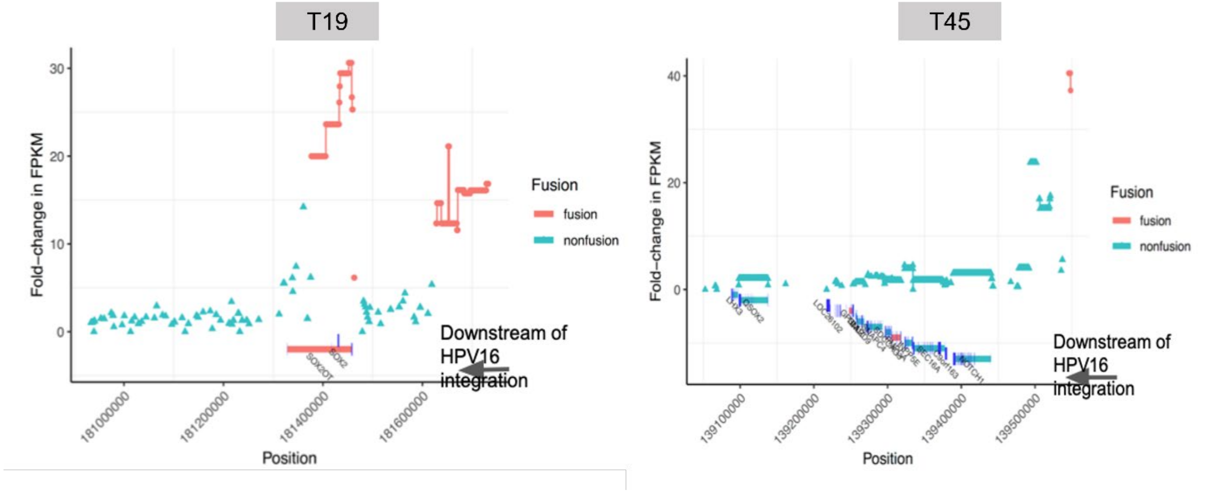
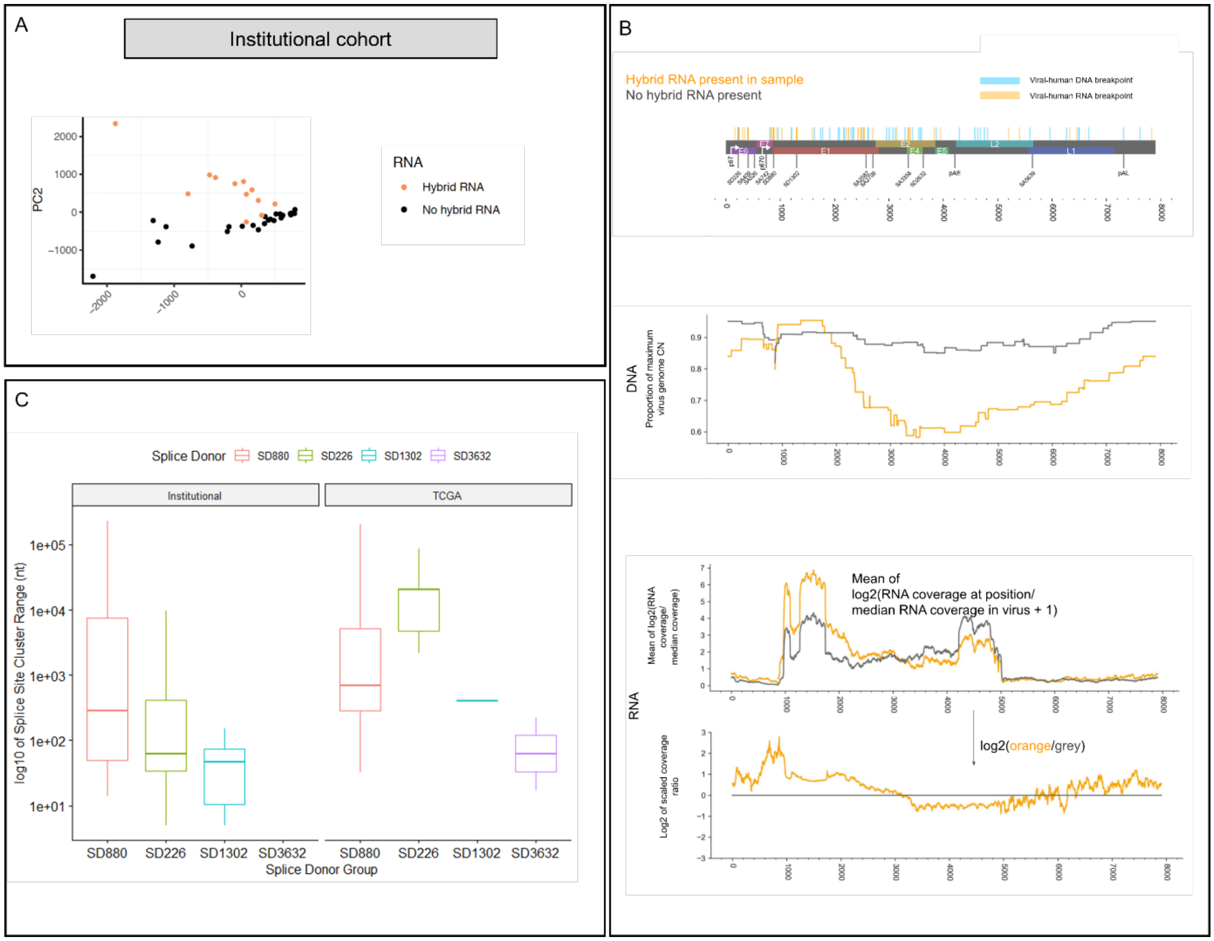


Figure 3.5: Downstream expression at site of integration of human/viral hybrid RNA sequence. Tumors T19 and T45 shows dramatic increase in TTC33 expression downstream of HPV-human hybrid sequence.

Figure 3.6: Hybrid genome and transcriptome status results in distinct splicing patterns and HPV oncogene expression levels. **a,** Principal component analysis of splice junction expression in individual tumors with annotation of tumors by presence or absence of human-viral hybrid transcripts. PCA of the tumor cohort supports the finding that hybrid transcriptome tumors (orange dots) exhibit distinct splicing patterns compared to non-hybrid transcriptome tumors (black dots). **b,** Top: genome map of HPV16, indicating human-viral DNA breakpoints and RNA breakpoints. Middle (DNA): Mean proportion of the CN at each position over the maximum HPV16 CN for each sample with hybrid RNA transcripts (orange) and those without (grey), based on AA copy number estimates. We observed selective enrichment of viral genomic copy number in the E6/E7 region, and 5' end of E1 in hybrid RNA tumors, while those without hybrid RNA showed much more uniform enrichment of viral copy number throughout the genome. Bottom (RNA): RNA-seq coverage for hybrid RNA tumors (orange) and those without hybrid RNA (grey). Coverage is rescaled by the median coverage across the virus, and the log2 value is shown. The lower plot shows the log2 ratio of scaled coverage between orange and grey from the upper plot, representing a log2 fold-change in mean scaled coverage. The highest peak overlaps SD880, indicating selectively increased transcription of that location in tumors with hybrid RNA. We also observed that the E4/E5 region, including the 5' end of L2 were far less likely to have selective enrichment for genomic copy number in hybrid RNA tumors, and that there was decreased expression of those regions as compared to tumors without hybrid RNA. **c,** Splice acceptor cluster quantification of hybrid transcripts for Institutional and TCGA cohorts stratified based on splice donor location. We found that the median splice acceptor range varied from 29 to 20,399 nucleotides wide across the TCGA and institutional cohort, and were narrowest for SD1302 (median 28.8 in TCGA; 406 in institutional) and widest for SD226 (20,399 in institutional cohort).


A
Institutional cohort
PC2
2000
1000
0
−1000
−2000
−1000
0
RNA
● Hybrid RNA
● No hybrid RNA
B
Hybrid RNA present in sample
No hybrid RNA present
Viral-human DNA breakpoint
Viral-human RNA breakpoint
E7
E6
E1
E2
E4
E5
L2
L1
p97
SD226
SA409
SA526
p670
SA742
SD880
SD1302
SA2582
SA2709
SA3358
SD3632
pAE
SA5639
pAL
0
1000
2000
3000
4000
5000
6000
7000
8000
DNA
Proportion of maximum virus genome CN
0.9
0.8
0.7
0.6
RNA
Mean of log2(RNA coverage/ median coverage)
Mean of
log2(RNA coverage at position/
median RNA coverage in virus + 1)
log2(orange/grey)
Log2 of scaled coverage ratio
7
6
5
4
3
2
1
0
-1
-2
-3
C
Splice Donor SD880 SD226 SD1302 SD3632
Institutional
TCGA
log10 of Splice Site Cluster Range (nt)
1e+05
1e+04
1e+03
1e+02
1e+01
SD880
SD226
SD1302
SD3632
SD880
SD226
SD1302
SD3632
Splice Donor Group

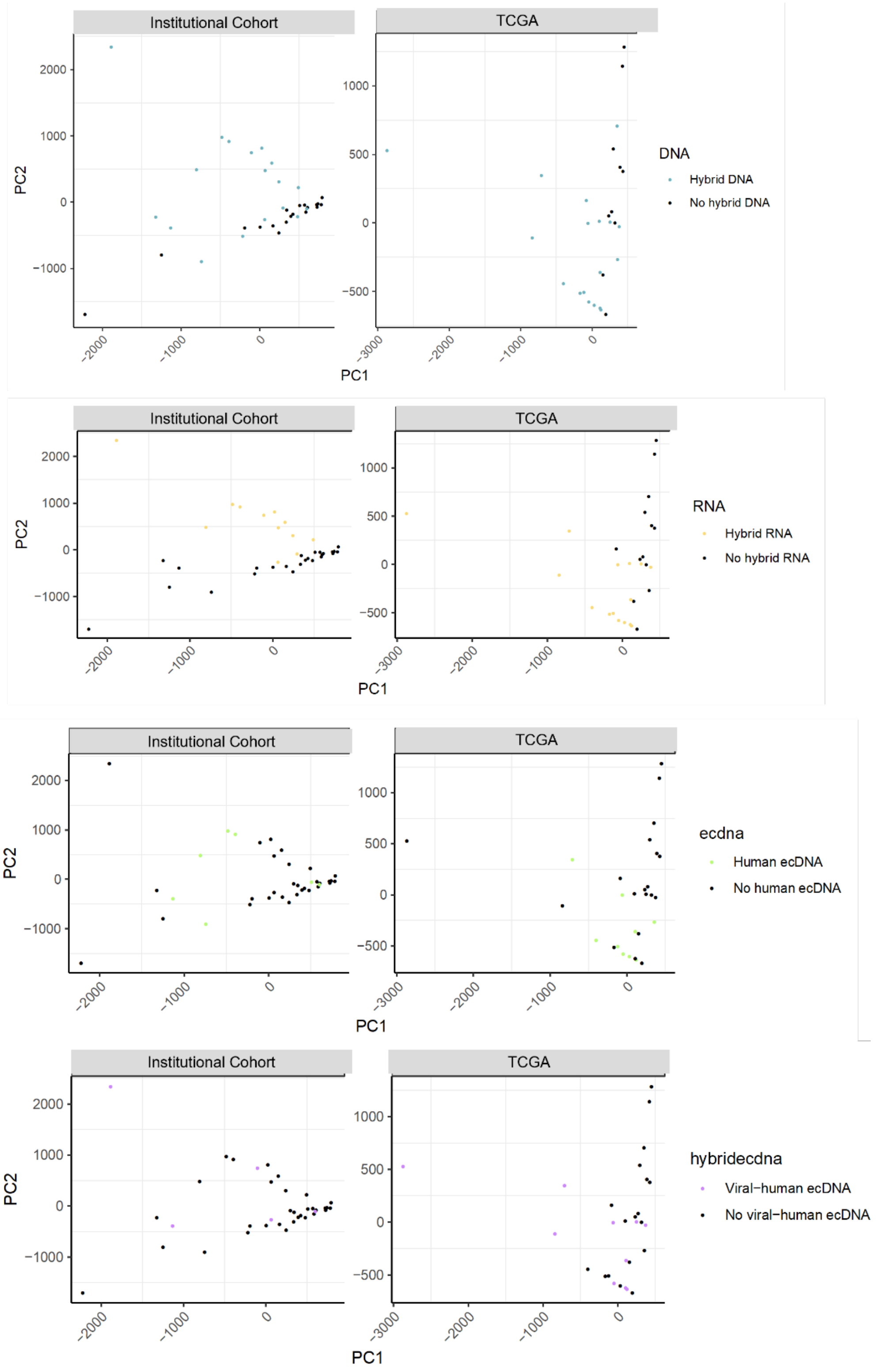
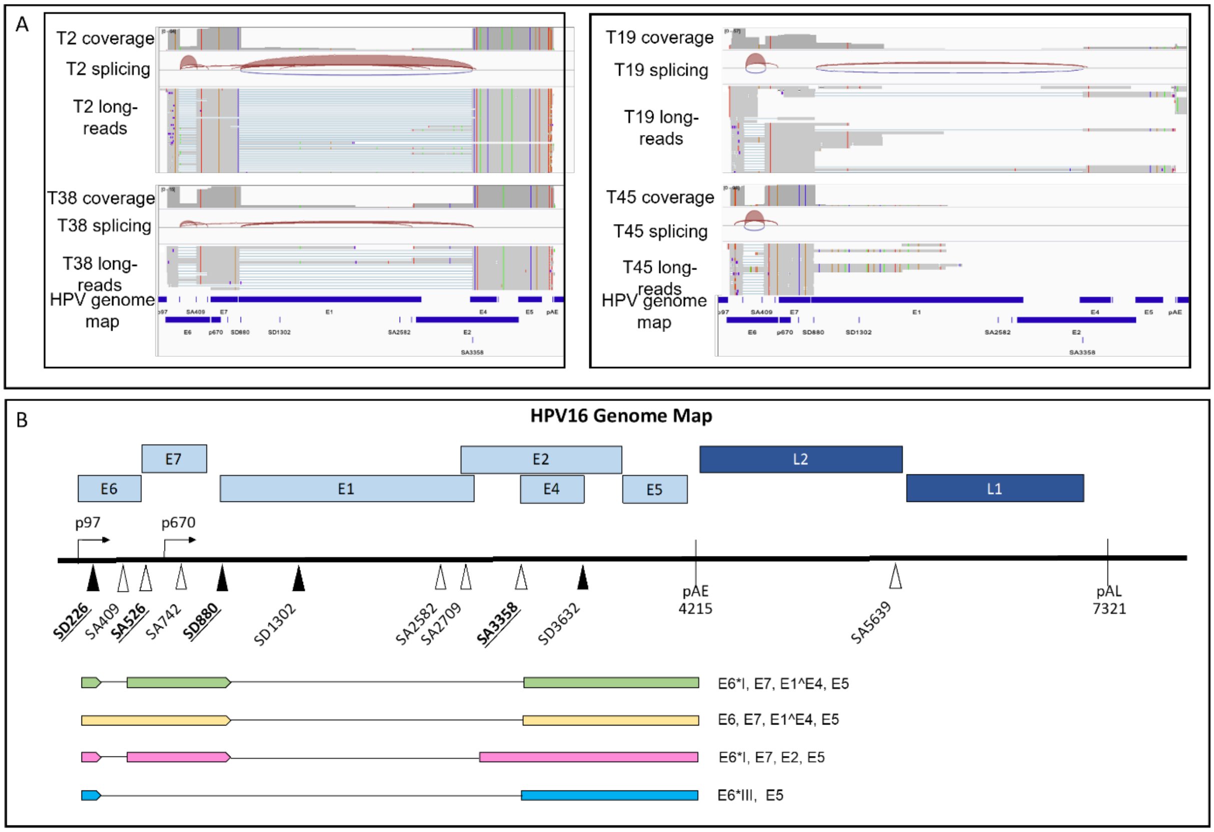


Figure 3.7: Principal component analyses of utilization of canonical HPV splice sites in institutional and TCGA cohorts. Principal component analysis of splicing patterns based on hybrid DNA, hybrid RNA, human ecDNA, and hybrid ecDNA status.

Figure 3.8: HPV16 full length transcript structure in primary HPVOPC tumors. **a,** Integrated Genome Viewer graphics are displayed of long-read RNA whole genome poly A transcript sequencing on a subset of two tumors without hybrid transcripts (T2 and T38), one human ecDNA tumor with hybrid transcript expression (T19), and one hybrid ecDNA tumor with hybrid transcripts (T45). **b,** The most common long read Iso-Seq transcripts mapping to HPV exclusively are depicted with their coding potential. The single most common full-length transcript in non-hybrid tumors was 1,476 nt long, beginning at the p97 promoter with splicing at SD226-SA409 and SD880-SA3358 extending to the early polyA tail, with coding potential for the E6 oncoprotein variant E6*I defined by SD226-SA409, full-length E7, full-length E4, and full-length E5. **c,** Poly-A tail-based long-read RNA sequencing of non-integrated and integrated tumors with mapping of full-length transcripts to HPV-16 genome demonstrates the transcriptome patterns in primary tumors. Identical isoforms have been color-coded. Read counts fewer than 5 were discarded. Full length coverage counts per Iso-Seq protocol are not proportionate to transcript quantity. Tumors without hybrid transcripts (T2 and T38) have distinct transcriptomes from those with hybrid transcripts and with ecDNA (T19, T45).

Integrated Genome Viewer Representations of Long-Read RNAseq on Non-hybrid and Hybrid HPV+ OPC



C

Tumor	Canonical circular HPV DNA	Noncanical circular HPV DNA	Hybrid DNA	Hybrid RNA	Human ecDNA	Hybrid ecDNA	Coding Potential	Read Length	Start	Stop	Start	Stop	Start	Stop
T2	■						E6*I, E7, E1^E4, E5	1460	97	SD226	SA409	SD880	SA3358	pAE
							E6, E7, E1^E4, E5	1642	97	SD880	SA3358	pAE		
							E6*III, E5	988	97	SD226	SA3358	pAE		
							E6, E7, E2, E5	2291	97	SD880	SA2709	pAE		
							E6*II, E7, E1^E4, E5	1343	97	SD226	SA526	SD880	SA3358	pAE
							E6*I, E7, E2, E5	2109	97	SD226	SA409	SD880	SA2709	pAE
							E6, E7, E1*	2008	97	2105				
							E6*I, E7, E1^E4, E5	1482	75	SD226	SA409	SD880	SA3358	pAE
							E6, E7, E1, E2, E4, E5	4118	97	pAE				
							E6*I, E7, E1^E4, E5	1543	14	SD226	SA409	SD880	SA3358	pAE
							E7, E1^E4, E5	1410	147	SD226	SA409	SD880	SA3358	pAE
T38	■						E6*I, E7, E1^E4, E5	1460	97	SD226	SA409	SD880	SA3358	pAE
							E6, E7, E1^E4, E5	1642	97	SD880	SA3358	pAE		
							E6*III, E5	988	97	SD226	SA3358	pAE		
							E6, E7, E2, E5	2291	97	SD880	SA2709	pAE		
							E6*II, E7, E1^E4, E5	1343	97	SD226	SA526	SD880	SA3358	pAE
							E6*I, E7, E2, E5	2109	97	SD226	SA409	SD880	SA2709	pAE
							E6, E7, E1*	2008	97	2105				
T19	■		■	■	■		E6*I, E7, E1^E4, E5	1460	97	SD226	SA409	SD880	SA3358	pAE
							E6, E7, E1^E4, E5	1642	97	SD880	SA3358	pAE		
							E6*I, E7	660	97	SD226	SA409	938		
T45			■	■		■	E6*I, E7, E1*	1830	97	SD226	SA409	2108		
							E7, E1*	1998	107	2105				

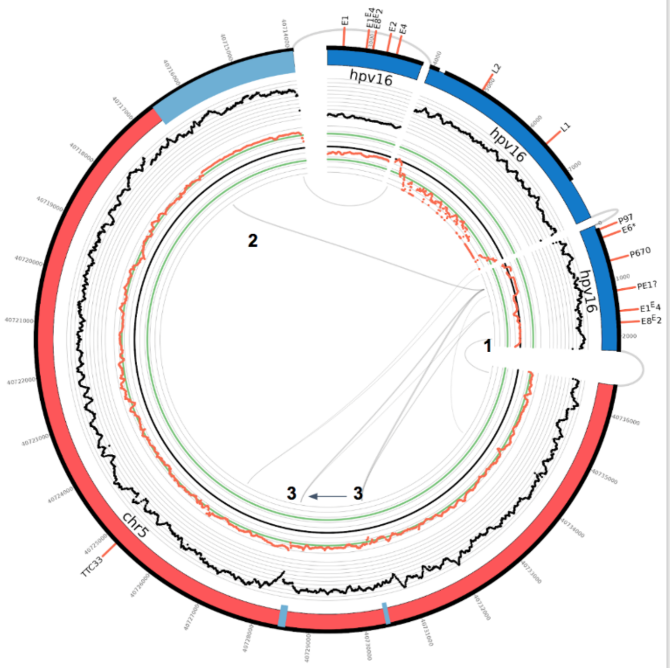
- T41 structure

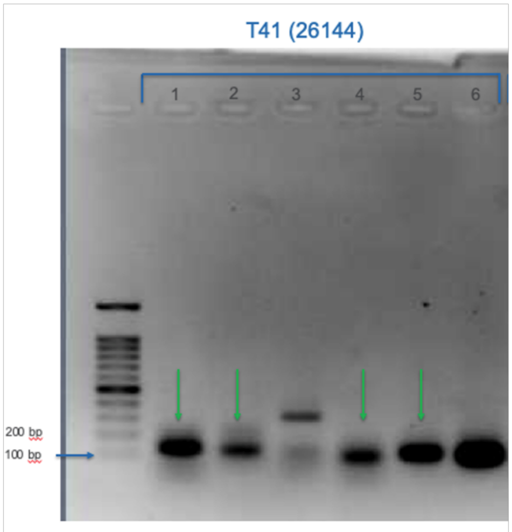
HPV genes in same orientation as TTC33 gene,
HPV sequence in blue, Human sequence in red

1) hpv16:2168-2258 (E1); chr5:40736402-40736532 (-)
168 reads
TGTTTTTAAGGTATCAAGGTGTAGAGTTTATGTCATT
TTTAACTGCATTAAAAAGATTTTTG
CAGTTGTATGTGTATGAAGCAGGAATATAAATTAGA
ATTTACCATGCATTTTGTTCTTGTCTAGGAAAGTATA
CCTACTAATGTAAATTTGAGGGTGATCTTGCTGTGT
GACCCACATTCTGCAGATTACT

2) hpv16:805-881 (E6); chr5:40716472-40716601(-)
CTGTTAATGGGCACACTAGGAATTGTGTGCCCCATC
TGTTCTCAGAAACCATAATCTACCATGGCTGATCCT
GCAGGCAATTCGAAGTTTTCAAGTAGCCCTTCACAT
CTATCCAATGAACCCTGAAATATGGAAAGAAGACCT
CTCTTGGGCAAGAACGCTCCAGGAGCAGCAGAAGG
TAGCACAGAGGATTAAAAAAAGTGAA

3)hpv884-887;chr5:40728533-40728578;40730364-40730417
CAGAAACCATAATCTACCATGGCTGATCCTG
CAGATATCGGGAGGCAATTCAGAAGTGGGATGAAGCACTACAG
TTAACTCCAAATGATGCTACCCTATACGAGATGAAATCACAGGTG
CTAATGTCTCTTCATGAAATGTTCCCAGCAGTACAT





Lane	Primers	Exp Amplicon
1Kb ladder		
1	T41-1-F1/R1	145
2	T41-1-F2/R2	132
3	T41-2-F1/R1	109
4	T41-3-F1/R1	85
5	T-41-3-F2/R2	108
6	E7F/R (pos control)	100

Sequenced read from T41-1-F1/R1:

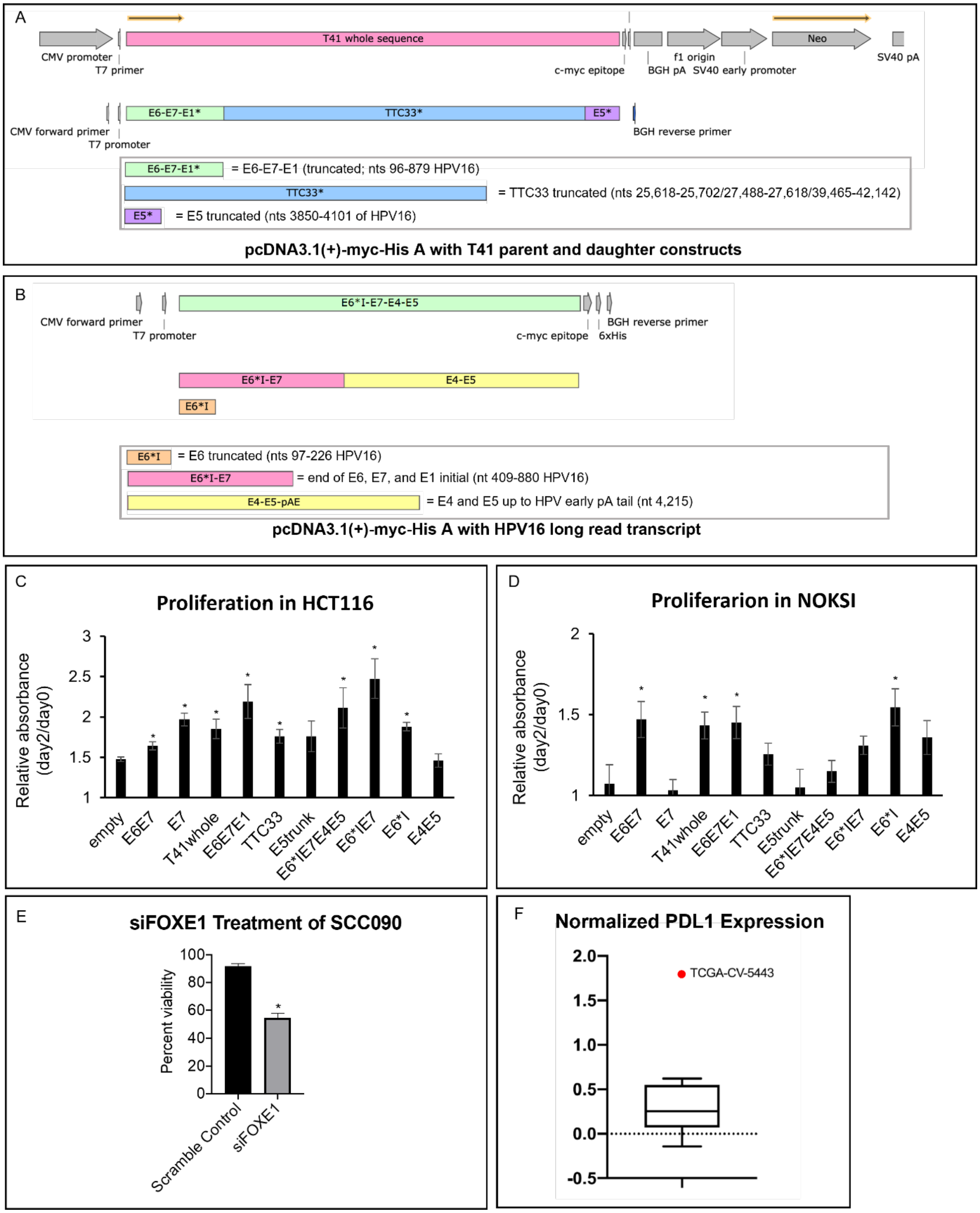
ATGTTTTATTAATTTTTGCAGTTGTATGTGTATGAAGCAGGAATATAAATTAGAATTTACCATGCATT
TTGTTCTTGTCTAGGAAAGTATCCTACTAATGTAAATTGAGGGGTNTGCCGG

Target junction from T41:

1) hpv16:2168-2258 (E1); chr5:40736402-40736532 (-) 168 reads
TGTTTTTAAGGTATCAAGGTGTAGAGTTTATGTCATTTTTAACTGCATTAAAAAGATTTTTG
CAGTTGTATGTGTATGAAGCAGGAATATAAATTAGAATTTACCATGCATTTTGTTCTTGTCTAGGAAAG

Figure 3.9: Verification of full-length transcript identified by Amplicon Architect. Three putative splice reads are displayed with viral sequence in blue and human sequence in red. RT-PCR demonstrates the expected size amplicon in agarose gel, followed by excision of gel and Sanger sequencing confirmation of expected sequence.

Figure 3.10: Functional activity of hybrid ecDNA structures and full-length HPV transcripts. **a,** T41 entire transcript (E6-E7-E1-TTC33*-E5*) as well as daughter constructs were cloned into a pcDNA 3.1(+)-myc-His A vector (Genscript, Inc.). **b,** Most common full length HPV16 RNA (E6*I-E7-E4-E5) as well as daughter constructs were cloned into a pcDNA 3.1(+)-myc-His A vector (Genscript, Inc.). **c,** Transfection of pcDNA cloned with T41 fusion transcript and HPV transcript promoted the proliferation in HCT116. The proliferation assays after transient transfection of pcDNA3.1 cloned with empty, E6*I, E7, E6*IE7, E6E7, E6*IE7E4E5, TTC33, E6E7E1, E5truncated, E4E5 and T41 vector (E6-E7-E1-TTC33*-E5*) were performed on HCT116. Proliferation was normalized to day0, and relative absorbance of day2/day0 data was shown. E6*I, E7, E6*IE7, E6E7, E6*IE7E4E5, TTC33, E6E7E1 and T41 vector promoted the proliferation significantly compared to the empty vector (p =0.02, 3x10^{-4}, 6x10^{-3}, 0.02, 0.03, 2x10^{-4}, 7x10^{-3}, and 0.01, respectively, *p<0.05, Student's *t*-test. Error bars represent standard error. **d,** Transfection of pcDNA cloned with T41 fusion transcript and HPV transcript promoted the proliferation in NOKSI cells. The proliferation assays after transient transfection of pcDNA3.1 cloned with empty, E6*I, E7, E6*IE7, E6E7, E6*IE7E4E5, TTC33, E6E7E1, E5truncated, E4E5 and T41 vector (E6-E7-E1-TTC33*-E5*) were performed on NOKSI cells. Proliferation was normalized to day0, and relative absorbance of day2/day0 data was shown. E6*I, E6E7, E6E7E1 and the entire intact hybrid transcript from T41 increased proliferation (*p*=0.02, 0.01, 0.03, and 0.04, respectively, **p*<0.05, Student's *t*-test. Error bars represent standard error. **e,** Proliferation assay after knockdown of FOXE1 was performed on SCC090. Knockdown of FOXE1 using siRNA of FOXE1 (Santa Cruz biotech) significantly suppressed the proliferation compared to the scramble control (Santa Cruz biotech). Percent viability was normalized to day0. *p<0.05, Student's t-test. Error bars represent standard error. **f,** Data derived via reverse phase protein arrays (RPPA) corresponding to head and neck squamous cell carcinoma samples included in TCGA were extracted from The Cancer Proteome Atlas (TCPA). Nine tumor samples identified as being HPV+ in the TCGA cohort had RPPA-derived expression data available in TCPA. Mean PDL1 protein expression in TCGA-CV-5443 (red data point outlier) was increased 7.6 X relative to the mean of the other eight TCGA tumors represented by standard box and whisker plot including median horizontal line inside box, box spanning interquartile range, and whiskers extending to highest and lowest value of these eight tumors (one sample T-test p<0.001).


A
T41 whole sequence
Neo
CMV promoter
T7 primer
c-myc epitope
f1 origin
BGH pA
SV40 early promoter
SV40 pA
E6-E7-E1*
TTC33*
E5*
CMV forward primer
T7 promoter
BGH reverse primer
E6-E7-E1* = E6-E7-E1 (truncated; nts 96-879 HPV16)
TTC33* = TTC33 truncated (nts 25,618-25,702/27,488-27,618/39,465-42,142)
E5* = E5 truncated (nts 3850-4101 of HPV16)
pcDNA3.1(+)-myc-His A with T41 parent and daughter constructs
B
E6*I-E7-E4-E5
CMV forward primer
T7 promoter
c-myc epitope
BGH reverse primer
6xHis
E6*I-E7
E4-E5
E6*I
E6*I = E6 truncated (nts 97-226 HPV16)
E6*I-E7 = end of E6, E7, and E1 initial (nt 409-880 HPV16)
E4-E5-pAE = E4 and E5 up to HPV early pA tail (nt 4,215)
pcDNA3.1(+)-myc-His A with HPV16 long read transcript
C
Proliferation in HCT116
Relative absorbance (day2/day0)
3
2.5
2
1.5
1
empty
E6E7
E7
T41whole
E6E7E1
TTC33
E5trunk
E6*IE7E4E5
E6*IE7
E6*I
E4E5
*
D
Proliferarion in NOKSI
Relative absorbance (day2/day0)
2
1.5
1
empty
E6E7
E7
T41whole
E6E7E1
TTC33
E5trunk
E6*IE7E4E5
E6*IE7
E6*I
E4E5
*
E
siFOXE1 Treatment of SCC090
Percent viability
100
80
60
40
20
0
Scramble Control
siFOXE1
*
F
Normalized PDL1 Expression
2.0
1.5
1.0
0.5
0.0
-0.5
TCGA-CV-5443

3.7 References

1. Annual Report to the Nation 2019: Overall Cancer Statistics. National Cancer Institute: Surveillance, Epidemiology, and End Results Program. https://seer.cancer.gov/report_to_nation/statistics.html. Accessed March 1, 2020.

2. 04/22/2020. American Cancer Society. Cancer Facts & Figures 2020. Atlanta: American Cancer Society; 2020. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf>. 04/22/2020.

3. A. K. Chaturvedi, E. A. Engels, R. M. Pfeiffer, B. Y. Hernandez, W. Xiao, E. Kim, B. Jiang, M. T. Goodman, M. Sibug-Saber, W. Cozen, L. Liu, C. F. Lynch, N. Wentzensen, R. C. Jordan, S. Altekruse, W. F. Anderson, P. S. Rosenberg, M. L. Gillison. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J Clin Oncol* 2011;**29**(32):4294-301 doi 10.1200/jco.2011.36.4596.

4. S. V. Graham. Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies. *Future Microbiol* 2010;**5**(10):1493-506 doi 10.2217/fmb.10.107.

5. K. K. Ang, J. Harris, R. Wheeler, R. Weber, D. I. Rosenthal, P. F. Nguyen-Tân, W. H. Westra, C. H. Chung, R. C. Jordan, C. Lu, H. Kim, R. Axelrod, C. C. Silverman, K. P. Redmond, M. L. Gillison. Human papillomavirus and survival of patients with oropharyngeal cancer. *The New England journal of medicine* 2010;**363**(1):24-35 doi 10.1056/NEJMoa0912217.

6. 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, P. S. Mischel. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* 2017;**543**(7643):122-5 doi 10.1038/nature21356.

7. 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* 2019;**575**(7784):699-703 doi 10.1038/s41586-019-1763-5.

8. 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. Nat. Genet. 52, 891–897 (2020). Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* 2020;**52**(9):891-7 doi 10.1038/s41588-020-0678-2.

9. C. Bailey, M. J. Shoura, P. S. Mischel, C. Swanton. Extrachromosomal DNA-relieving heredity constraints, accelerating tumour evolution. *Ann Oncol* 2020;**31**(7):884-93 doi 10.1016/j.annonc.2020.03.303.

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, P. S. Mischel. Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. *Science* 2014;**343**(6166):72-6 doi 10.1126/science.1241328.

11. V. Deshpande, J. Luebeck, N.-P. D. Nguyen, M. Bakhtiari, K. M. Turner, R. Schwab, H. Carter, P. S. Mischel, V. Bafna. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat Commun* 2019;**10**(1):392 doi 10.1038/s41467-018-08200-y.

12. I. J. Groves, N. Coleman. Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us? *J Pathol* 2018;**245**(1):9-18 doi 10.1002/path.5058.

13. I. M. Morgan, L. J. DiNardo, B. Windle. Integration of Human Papillomavirus Genomes in Head and Neck Cancer: Is It Time to Consider a Paradigm Shift? *Viruses* 2017;**9**(8) doi 10.3390/v9080208.

14. T.J. Nulton, A.L. Olex, M. Dozmorov, I.M. Morgan, B. Windle. Analysis of The Cancer Genome Atlas sequencing data reveals novel properties of the human papillomavirus 16 genome in head and neck squamous cell carcinoma. *Oncotarget* 2017;**8**(11):17684-99 doi 10.18632/oncotarget.15179.

15. K. Akagi, J. Li, T. R. Broutian, H. Padilla-Nash, W. Xiao, B. Jiang, J. W. Rocco, T. N. Teknos, B. Kumar, D. Wangsa, D. He, T. Ried, D. E. Symer, M. L. Gillison. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res* 2014;**24**(2):185-99 doi 10.1101/gr.164806.113.

16. A. Holmes, S. Lameiras, E. Jeannot, Y. Marie, L. Castera, X. Sastre-Garau, A. Nicolas. Mechanistic signatures of HPV insertions in cervical carcinomas. *NPJ Genom Med* 2016;**1**:16004 doi 10.1038/npjgenmed.2016.4.

17. H. Cho, J. Davis, X. Li, K. S. Smith, A. Battle, S. B. Montgomery. High-resolution transcriptome analysis with long-read RNA sequencing. *PLoS One* 2014;**9**(9):e108095 doi 10.1371/journal.pone.0108095.

18. A. M. Patch, K. Nones, S. H. Kazakoff, F. Newell, S. Wood, C. Leonard, O. Holmes, Q. Xu, V. Addala, J. Creaney, B. W. Robinson, S. Fu, C. Geng, T. Li, W. Zhang, X. Liang, J. Rao, J. Wang, M. Tian, Y. Zhao, F. Teng, H. Gou, B. Yang, H. Jiang, F. Mu, J. V. Pearson, N. Waddell. Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. *PLoS One* 2018;**13**(1):e0190264 doi 10.1371/journal.pone.0190264.

19. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15-21 doi 10.1093/bioinformatics/bts635.

20. T. Guo, A. Sakai, B. Afsari, M. Considine, L. Danilova, A. V. Favorov, S. Yegnasubramanian, D. Z. Kelley, E. Flam, P. K. Ha, Z. Khan, S. J. Wheelan, J. S. Gutkind, E. J. Fertig, D. A. Gaykalova, J. Califano. A Novel Functional Splice Variant of AKT3 Defined by Analysis of Alternative Splice Expression in HPV-Positive Oropharyngeal Cancers. *Cancer Res* 2017;**77**(19):5248-58 doi 10.1158/0008-5472.can-16-3106.

21. K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, J. Liu. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* 2010;**38**(18):e178 doi 10.1093/nar/gkq622.

22. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;**11**(9):1650-67 doi 10.1038/nprot.2016.095.

23. H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Quantitative Biology* 2013(https://arxiv.org/abs/1303.3997).

24. E. Talevich, A. H. Shain, T. Botton, B. C. Bastian. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* 2016;**12**(4):e1004873 doi 10.1371/journal.pcbi.1004873.

25. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**(1):24-6 doi 10.1038/nbt.1754.

26. S. Ardui, A. Ameur, J. R. Vermeesch, M. S. Hestand. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* 2018;**46**(5):2159-68 doi 10.1093/nar/gky066.

27. N.P. Nguyen, V. Deshpande, J. Luebeck, P. S. Mischel, 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* 2018;**46**(7):3309-25 doi 10.1093/nar/gky180.

28. S. Wu, V. Bafna, P. S. Mischel. Extrachromosomal DNA (ecDNA) in cancer pathogenesis. *Curr Opin Genet Dev* 2021;**66**:78-82 doi 10.1016/j.gde.2021.01.001.

29. S. P. Patel, R. Kurzrock. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Mol Cancer Ther* 2015;**14**(4):847-56 doi 10.1158/1535-7163.Mct-14-0983.

30. R. Plebani, G. R. Oliver, M. Trerotola, E. Guerra, P. Cantanelli, L. Apicella*, et al.* Long-range transcriptome sequencing reveals cancer cell growth regulatory chimeric mRNA. *Neoplasia* 2012;**14**(11):1087-96 doi 10.1593/neo.121342.

31. M. Hosonaga, H. Saya, Y. Arima. Molecular and cellular mechanisms underlying brain metastasis of breast cancer. *Cancer and Metastasis Reviews* 2020;**39**(3):711-20 doi 10.1007/s10555-020-09881-y.

32. Y. Pan, L. Liu, Y. Cheng, J. Yu, Y. Feng. Amplified LncRNA PVT1 promotes lung cancer proliferation and metastasis by facilitating VEGFC expression. *Biochem Cell Biol* 2020:1-7 doi 10.1139/bcb-2019-0435.

33. D. Smotkin, H. Prokoph, F. O. Wettstein. Oncogenic and nononcogenic human genital papillomaviruses generate the E7 mRNA by different mechanisms. *J Virol* 1989;**63**(3):1441-7.

34. M. L. Gillison, K. Akagi, W. Xiao, B. Jiang, R. K. L. Pickard, J. Li, B. J. Swanson, A. D. Agrawal, M. Zucker, B. Stache-Crain, A. K. Emde, H. M. Geiger, N. Robine, K. R. Coombes, D. E. Symer. Human papillomavirus and the landscape of secondary genetic alterations in oral cancers. *Genome Res* 2019;**29**(1):1-17 doi 10.1101/gr.241141.118.

35. M. T. E. Cornelissen, H. L. Smits, M. A. Briet, J. G. Van den Tweel, A. P. H. B. Struyk, J. Van der Noordaa, J. Ter Schegget. Uniformity of the splicing pattern of the E6/E7 transcripts in human papillomavirus type 16-transformed human fibroblasts, human cervical premalignant lesions and carcinomas. *J Gen Virol* 1990;**71 (Pt 5)**:1243-6 doi 10.1099/0022-1317-71-5-1243.

36. S. Tang, M. Tao, J. P. McCoy, Z.-M. Zheng. The E7 oncoprotein is translated from spliced E6*I transcripts in high-risk human papillomavirus type 16- or type 18-positive cervical cancer cell lines via translation reinitiation. *J Virol* 2006;**80**(9):4249-63 doi 10.1128/jvi.80.9.4249-4263.2006.

37. L. Olmedo-Nieva, J. O. Munoz-Bello, A. Contreras-Paredes, M. Lizano. The Role of E6 Spliced Isoforms (E6*) in Human Papillomavirus-Induced Carcinogenesis. *Viruses* 2018;**10**(1) doi 10.3390/v10010045.

38. X. Li, C. Johansson, C. Cardoso Palacios, A. Mossberg, S. Dhanjal, M. Bergvall, S. Schwartz. Eight nucleotide substitutions inhibit splicing to HPV-16 3'-splice site SA3358 and reduce the efficiency by which HPV-16 increases the life span of primary human keratinocytes. *PLoS One* 2013;**8**(9):e72776 doi 10.1371/journal.pone.0072776.

39. M. Somberg, S. Schwartz. Multiple ASF/SF2 sites in the human papillomavirus type 16 (HPV-16) E4-coding region promote splicing to the most commonly used 3'-splice site on the HPV-16 genome. *J Virol* 2010;**84**(16):8219-30 doi 10.1128/jvi.00462-10.

40. C. Johansson, S. Schwartz. Regulation of human papillomavirus gene expression by splicing and polyadenylation. *Nat Rev Microbiol* 2013;**11**(4):239-51 doi 10.1038/nrmicro2984.

41. C. E. Vaisman, O. Del Moral-Hernandez, S. Moreno-Campuzano, E. Aréchaga-Ocampo, R. Bonilla-Moreno, I. Garcia-Aguiar, L. Cedillo-Barron, J. Berumen, P. Nava, N. Villegas-Sepúlveda. C33-A cells transfected with E6*I or E6*II the short forms of HPV-16 E6, displayed opposite effects on cisplatin-induced apoptosis. *Virus Res* 2018;**247**:94-101 doi 10.1016/j.virusres.2018.02.009.

42. I. Nindl, K. Rindfleisch, B. Lotz, A. Schneider, M. Dürst. Uniform distribution of HPV 16 E6 and E7 variants in patients with normal histology, cervical intra-epithelial neoplasia and cervical cancer. *Int J Cancer* 1999;**82**(2):203-7 doi 10.1002/(sici)1097-0215(19990719)82:2<203::aid-ijc9>3.0.co;2-9.

43. M. G. Brattain, W. D. Fine, F. M. Khaled, J. Thompson, D. E. Brattain. Heterogeneity of malignant cells from a human colonic carcinoma. *Cancer Res* 1981;**41**(5):1751-6.

44. D. Martin, M. C. Abba, A. A. Molinolo, L. Vitale-Cross, Z. Wang, M. Zaida, N. C. Delic, Y. Samuels, J. G. Lyons, J. S. Gutkind. The head and neck cancer cell oncogenome: a platform for the development of precision molecular therapies. *Oncotarget* 2014;**5**(19):8906-23 doi 10.18632/oncotarget.2417.

45. J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J. Y. Yang, B. M. Broom, R. G. W. Verhaak, D. W. Kane, C. Wakefield, J. N. Weinstein, G. B. Mills, H. Liang. TCPA: a resource for cancer functional proteomics data. *Nat Methods* 2013;**10**(11):1046-7 doi 10.1038/nmeth.2650.

46. M. Parfenov, C. S. Pedamallu, N. Gehlenborg, S. S. Freeman, L. Danilova, C. A. Bristow, S. Lee, A. G. Hadjipanayis, E. V. Ivanova, M. D. Wilkerson, A. Protopopov, L. Yang, S. Seth, X. Song, J. Tang, X. Ren, J. Zhang, A. Pantazi, N. Santoso, A. W. Xu, H. Mahadeshwar, D. A. Wheeler, R. I. Haddad, J. Jung, A. I. Ojesina, N. Issaeva, W. G. Yarbrough, D. N. Hayes, J. R. Grandism, A. K. El-Naggar, M. Meyerson, P. J. Park, L. Chin, J. G. Seidman, P. S. Hammerman, R. Kucherlapati. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc Natl Acad Sci U S A* 2014;**111**(43):15544-9 doi 10.1073/pnas.1416074111.

47. A. De Leo, A. Calderon, P. M. Lieberman. Control of Viral Latency by Episome Maintenance Proteins. *Trends Microbiol* 2020;**28**(2):150-62 doi 10.1016/j.tim.2019.09.002.

48. L. Jose, E. J. Androphy, M. DeSmet. Phosphorylation of the Human Papillomavirus E2 Protein at Tyrosine 138 Regulates Episomal Replication. *J Virol* 2020;**94**(14) doi 10.1128/JVI.00488-20.

49. N. Häfner, C. Driesch, M. Gajda, L. Jansen, R. Kirchmayr, I. B. Runnebaum, M. Dürst. Integration of the HPV16 genome does not invariably result in high levels of viral oncogene transcripts. *Oncogene* 2008;**27**(11):1610-7 doi 10.1038/sj.onc.1210791.

50. D. Hong, J. Liu, Y. Hu, X. Lu, B. Li, Y. Li, D. Hu, W. Lu, X. Xie, X. Cheng. Viral E6 is overexpressed via high viral load in invasive cervical cancer with episomal HPV16. *BMC Cancer* 2017;**17**(1):136 doi 10.1186/s12885-017-3124-9.

51. P. Paget-Bailly, K. Meznad, D. Bruyère, J. Perrard, M. Herfs, A. C. Jung, C. Mougin, J. L. Prétet, A. Baguet. Comparative RNA sequencing reveals that HPV16 E6 abrogates the effect of E6*I on ROS metabolism. *Sci Rep* 2019;**9**(1):5938 doi 10.1038/s41598-019-42393-6.

52. D. Pim, P. Massimi, L. Banks. Alternatively spliced HPV-18 E6* protein inhibits E6 mediated degradation of p53 and suppresses transformed cell growth. *Oncogene* 1997;**15**(3):257-64 doi 10.1038/sj.onc.1201202.

53. M. Filippova, M. M. Johnson, M. Bautista, V. Filippov, N. Fodor, S. S. Tungteakkhun, K. Williams, P. J. Duerksen-Hughes. The large and small isoforms of human papillomavirus type 16 E6 bind to and differentially affect procaspase 8 stability and activity. *J Virol* 2007;**81**(8):4116-29 doi 10.1128/jvi.01924-06.

54. B. Roggenbuck, P. M. Larsen, S. J. Fey, D. Bartsch, L. Gissmann, E. Schwarz. Human papillomavirus type 18 E6*, E6, and E7 protein synthesis in cell-free translation systems and comparison of E6 and E7 in vitro translation products to proteins immunoprecipitated from human epithelial cells. *J Virol* 1991;**65**(9):5068-72.

55. C. Artaza-Irigaray, A. Molina-Pineda, A. Aguilar-Lemarroy, P. Ortiz-Lazareno, L. P. Limón-Toledo, A. L. Pereira-Suárez, W. Rojo-Contreras, L. F. Jave-Suárez. E6/E7 and E6(*) From HPV16 and HPV18 Upregulate IL-6 Expression Independently of p53 in Keratinocytes. *Front Immunol* 2019;**10**:1676 doi 10.3389/fimmu.2019.01676.

56. V. M. Williams, M. Filippova, V. Filippov, K. J. Payne, P. Duerksen-Hughes. Human papillomavirus type 16 E6* induces oxidative stress and DNA damage. *J Virol* 2014;**88**(12):6751-61 doi 10.1128/jvi.03355-13.

57. S. Kosel, S. Burggraf, W. Engelhardt, B. Olgemoller. Increased levels of HPV16 E6*I transcripts in high-grade cervical cytology and histology (CIN II+) detected by rapid real-time RT-PCR amplification. *Cytopathology* 2007;**18**(5):290-9 doi 10.1111/j.1365-2303.2007.00481.x.

58. D. Martinez-Zapien, F. X. Ruiz, J. Poirson, A. Mitschler, J. Ramirez, A. Forster, A. Cousido-Siah, M. Masson, S. Vande Pol, A. Podjarny, G. Travé, K. Zanier. Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. *Nature* 2016;**529**(7587):541-5 doi 10.1038/nature16481.

59. N. Hemmat, H. B. Baghi. Human papillomavirus E5 protein, the undercover culprit of tumorigenesis. *Infect Agent Cancer* 2018;**13**:31 doi 10.1186/s13027-018-0208-3.

60. S. Miyauchi, P. D. Sanders, K. Guram, S. S. Kim, F. Paolini, A. Venuti, E. E. W. Cohen, J. S. Gutkind, J. A. Califano, A. B. Sharabi. HPV16 E5 Mediates Resistance to PD-L1 Blockade and Can Be Targeted with Rimantadine in Head and Neck Cancer. *Cancer Res* 2020;**80**(4):732-46 doi 10.1158/0008-5472.Can-19-1771.

61. S. Ren, D. A. Gaykalova, T. Guo, A. V. Favorov, E. J. Fertig, P. Tamayo, J. L. Callejas-Valera, M. Allevato, M. Gilardi, J. Santos, T. Fukusumi, A. Sakai, M. Ando, S. Sadat, C. Liu, G. Xu, K. M. Fisch, Z. Wang, A. A. Molinolo, J. S. Gutkind, T. Ideker, W. M. Koch, J. A. Califano. HPV E2, E4, E5 drive alternative carcinogenic pathways in HPV positive cancers. *Oncogene* 2020;**39**(40):6327-39 doi 10.1038/s41388-020-01431-8.

62. N. Wentzensen, R. Ridder, R. Klaes, S. Vinokurova, U. Schaefer, M. Doeberitz. Characterization of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions. *Oncogene* 2002;**21**(3):419-26 doi 10.1038/sj.onc.1205104.

63. Y. Zhu, A. D. Gujar, C. H. Wong, H. Tjong, C. Y. Ngan, L. Gong, Y. A. Chen, H. Kim, J. Liu, M. Li, A. Mil-Homens, R. Maurya, C. Kuhlberg, F. Sun, E. Yi, A. C. deCarvalho, Y. Ruan, R. G. W. Verhaak, C. L. Wei. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell* 2021;**39**(5):694-707.e7 doi 10.1016/j.ccell.2021.03.006.

64. J. M. Molkentine, D. P. Molkentine, K. A. Bridges, T. Xie, L. Yang, A. Sheth, T. P. Heffernan, D. A. Clump, A. Z. Faust, R. L. Ferris, J. N. Myers, M. J. Frederick, K. A. Mason, R. E. Meyn, C. R. Pickering, H. D. Skinner. Targeting DNA damage response in head and neck cancers through abrogation of cell cycle checkpoints. *Int J Radiat Biol* 2020:1-8 doi 10.1080/09553002.2020.1730014.

65. L. Zeng, A. Nikolaev, C. Xing, D. L. Della Manna, E. S. Yang. CHK1/2 Inhibitor Prexasertib Suppresses NOTCH Signaling and Enhances Cytotoxicity of Cisplatin and Radiation in Head and Neck Squamous Cell Carcinoma. *Mol Cancer Ther* 2020;**19**(6):1279-88 doi 10.1158/1535-7163.Mct-19-0946.

66. N. Tanaka, A. A. Patel, J. Wang, M. J. Frederick, N. N. Kalu, M. Zhao, A. L. Fitzgerald, T. X. Xie, N. L. Silver, C. Caulin, G. Zhou, H. D. Skinner, F. M. Johnson, J. N. Myers, A. A. Osman. Wee-1 Kinase Inhibition Sensitizes High-Risk HPV+ HNSCC to Apoptosis Accompanied by Downregulation of MCl-1 and XIAP Antiapoptotic Proteins. *Clin Cancer Res* 2015;**21**(21):4831-44 doi 10.1158/1078-0432.Ccr-15-0279.

CHAPTER 4: Extrachromosomal DNA is a driver of the malignant transformation from Barrett's esophagus to adenocarcinoma

4.1 Introduction

Esophageal adenocarcinoma (EAC) arises from a precursor lesion, Barrett's esophagus (BE) (1,2). BE may carry many genomic abnormalities, however somatic genomic copy number changes frequently precede and predict malignancy in BE, often by multiple years (3). Somatic copy number changes come in many forms. Prior work has established one particular form of copy number change, the presence of focal somatic copy number amplifications (fsCNAs) of up to 10 Mbp, to be associated with malignancy, particularly when oncogenes are involved (4,5). Oncogene-bearing fsCNA represent a hallmark of the cancer genome (4,6–10).

FsCNA may arise by many related mechanisms including, but not limited to, extrachromosomal circular DNA (ecDNA) (11–13), breakage-fusion-bridge (BFB) cycles (14,15), chromothripsis (16,17), tandem short template jumps (18), and other forms of genomic instability which enable the formation of double-stranded DNA breaks (19).

Circular extrachromosomal DNA (ecDNA) is a frequent cause of oncogene amplification (5), epigenetic and transcriptional dysregulation (20,21), resulting in more aggressive tumors with poorer patient survival, regardless of cancer type (5). While previous studies have characterized the frequency of ecDNA in extant tumors, cell lines, and even have induced the formation and evolution of ecDNA in cancer cell lines by drug application (22), less is known about its natural origins in tumors or its presence in pre-cancerous tissue. Whether ecDNA represents an event occurring early or later in tumorigenesis is incompletely understood, as is ecDNA's role as a passenger or a driver during malignant transformation.

Barrett’s esophagus presents a unique opportunity to study the origins of ecDNA, as BE tissue can exist in patients for decades prior to malignant transformation into EAC (3). However, the progression of BE to EAC is relatively uncommon in patients (0.3% per annum), and current standards of care involve the use of histopathology to inform intervention (typically endoscopic resection or radiofrequency ablation). Barrett’s lesions are both clonal and typically very diverse in their genomic landscape. As a result, the molecular phenotypes associated with worsened histological states are incompletely understood in BE.

To begin to better understand the origins of ecDNA, we analyzed 572 Barrett’s esophagus and esophageal adenocarcinoma samples derived from 306 cancer outcome and non-cancer outcome patients, from multiple independent cohorts. We utilized data from the Fred Hutchinson Cancer Research Center (FHCRC, Seattle, USA), the Medical Research Council Cancer Unit (MRCCU, Cambridge, UK), as well as EAC tumors in the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) databases and examined the role of ecDNA in malignant transformation of BE.

4.2 Results

4.2.1 FHCRC cohort

To understand the role of ecDNA in malignant transformation, we investigated a BE cohort from the Fred Hutchinson Research Center (FHCRC). The FHCRC cohort was derived from a BE surveillance study. Biopsies were primarily taken from two retrospectively assigned time-points, referred to as T1 and T2, and WGS data was generated for two levels of biopsies in the esophagus at each time-point (Figure 4.1a). Forty patients who developed esophageal adenocarcinoma were assigned a “cancer outcome” (CO) label. Among the CO patients, T2 was the endoscopic sampling time-point where cancer was discovered. Forty patients who did not develop esophageal adenocarcinoma were assigned a “non-cancer outcome” (NCO) label. The distribution of time intervals between T1 and T2 were designed to be similar between NCO and CO patients (median 2.8 years versus 2.1 years, respectively, Mann-Whitney U test, p-value=0.057, test statistic=635.5, H_a=less, Figure 4.2).

In total, the FHCRC dataset included 320 esophageal WGS samples collected at time-points T1 and T2 from CO and NCO patients (Figure 4.1b). In addition, 20 long-term follow-up samples from 10 NCO patients (median 9.6 years after T2) were collected, and at least one matched normal sample was collected for each patient, including 62 normal blood and 25 normal gastric samples - resulting in 427 sequencing samples with an average sequencing depth of 74x for esophageal samples and 38x for normals.

The FHCRC BE samples were prepared and analyzed by the following steps (Figure 4.1c); crypt isolation to separate BE tissue from the basement membrane, WGS of the isolated crypts, CNV calling on aligned WGS data (Methods – “FHCRC cohort data generation”), identification of seed regions for focal amplification analysis. AmpliconArchitect has recently been used to identify extrachromosomal DNA from mapped WGS reads with high precision and recall (5, 23). Therefore, focal amplifications were characterized using AmpliconArchitect (Methods – “FHCRC cohort focal amplification detection”).

Histology and sequencing biopsies were collected separately in the FHCRC study (Figure 4.1d). To match the two types of biopsies we defined both an on-level histology and a windowed (+/-1cm) histology based on matching the distance (or “height”) of the biopsies from the gastroesophageal junction (GEJ). In cases where multiple histology biopsies existed on the same level or in the same window, the highest disease-state was considered.

In total, we identified ecDNA in 13/40 (33%) of FHCRC CO patients, and 1/40 (3%) of FHCRC NCO patients, identifying ecDNA at multiple time-points in 7/14 (50%) of ecDNA+ patients (Figure 4.1e), including in 8 patients with pre-cancerous tissue but no detectable cancer. The single NCO patient with ecDNA (patient 303), passed away 2.5 years after the last endoscopy (T2). While the patient is characterized as NCO, it is unknown whether the patient developed EAC in the two years prior to death, and the patient is denoted by a gold star in the study. Across all four biopsies from either time-point, we found 8/14 (57%) ecDNA-positive patients who had ecDNA in multiple biopsies. EcDNA-positive patients showed a statistically significant increase in maximum histology between T1 to T2 (Figure 4.1f, Wilcoxon signed-rank test, p-value=4.7x10^{-4}, test statistic=1.5, H$_a$=less). In contrast, we identified no ecDNA in matched normal blood or gastric samples from NCO and CO (Figure 4.3a-b), nor did we identify any ecDNA in the 20 NCO long-term follow-up samples (Figure 4.3c).

We examined associated histology data in the FHCRC cohort to understand its relationship with ecDNA status in both time-points. At time-point T1, where none of the CO patients had been diagnosed with cancer, we found that 100% (6/6) of ecDNA-positive biopsies showed high grade dysplasia (HGD) in their on-level histology. By contrast, in 54% (25/46) of the ecDNA-negative biopsies, HGD was not detected (Figure 4.1g, odds ratio (OR)=15.4, Fisher’s exact test, p-value=0.015, H$_a$=greater). Examining the CO patients again at T2, where cancer was first identified in the patients, 82% (9/11) of ecDNA-positive biopsies had EAC in the on-level histology, and the remaining biopsies associated with HGD. Among ecDNA-negative biopsies, however, only 47% (20/43) were associated with EAC, and 21% (9/43) remained associated with BE only (Figure 4.1h, OR=5.2, Fisher’s exact test, p-value=0.037, H$_a$=greater).

Genomic amplifications are a known driver of EAC progression from BE (3, 24). Grouping FHCRC NCO and CO samples at time-point T2 and using histology assigned from a windowed approach, we found that non-ecDNA fsCNA were associated with a higher incidence of HGD or EAC (Figure 1i, OR=4.7, Fisher's exact test, p-value=1.1×10^{-4}, H_a=greater). However, all ecDNA-positive samples at that time-point associated even more strongly with HGD or EAC (OR=29.4, Fisher's exact test, p-value=2.0×10^{-4}, H_a=greater), illustrating that focal amplification type has a profound effect on worsened disease state.

4.2.2 MRCCU cohort

We also sought to characterize ecDNA in an independent, curated cross-sectional group of BE & early EAC patients. The MRCCU study involved manual curation of samples from 117 patients in the United Kingdom (UK). Based on highest disease stage per patient as assessed by histology (Figure 4.4a, Methods – "MRCCU sample selection"), we stratified patients into three groups: 1) 42 patients whose worst disease state never progressed past non-dysplastic Barrett's esophagus (mean follow-up length 6.2 years). 2) 25 patients who developed HGD and never progressed past that point due to interventional therapy. 3) 50 patients who developed BE-adjacent EAC. The BE-only category of the MRCCU cohort is similar to FHCRC NCO patients with BE as the maximum histology, while the HGD category of the MRCCU cohort is similar to a combination of FHCRC CO (time-point T1) and NCO (time-point T2) patients who have HGD as the maximum histology. The MRCCU EAC category is most similar to FHCRC CO patients at T2. The biopsies or resection samples characterized at the patient's most severe histological state were the same samples used to generate WGS data with 50X average depth (Methods – "MRCCU cohort sequencing data"). The PrepareAA pipeline was used to identify seed regions, with AA to characterize the architecture of the focal amplifications (Figure 4.4b, Methods – "MRCCU and ICGC focal amplification detection").

In the MRCCU BE-only (without progression) category, we found no ecDNA in 42 samples (Figure 4.4c), while in the MRCCU HGD category, where the patients subsequently underwent therapeutic intervention shortly after the first identification of HGD, we identified 1 ecDNA+ patient out of 25 total (4%) (Figure 2c). In the BE-adjacent EAC category we identified 12 ecDNA+ patients out of 50 (24%), further reinforcing that ecDNA is not a characteristic of non-progressing Barrett's esophagus, but instead is found commonly in pre-cancers undergoing malignant transformation and is also found in early EAC tumors.

We associated ecDNA status and other (non-ecDNA) fsCNA status with the matched histological findings (Figure 4.4d) to determine whether ecDNA presence was associated with worsened histology in the matched WGS/histology samples in the MRCCU cohort. When compared to samples without any fsCNA, samples with non-ecDNA fsCNA associated with higher-grade (HGD or EAC) histological status (OR=11.8, Fisher's exact test, p-value=7.5x10^{-5}, H$_a$=greater). However, ecDNA-positive samples, when compared to ecDNA-negative samples with both being agnostic to other fsCNAs, showed an even stronger association with higher-grade histological state (OR=18.4, Fisher's exact test, p-value=2.0x10^{-3}, H$_a$=greater).

4.2.3 Association of ecDNA and other genomic lesions with histology and cancer outcome

To understand the relationship between genome instability and the origin of focal amplifications in the pre-cancer samples and early EACs, we examined related genomic features, including breakage-fusion bridge (BFB) presence, non-ecDNA & non-BFB fsCNA, *TP53* gene disruption, whole-genome duplication (WGD) status, and chromothripsis presence in both CO patients (Figure 4.5) and NCO patients (Figure 4.6). We performed a similar analysis in the MRCCU patients as well, characterizing BFB status, other fsCNA (non-ecDNA & non-BFB), and *TP53* gene disruption (Figure 4.7).

Profiling the focal amplification types in both FHCRC and MRCCU cohorts revealed multiple types of focal amplifications had strong associations with worsened histological status (Figure 4.8). In the FHCRC cohort time-point T2 samples, when ecDNA was present in the windowed histology it was always associated with HGD or worse (OR=29.4) and in MRCCU samples when ecDNA was present was also always associated with HGD or worse (OR=18.4).

Prior loss of *TP53*, even partially, enables the development of genomic instability (25, 26), and we found in both FHCRC and MRCCU cohorts a strong association between *TP53* disruption and ecDNA-positive status (Fisher's exact test, p-values 1.1×10^{-4} and 0.023, respectively for FHCRC and MRCCU, H_a=greater), as well as between *TP53* disruption and BFB-positive status (Figure 4.9a-d, Fisher's exact test, p-values=4.8×10^{-6} and 1.8×10^{-4} respectively, H_a=greater) – despite different methodologies being used to analyze the status of *TP53* in the two cohorts (Methods – "*TP53* disruption analysis").

In the FHCRC data we found that *TP53* disruption had a statistically significant association with increased frequency of WGD (Figure 4.10a, Fisher's exact test, p-value=1.4×10^{-21}, H_a=greater) as well as a statistically significant association with increased frequency of chromothripsis (Figure 4.10b, Fisher's exact test, p-value=8.2×10^{-7}, H_a=greater). EcDNA, non-ecDNA fsCNA, WGD, chromothripsis all preferentially occurred in *TP53* disrupted samples, suggesting that *TP53* disruption frequently precedes those events. We subsequently conditioned on *TP53* disruption and examined the association of ecDNA with WGD, and also examined the association of ecDNA with chromothripsis in FHCRC patients. We found that while ecDNA was more frequent in WGD samples (17% in WGD-negative, 28% in WGD-positive) and more frequent in chromothriptic samples (19% chromothripsis-negative vs 33% chromothripsis-positive), there was not a statistically significant relationship between increased frequency of ecDNA and WGD or chromothripsis (Figure 4.10c-d, Fisher's exact test, p-values=0.14 and 0.12, respectively, H_a=greater). It is also important to note that *TP53* disruption was not found to be an exclusive phenomenon in cancer-outcome patients (9/40, 23% in FHCRC NCO - Figure 4.6, 8/42 19% in MRCCU BE, Figure 4.7). Together these results indicate that ecDNA may by a number of different mechanisms and the destabilizing effect of prior *TP53* disruption on the genome is tightly linked to ecDNA formation.

We compared ecDNA occurrence to the presence of other genomic lesions in the FHCRC data to assess the specificity with which they associated with cancer outcome status. We assessed FHCRC pre-cancer samples from NCO patients at time-points T1 & T2 (patient 303 censored due to poor survival) and pre-cancer CO patients at time-point T1 (Supplemental Figure 9a). Compared to ecDNA status (OR = 17.7), we found that WGD, *TP53* disruption, non-ecDNA fsCNA, and chromothripsis, showed less-specific association with cancer outcome (ORs=8.1, 4.6, 4.5, 2.1, respectively). When restricting the analysis solely to FHCRC CO patients at pre-cancer time-point T1 (Figure 4.11b), we found ecDNA and chromothripsis had the strongest association with high-grade dysplasia (OR = 15.4 in both), as compared to WGD, *TP53* disruption, and non-ecDNA fsCNA presence (ORs = 14.1, 4.2, 2.4, respectively).

4.2.4 Tracking clonal ecDNA

The longitudinal, multiregional biopsy collection in the FHCRC cohort allowed us to track ecDNA across multiple time-points in the same individual. To quantify the similarities between two genomically overlapping focal amplifications (Figure 4.12a), we developed an amplicon similarity score ranging from 0 to 1, where 0 = no similarity and 1 = identical, to quantify overlaps between amplified genomics coordinates as well as overlaps between breakpoint junctions (Figure 4.12b, Methods - “Amplicon similarity score”, Figure 4.13a-d). The amplicon similarity score can be computed in two ways – a directional similarity, and a symmetric similarity (Figure 4.12c). Importantly, the directional similarity quantifies one amplicon’s similarity as a subset of regions and breakpoints identified in the other, while the symmetric similarity score represents the mean of both directional similarities. A key benefit of the directional similarity is that it enables quantification of the similarity of genomic regions and breakpoints from a more diverged amplicon to parent amplicon. This enabled a quantitative analysis to determine if two amplicons arose from a common origin, where the parental amplicon’s genomic segments and breakpoints formed a subset of those found in the other amplicon.

In the symmetric similarity scoring, the statistical significance of the overlap between two amplicons is computed against a background panel of genomically overlapping amplicons from unrelated, independent origins. We fit the empirical distribution of overlapping unrelated amplicon similarity scores with a beta distribution using a maximum likelihood estimation approach to create a robust model from which to assess statistical significance of overlaps.

Applying the amplicon similarity scoring to multiple focal amplifications from the same patient, we found that overlapping focal amplification had significantly higher similarity scores than overlapping focal amplifications from different patients (Figure 4.12d, Mann-Whitney U test, p-value=$1.2\text{x}10^{-24}$, test statistic=771.5, H_a=greater). We next identified clonal ecDNA and other focal amplifications present at multiple time-points in pre-cancer and EAC samples. Importantly, when genomically overlapping ecDNA amplicons were detected in multiple samples from the same patient, the ecDNA amplicons shared high similarity, evaluated against the null distribution of overlapping focal amplifications from independent origins in three different studies (Methods – “Amplicon similarity score”, Figure 4.12e). Of 11 genomically overlapping ecDNA pairs from the same patients, 10 fell in the 95th percentile or higher of the null distribution model. This suggests that ecDNA detected in pre-cancer are frequently maintained through the transition to cancer. Furthermore, when overlapping ecDNA are identified in multi-region sampling of BE tissue or EAC tumors, they are frequently resulting from a common origin.

To track the emergence of ecDNA even more comprehensively, in FHCRC CO patient 391 (Figure 4.12f), we leveraged 6 additional biopsies, beyond the four collected at T1 and T2, and identified two distinct ecDNA species which appeared in multiple samples within and across time-points (Figure 4.12g). While the patient had high-grade dysplasia in multiple regions in the first endoscopy, we did not detect ecDNA in nearby WGS samples. However, from the second surveillance endoscopy, performed 5.6 years later, we identified ecDNA (ecDNA-1) in the one sample submitted for WGS with high-grade dysplasia in the windowed histology. From the third surveillance endoscopy performed 6.5 months later, we identified two distinct species of ecDNA (ecDNA-1 and ecDNA-2). While ecDNA-1 appeared at multiple levels of the esophagus (Figure 4.12h), including on levels associated with both HGD and EAC, it did not carry canonical oncogenes. On the other hand, ecDNA-2 carried three canonical oncogenes, *RIM2*, *SOCS1*, and *CIITA* (Figure 4.12i), and was associated with EAC histology.

EcDNA-1 appeared in six biopsies/resections, and the mean similarity score of the ecDNA across all fifteen possible pairings was 0.935 (null beta distribution p-value=1.3×10^{-9}). EcDNA-2 appeared in three biopsies/resections and the mean similarity score of the ecDNA across all three possible pairings was 0.865 (null beta distribution p-value=$3.7x10^{-7}$), demonstrating the clonality of both species.

We additionally highlighted two patients with recurrently detected oncogene-carrying ecDNA. In these cases, ecDNA was first detected at an earlier disease state than the EAC-state associated with the two ecDNA in patient 391. In FHCRC CO patient 169, we identified a candidate focal amplification on chr18 in a time-point T1 WGS sample with HGD in the histology window (Figure 4.14a). Analysis of the focal amplification region identified an ecDNA carrying *GATA6* (Figure 4.14b). A second WGS sample at time-point T2, taken at the same distance from the GEJ (3cm) 2.74 years later revealed a focal amplification in the same region with EAC in the sample's histology window (Figure 4.14c). Analysis of the region also revealed an identical 2.23 Mbp ecDNA (Figure 4.14d-e). This ecDNA appeared in three of four biopsies from patient 169 and the mean similarity score of the ecDNA between all three pairings was 1.000 (null beta distribution p-value=$1.1x10^{-16}$). In a second example from FHCRC CO patient 740, we identified a candidate focal amplification on chr6 in a time-point T1 WGS sample with HGD in the histology window (Figure 4.14f). Analysis of the focal amplification region identified a low CN ecDNA carrying *POU5F1*, *HMGA1* and *PIM1* (Figure 4.14g). A second WGS sample at time-point T2, taken at the same distance from the GEJ (2cm) 1.08 years later revealed a focal amplification in the same region with EAC in the sample's histology window (Figure 4.14h). Analysis of the region revealed the same oncogenes with increased copy number as compared to the first time-point (Figure 4.14i), and we recovered a conserved core ecDNA cycle from the focal amplification carrying *HMGA1* and having length 1.22 Mbp seen in both samples from time-point T1 and time-point T2 (Figure 4.14j, ecDNA region similarity score = 0.729, null beta distribution p-value=$7.9x10^{-5}$). Taken together, the tracking of ecDNA during malignant transformation demonstrates ecDNA formation is a truncal event.

4.2.5 The characteristics of ecDNA in BE and BE-derived EAC

Barrett's esophagus is thought to be the precursor of all esophageal adenocarcinomas (27), and thus we sought to also understand the relationship between ecDNA appearing in BE and EAC. To improve power, we additionally analyzed 109 esophageal carcinoma samples from ICGC (89 samples) and TCGA (20 samples).

We identified increasing rates of ecDNA in patients from the MRCCU EAC cohort (24%), FHCRC CO cohort (33%), and the combined ICGC & TCGA cohorts (43%) (Figure 4.15a), despite the different cohorts representing different levels of surveillance and sampling density. We found that 26/84 (31%) of combined BE & EAC ecDNA-positive samples in our study contained more than one distinct species of ecDNA (Figure 4.15b), suggesting that ecDNA heterogeneity and competition between multiple distinct ecDNA may play a role in EAC evolution.

We found that the copy number of ecDNA in EAC is significantly higher than in pre-cancer (Mann-Whitney U test, p-value=0.035, test statistic=503.0, H_a=greater), indicating positive selection of ecDNA during the malignant transformation (Figure 4.15c). Furthermore, we found that while the length of ecDNA were not significantly different between BE and EAC states (Figure 4.16) (Mann-Whitney U test, p-value=0.12, test statistic=585.0, H_a=less), the complexity of structural rearrangements in ecDNA-derived regions of the genome increased between pre-cancer and EAC (Figure 4.15d, Mann-Whitney U test, p-value=0.024, test statistic=429.0, H_a=less). The results highlighted the impact of positive selection in mediating continued evolution and increasing heterogeneity of ecDNA during the malignant transformation.

We also characterized the gene contents of the ecDNAs, and their copy number changes during disease progression. EcDNA tended to have a significantly larger number of oncogenes than non-ecDNA focal amplifications (1.18 oncogenes per ecDNA versus, 0.96 oncogenes per non-ecDNA fsCNA, Mann-Whitney U test, p-value=0.024, test statistic=28258.5, H_a=greater). We examined the changing copy number of ecDNA-borne genes in the clonal, recurrently detected ecDNA from the FHCRC study, across time-points T1 and T2. We found that five of the six recurrently detected FHCRC ecDNA, had at least one gene which increased in copy number between the two time-points (Wilcoxon signed-rank test, p-value=0.023, test statistic=1.0, H_a=less) (Figure 4.15e). In all samples analyzed in our study, when compared to non-ecDNA focal amplifications, we found that ecDNA permitted a greater maximum oncogene copy number than non-ecDNA focal amplifications (Figure 4.15f, Mann-Whitney U test, p-value=1.1×10^{-4}, test statistic=3922.5, H_a=greater), with some genes surpassing a CN > 100. When considering the highest CN oncogene per ecDNA, ecDNA also permitted greater diversity in maximum copy number (non-ecDNA fsCNA oncogene CN variance = 130.4, ecDNA oncogene CN variance = 739.3, Levene's test, p-value=1.6×10^{-4}, test statistic=14.7). Interestingly, across all samples, the highest CN oncogene in ecDNA tended to be *ERBB2* – one of the most frequently amplified oncogenes in BE and EAC (24, 28, 29).

While ecDNA-borne oncogenes were frequently amplified in a recurrent fashion, we noted that in ecDNA-positive tumors, a significant fraction of the ecDNA-borne oncogenes were not detected on non-ecDNA fsCNA (Figure 4.15g), suggesting that ecDNA may permit a wider variety of oncogene amplifications than other mechanisms such as BFB. We found that ecDNA carried 70 unique oncogenes across 122 distinct ecDNA (0.57 unique genes per ecDNA), while non-ecDNA fsCNA carried 149 unique oncogenes across 419 distinct amplicons (0.36 unique genes per non-ecDNA fsCNA). We also found that despite the high diversity of oncogenes, the genomic intervals amplified on ecDNA were non-random. We compared ecDNA-derived regions to oncogene intervals known to associate with BE and EAC, and found a significant overlap, once again suggesting that oncogene-containing ecDNAs are positively selected (ISTAT interval overlap test, $p=1.7\times10^{-4}$, Figure 4.15h, Figure 4.17). Among oncogenes identified in all cohorts, we observed a number of canonical BE- and EAC-associated oncogenes amplified on ecDNA (Figure 4.15i).

We asked whether ecDNA was more common in early- or late-stage EAC tumors. Using the AJCC cancer stage criteria we found that ecDNA was not significantly more enriched in the 28 stage I-II EACs from ICGC & TCGA (15/28 ecDNA+, 54%) than in the 28 stage III-IV EACs (11/38 ecDNA+, 39%) in that group (Fisher's exact test, p-value=0.21, H_a=greater). However, we did find that there was a significantly greater fraction of ecDNA-positive EACs in ICGC & TCGA than in BE pre-cancers fated to become cancer (FHCRC CO T1, 7/40, 18% vs. MRCCU, ICGC & TCGA EACs, 59/159, 37%, respectively, Fisher's exact test, p-value=0.013, H_a=greater), indicating that ecDNA formation may occur continually though tumorigenesis. Additionally, we found ecDNA-positive status had a significant association with tumor invasion, as scored by AJCC T-status criteria (Fisher's exact test, p-value=0.019, H_a=greater) and ecDNA was found more frequently in moderately- to severely-invasive tumors (Figure 4.15j).

4.3 Discussion

While genomic amplifications are known to predict cancer outcome (3), we found that not all focal amplifications are equally associated with worsened disease state. Our study identified extrachromosomal DNA as an early driver of malignant transformation in esophageal adenocarcinoma, detectable not only in mid- to late-stage tumors but also in microscopic early-stage tumors, and frequently in pre-cancerous tissue. We demonstrated these findings through an analysis of patients from multiple independent cohorts, utilizing data from the Fred Hutchinson Cancer Research Center (USA), the Medical Research Council Cancer Unit (UK), as well as EAC tumors in the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) databases. Analysis of the contents of ecDNA in Barrett's and Barrett's-derived cancers demonstrates ecDNA as a potent medium for the focal amplification of a wide variety of oncogenes.

In all samples, we considered associated histology data, linking ecDNA to more severe histological states. We included data derived from longitudinal surveillance of Barrett's patients and included cross-sectional data representing patients with multiple distinct histological states to provide a comprehensive survey of the genomic and histological associations of ecDNA formation. Whenever ecDNA appeared in the FHCRC cohort, it was associated with worsened histological state than non-ecDNA samples, demonstrating that ecDNA in pre-cancer is not a genomic curiosity, but rather an important genomic lesion which drives malignant changes in pre-malignant tissue phenotypes.

Through longitudinal, multiregional sampling of individual patients, we identified oncogene carrying ecDNA which were conserved between pre-cancer and cancer states, demonstrating the truncal nature of ecDNA in tumor development. Our finding that ecDNA copy numbers were increased in cancer versus pre-cancer suggests continued positive selection of ecDNA in tumors.

The strong association of ecDNA to more severe histology, the preferential amplification of oncogenes on ecDNA, and the rapid evolution of ecDNA copy number and structure all point to ecDNA as a critical indicator for intervention in Barrett's derived esophageal adenocarcinoma and BE-associated HGD. We propose that these same principles of ecDNA as a frequent early driver of malignancy apply broadly to other cancer types. Many BE-associated ecDNA contain highly similar sets of oncogenes as found in the ecDNA of other cancers, likely conferring highly similar benefits to tumor-cell fitness. We believe it is highly likely that ecDNA similarly exists in other pre-cancers, ultimately acting as a key driver of the malignant transformation regardless of tissue type.

4.4 Methods

4.4.1 FHCRC study data

The full methodology for the study design and data generation from the FHRCR cohort is presented in Paulson et al. 2021. In brief, surveillance followed the Seattle protocol. Two groups of 40 patients with cancer and non-cancer outcomes during the study period were identified, where inclusion criteria required 2 or more endoscopies where BE segments of 2 cm or more were present, and no prior ablative, photodynamic, or surgical intervention for the disease had been performed. At each of the two primary study timepoints (T1, T2), biopsies and one-third and two-thirds of the height of the BE segment (measured from the GEJ) were collected for subsequent WGS. If EAC was present, it was preferentially chosen as one of the two T2 biopsies. Ten NCO patients, two additional biopsies were collected at a long-term follow up timepoint. For each biopsy, the epithelial layer was isolated from stroma. Matched normal were derived from blood, and additional normal gastric fundus tissue was sequenced in seven patients.

WGS was performed using an Illumina HiSeqX sequencer, following the WGS library preparation protocol outlined in Paulson et al. 2021. Reads were then aligned with BWA-MEM

(30) (version 0.6.2-r126) to GRCh37 (1000 Genomes Project human_g1k_v37 with decoy sequence hs37d5). BAM files went subsequent indel realignment with GATK IndelRealigner (31) (version 3.4-0-g7e26428).

Whole genome duplication calls were specified by examining WGS from samples to determine if >1000 Mbp of the genome showed evidence of somatic copy number amplification. Chromothripsis status was called using JaBbA (32), as presented in Paulson et al. 2021.

4.4.2 FHCRC cohort focal amplification detection

We utilized the PrepareAA wrapper (https://github.com/jluebeck/PrepareAA) to detect focal amplifications in the FHCRC cohort. The wrapper pipeline for seed detection incorporated CNVKit (33) (version 0.96) run in tumor-normal mode to call somatic CNVs against the matched normal WGS samples for each patient (when multiple normal samples were available, one was selected arbitrarily). Normal samples also underwent the same pipeline in unpaired mode for standalone CNV detection. The CNV calls were then provided the amplified_intervals.py script and filtered based on regions having CN > 4.3 (4.0 for normals) and size > 50kbp (10kbp for normals) to produce sets of seed regions. For each patient, seed regions from all that patient's samples were combined into a set of patient-specific regions, such that AA would be run on the same regions for every sample. The wrapper then invoked AmpliconArchitect (23) (version 1.2) in default mode on the WGS BAM files to examine seed regions and profile the architecture of the focal amplifications. The resulting graph and cycles output files were provided to AmpliconClassifier (AC) (version 0.4.5) to produce classifications of the AA amplicons for ecDNA, BFB, complex non-cyclic and linear focal amplifications (Methods - "Amplicon classification and ecDNA detection"). AC also specified bed files corresponding to the classified regions and annotated the identity of genes on the focal amplifications.

4.4.3 MRCCU cohort ethics approval

The MRCCU cohort consists of Barrett's Esophagus (BE) cases with 42 patients having low grade disease, 25 with high grade disease, and 50 early-stage (T1) esophageal adenocarcinoma (EAC). Patients with low grade BE and high grade BE underwent surveillance at Cambridge University Hospitals NHS Trust and consented prospectively to a biomarker and genomic characterization study (Cell Determinants Biomarker, REC no. 01/149, BEST2 REC no. 10/H0308/71). Early-stage EAC patients were recruited for the EAC International Cancer Genome Consortium (ICGC) study, for which samples were collected through the UK-wide Oesophageal Cancer Classification and Molecular Stratification (OCCAMS, Rec. no. 10-H0305-1) consortium. Ethical approval for these trials were from the East of England-Cambridge Central Research Ethics Committee.

4.4.4 MRCCU sample selection and sequencing data.

For all MRCCU samples, strict pathology consensus review was carried out, with 30% of pathological cellularity required for Barrett's samples and 70% percent for early-stage cancers. BE research samples were collected every 2cm of the BE segment at endoscopy and snap-frozen. A snap frozen section was taken from each BE sample to determine the grade of dysplasia. If more than one grade was present in a sample, it was classified according to the highest grade. In cases which progressed to multiple different disease stages, the highest grade of dysplasia in the case's follow-up was used for sequencing. Patients who received prior ablative treatment or had BE adjacent to cancer were excluded. Samples with squamous contamination were excluded.

Early-stage EAC samples were prospectively collected as endoscopic biopsies or resection specimens. All tissue samples were snap frozen and blood or normal squamous epithelium (at least 5cm from the tumor) were used as germline reference as previously described (29).

We utilized sequencing data generated in Katz-Summercorn et al. 2021. Reads were aligned with BWA-MEM to GRCh37 (1000 Genomes Project human_g1k_v37 with decoy sequences hs37d5).

4.4.5 MRCCU and ICGC focal amplification detection

Both MRCCU BAM files and ICGC BAM files were aligned to (1000 Genomes Project human_g1k_v37 with decoy sequences hs37d5). Absolute copy number (CN) profiles were generated using ASCAT (34) (v2.3). Genomic regions with a total CN > 4.5 and interval size > 10kbp were identified, merged, and refined with the amplified_intervals.py script. Each seed region was given to AA separately to improve runtime on each sample. AA was run in the default explore mode to reconstruct amplicon structures and amplicons formed by the same regions were deduplicated based on genomic overlap such that the highest-level classification amplicon was kept (ranked by ecDNA, BFB, complex non-cyclic, and then linear), ties being broken by largest amplicon size.

4.4.6 TCGA data processing

We utilized the Dockerized PrepareAA wrapper to detect focal amplifications in the TCGA cohort. The wrapper pipeline for seed detection incorporated CNVKit (version 0.9.7) run in unpaired mode to detect CNVs. The CNV calls were then provided the amplified_intervals.py script and filtered based on regions having CN > 4.5 and size > 50kbp to produce a set of seed regions. We used AmpliconArchitect (version 1.2) to infer the architecture of amplicons, The pipeline was run on 20 TCGA-ESCA EAC tumor whole genome sequencing BAMs, aligned to GRCh37, through the Institute for Systems Biology Cancer Genomics Cloud (https://isb-cgc.appspot.com/) which provides a cloud-based platform for TCGA data analysis.

4.4.7 Amplicon classification and ecDNA detection

We utilized AmpliconClassifier (AC) (https://github.com/jluebeck/AmpliconClassifier) (version 0.4.5) to perform classification of AA outputs into different types of focal amplifications and to extract coordinates of the genomic regions corresponding to those classifications. AC takes two inputs - the AA breakpoint graph file encoding genomic segment copy numbers and SV breakpoint junctions, as well as the AA cycles file encoding decompositions of the AA graph file into overlapping cyclic and/or non-cyclic paths weighted by the portion of the genomic CN they represent. AmpliconClassifier uses multiple heuristics to perform the classifications. First AC filters the paths < 10kbp, paths which significantly overlap low-complexity or repetitive regions, paths which overlap regions of the genome never exceeding CN 4.5 (not focally amplified), or which have a decomposed CN < δ (too low-frequency relative to other decompositions for reliable classification as focal amplifications). The decomposed CN (c_p) threshold, δ, for a path p, having a maximum genomic CN of m_p is defined as

$$\delta = \begin{cases} |p| = 1: & \min(1, \frac{m_p}{10}) \\ m_p > 7: & \min(3, \frac{m_p}{8}) \\ \quad else: & 2.5 \end{cases}$$

For each remaining path, AC computes a length-weighted CN, called *W*, which is the product of the length of the path (in kbp) and the decomposed path’s assigned copy number.

AC first assess non-filtered paths for the presence of BFB cycles using heuristics determined from manual examination of BFB-like focal amplifications in the FHCRC cohort and focal amplifications in previous studies (5, 23). AC computes the fraction of breakpoint graph discordant edges which are foldback, *f*, – i.e., inverted orientation having a genomic distance < 25kbp. AC then identifies decomposed paths containing foldback junctions between segments, and using all paths computes the set of consecutive segment pairs in the paths where the two boundaries of the segments together form a foldback junction. Each segment pair is assigned its own weight equal to the decomposed copy count of the path. If the proportion of BFB-like segment pairs over all segment pairs in all paths is less than 0.295, then the amplicon is not considered to contain a BFB. Furthermore, if the total weights of pairs which are "distal" (not foldback and > 5kbp jump between endpoints) divided by the total weight of all pairs is greater than 0.5, the amplicon is not considered to contain BFB. Lastly, if the total decomposed CN of all pairs is < 1.5, or if the total number of foldback segment pairs is < 3, or *f* < 0.25, or the decomposed CN weight of all BFB-like paths divided by the CN weight of all paths < 0.6, or the maximum genomic copy number of any region in the candidate BFB region is < 4, the amplicon is not considered to contain a BFB. If the amplicon has not failed any of these criteria, a BFB-positive status is assigned, and the BFB-like cycles (decomposed paths with a BFB foldback) are put into a set and kept separate from additional fsCNA detection inside the amplicon region.

Next, AC assess non-filtered, non-BFB paths for the presence of ecDNA cycles. If there is any cyclic path with decomposed CN > 5 and length > 100kbp, an ecDNA-positive status is assigned. If the total fraction of length-weighted CN, *W*, assigned to cycles exceeds 12% of the total length-weighted CN in the cycles file and more than 10kbp are inside the filtered cyclic paths, an ecDNA-positive status is assigned. Lastly, if the total length of complex cycles (cyclic paths with interior rearrangements > 5kbp) exceeds 50kbp and the region has CN > 4.5 an ecDNA-positive status is assigned. The ecDNA-like cyclic paths are then stored for subsequent analysis, including reporting of the genomic coordinates as a bed file and annotation of genes.

If the amplicon is not classified as BFB-positive and/or ecDNA-positive, and has paths consistent with focal amplification, then two other classifications are checked. If the fraction of *W* assigned to non-cyclic paths with rearrangements > 5kbp plus *W* assigned to cyclic paths is greater than 0.3 of total *W* in all paths, a complex non-cyclic label is assigned. If the ratio of *W* assigned to non-cyclic paths without rearrangements to *W* assigned to non-amplified paths is greater than 0.25, then the path is labeled complex non-cyclic if the breakpoint graph has > 4 discordant edges in amplified regions, otherwise a linear amplification label is assigned. If not resolved by these heuristics, the path type with the highest fraction of *W* is assigned.

4.4.8 Statistical testing and odds ratios

We used SciPy(35) (version 0.19.1) to conduct all statistical tests in the study, with the exception of the ecDNA/oncogene overlap significance test which utilized ISTAT (36) (version 1.0.0). When computing odds ratios, if any cell in the two-by-two table was zero, the Haldane correction was applied to every cell in the table.

4.4.9 Oncogene selection

We assembled a curated set of oncogenes by combining oncogene lists from multiple sources, to include standard and BE-specific oncogenes. The oncogene list from the ONGene database (37) was combined with two additional disease-specific sets. First, the EAC driver oncogene list reported in Frankell et al. (29), and second the list of EAC driver oncogenes reported in Paulson et al. 2021. We defined the EAC-specific oncogenes as the combined set of oncogenes reported by Frankell et al. and Paulson et al. 2021.

4.4.10 *TP53* disruption analysis

For the FHCRC cohort, *TP53* disruption status was determined in Paulson et al. 2021. In brief, mutations were defined as any moderate- to high-impact SNV or indel as reported by SNPeff (38). Deletions of at least one exon, or SVs affecting the *TP53* coding sequence or splice sites were also considered to disrupt *TP53*, as were copy number alterations affecting at least half of exonic regions. All alterations were verified manually using IGV (39) or Partek®. For the MRCCU cohort, *TP53* status was determined by identifying somatic coding variants (missense, frameshift, stop gain or splice site variants), using Strelka (40) v2.0.15 and Variant Effect Predictor (41) (VEP) version 78.

4.4.11 Amplicon similarity score

We compared overlapping focal amplifications to quantify amplicon similarity by quantifying the relative amounts of shared overlap in genomic coordinates and in SV breakpoint location. These calculations are implemented into the amplicon_similarity.py script, available in the AmpliconClassifier repository (https://github.com/jluebeck/AmpliconClassifier).

We defined symmetric and asymmetric amplicon similarity scores combining information from both the genomic interval overlap and the shared breakpoint junctions. An amplicon is defined as a collection of breakpoints (*B*), and genomic segments (*G*). Genomic overlap was evaluated on the basis of the number overlapping base-level coordinates in two intervals. Breakpoints were considered to be shared if the total distance between the two endpoints of each junction was in total measured to be less than *d* (default = 250bp). That is, for two breakpoints *x* and *y* with sorted endpoints (x_1,x_2) and (y_1,y_2), respectively, they must satisfy

$$|x_1 - y_1| + |x_2 - y_2| < d$$

The asymmetric amplicon similarity score between two amplicons A_1 and A_2 we defined as

$$Asym(A_1, A_2) = \frac{\alpha(G_1 \cap G_2)}{G_1} + \frac{(1-\alpha)(B_1 \cap B_2)}{B_1}$$

and similarly, the similarity of A_2 to A_1 is

$$Asym(A_2, A_1) = \frac{\alpha(G_2 \cap G_1)}{G_2} + \frac{(1-\alpha)(B_2 \cap B_1)}{B_2}$$

Where α is set to 0.25 by default. We then define a symmetric amplicon similarity score which is the average of the two asymmetric scores

$$Sym(A_1, A_2) = \frac{Asym(A_1, A_2) + Asym(A_2, A_1)}{2}$$

We computed symmetric amplicon similarity scores for a panel of amplicons from unrelated origins derived from Deshpande et al. (23), deCarvahlo et al. (42) and the amplicons from unrelated patients in the FHCRC cohort. We used the resulting distribution of 719 similarity scores for overlapping amplicons as a background null distribution. We computed the percentile of each new amplicon similarity score in this null distribution to quantify its similarity against the panel of overlapping amplicons from unrelated origins.

We also fit a beta distribution to the empirical null symmetric similarity score distribution, using a maximum likelihood estimation approach to fit the parameters of the model. The beta distribution was selected as it provides support on the interval [0, 1], provides a higher degree of flexibility in fitting various distributions given the two shape parameters, and enables a better estimation of small p-values than the empirical dataset. We performed negative log likelihood minimization using the SciPy (version 0.19.1) *fmin* function with initial parameter estimates (1.5, 10), and convergence occurred in 39 iterations.

As AmpliconArchitect may include flanking regions which are not focally amplified as part of the amplification itself, the amplicon similarity script filters from the calculation regions that are not focally amplified (CN < 4.5 default), SVs which join two elements less than 2500bp away, and it redundantly filters regions that are also present in the low-complexity or low-mappability database used by AmpliconArchitect.

4.4.12 Amplicon complexity score

AmpliconArchitect outputs a collection of (cyclic and/or non-cyclic) paths in the CN-aware breakpoint graph representing an approximate optimal balanced CN flow in the graph. As a result, non-trivial graphs may be decomposed into multiple paths, each having some copy-number assigned to the path, constrained by the total amount of CN flow available in the graph.

Each path has a copy number *c*, and a length in kilobase pairs, *s*. The total length-weighted copy number of all decomposed paths we call *T*, and is given by

$$T = \sum_{i=1}^{n} s_i c_i$$

Where c_i (s_i) is the copy number (length) of the *i*-th path. The values of c_i are pre-sorted in descending order for increasing *i*. For the decomposed paths of each amplicon graph, *G*, we computed a vector representing the fraction of total CN captured by each of the *n* decompositions. We denote this sorted collection as,

$$D = \left(\frac{s_1 c_1}{T}, \ldots, \frac{s_n c_n}{T}\right)$$

We noted that there may be many low-weight CN paths, representing non- or weakly-amplified paths extracted from the graph, and thus we defined a "residual", measured against the first percentile, *p*, (default = 80%) of weighted CN explained. We first define an index *j*, where *j* is the largest value such that

$$0 \leq j < n$$

$$\sum_{j=1}^{n} D_j < p$$

This implies that *j*+1 represents the first index such that sum of the first *j*+1 entries exceed *p*. The residual, ϵ, we defined as the weighted CN fractions above the first j+2 entries, is then given by

$$\epsilon = \sum_{i=j+2}^{n} D_i$$

We then defined an amplicon complexity score function *H(ϵ, D, k),* represented by the sum of entropies from the residual, the non-residual, and the total number of segments in the breakpoint graph, *k*.

$$H(\epsilon, D, k) = -\epsilon \ln \epsilon - \sum_{i=1}^{j+1} D_i \ln D_i - \ln \frac{1}{k}$$

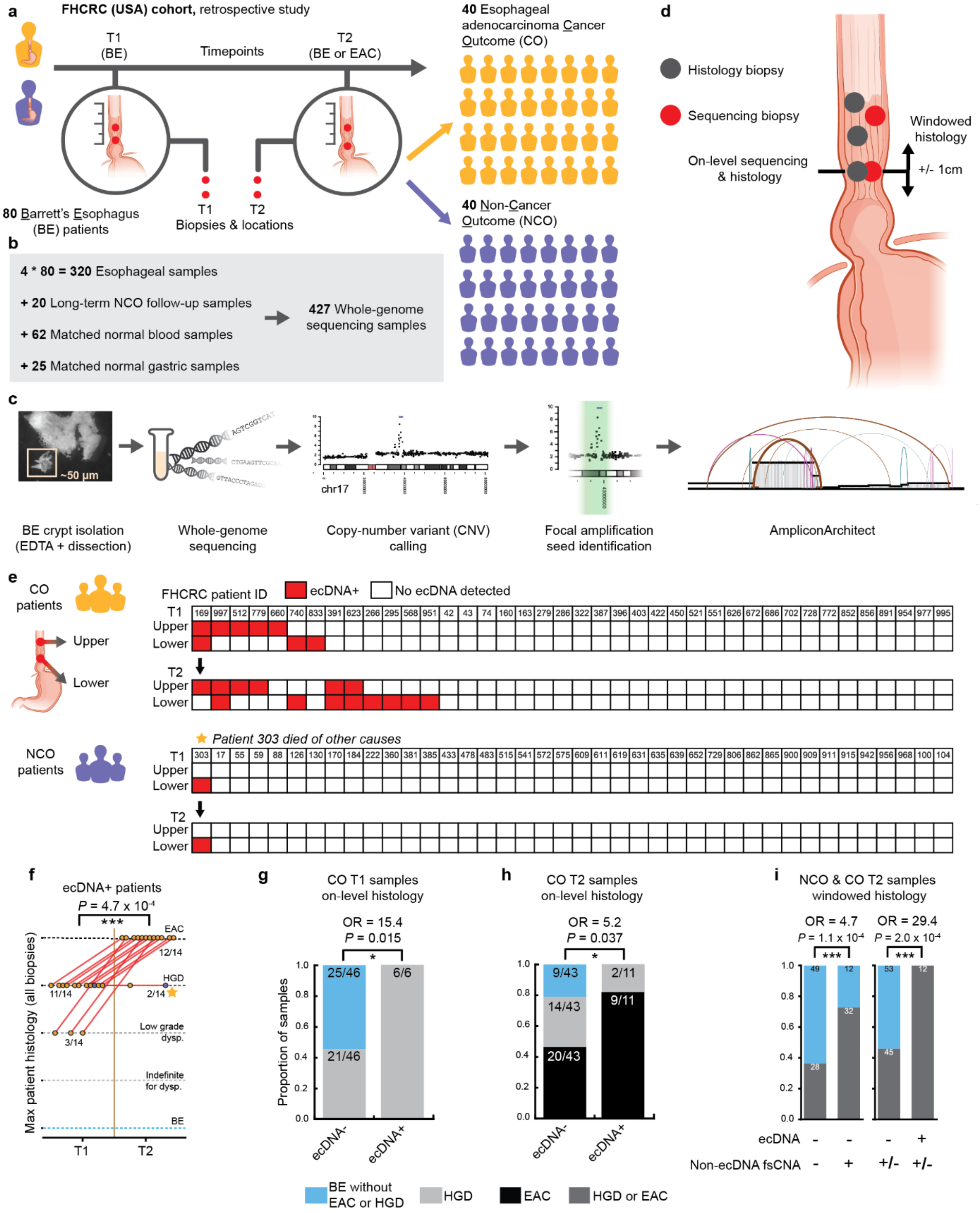
4.5 Acknowledgements

Chapter 4, in full is currently being prepared for submission for publication of the material and was co-authored by Jens Luebeck, Sihan Wu, Xiaohong Li, Alvin Ng, Patricia Galipeau, Carissa Sanchez, Annalise Katz-Summercorn, Hoon Kim, Sriganesh Jammula, Phoebe He, Howard Y. Chang, Roel Verhaak, Carlo C. Maley, Ludmil Alexandrov, Brian J. Reid, Rebecca Fitzgerald, Thomas Paulson, Vineet Bafna & Paul S. Mischel. The dissertation author was the primaryinvestigator and author of this material. The esophagus illustration

shown in our study was created by scientific illustrator Tami Tolpa (Tolpa Studios, www.tolpa.com). We also thank Shahab Sarmashghi for discussions about ISTAT.

4.6 Appendix

Figure 4.1: FHCRC BE study. **a)** Design of the FHCRC retrospective Barrett's Esophagus (BE) surveillance study, sample collection and separation of patients into cancer outcome (CO) and non-cancer outcome (NCO) status. **b)** Sequencing biopsies and histology biopsies were taken independently. Some histology and sequencing biopsies were taken on the same level of the esophagus (on-level), and some histology biopsies fell within +/- 1cm of the measured height of the sequencing biopsy. **c)** The number of sequencing samples generated in FHCRC study labeled by sample type. **d)** Experimental workflow for analyzing Barrett's WGS samples. **e**) AmpliconArchitect identified ecDNA in WGS samples from both sequencing samples, in both time-points and in both CO and NCO patients. **f)** The maximum histology discovered in any histology biopsy for ecDNA+ patients at T1 and T2. **g)** The proportion of T1 samples without EAC or HGD versus with HGD in CO patients (before developing cancer), segregated by ecDNA status of on-level sequencing biopsies. **h)** The proportion of T2 samples without HGD or EAC (BE) versus with HGD or EAC in CO patients (cancer first detected), segregated by ecDNA status of sequencing biopsies with on-level histology. **i)** The proportion of T2 samples without EAC or HGD (BE) versus with EAC or HGD in all FHCRC patients segregated by ecDNA status and non-ecDNA fsCNA status as measured from maximum windowed histology.



CO patients

FHCRC patient ID	169	997	512	779	660	740	833	391	623	266	295	568	951	42	43	74	160	163	279	286	322	387	396	403	422	450	521	551	626	672	686	702	728	772	852	856	891	954	977	995
T1 Upper	🟥	🟥	🟥	🟥	🟥	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
T1 Lower	🟥	□	□	□	□	🟥	🟥	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
T2 Upper	🟥	🟥	🟥	🟥	□	□	□	🟥	🟥	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
T2 Lower	□	🟥	□	□	□	🟥	□	🟥	🟥	🟥	🟥	🟥	🟥	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

NCO patients

FHCRC patient ID	303	17	55	59	88	126	130	170	184	222	360	381	385	433	478	483	515	541	572	575	609	611	619	631	635	639	652	729	806	862	865	900	909	911	915	942	956	968	100	104
T1 Upper	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
T1 Lower	🟥	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
T2 Upper	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
T2 Lower	🟥	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

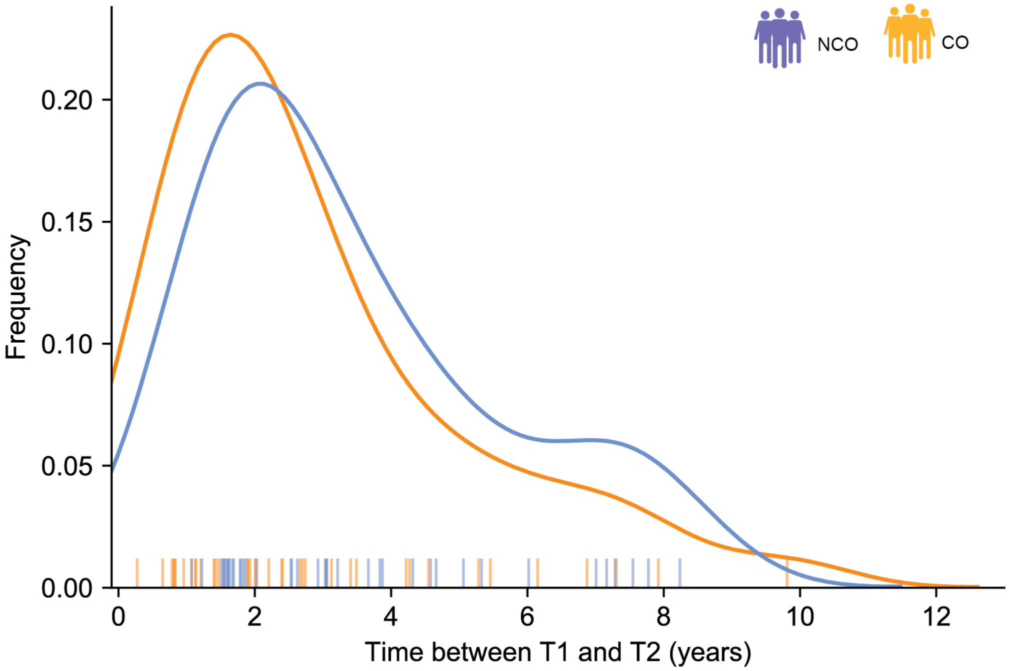
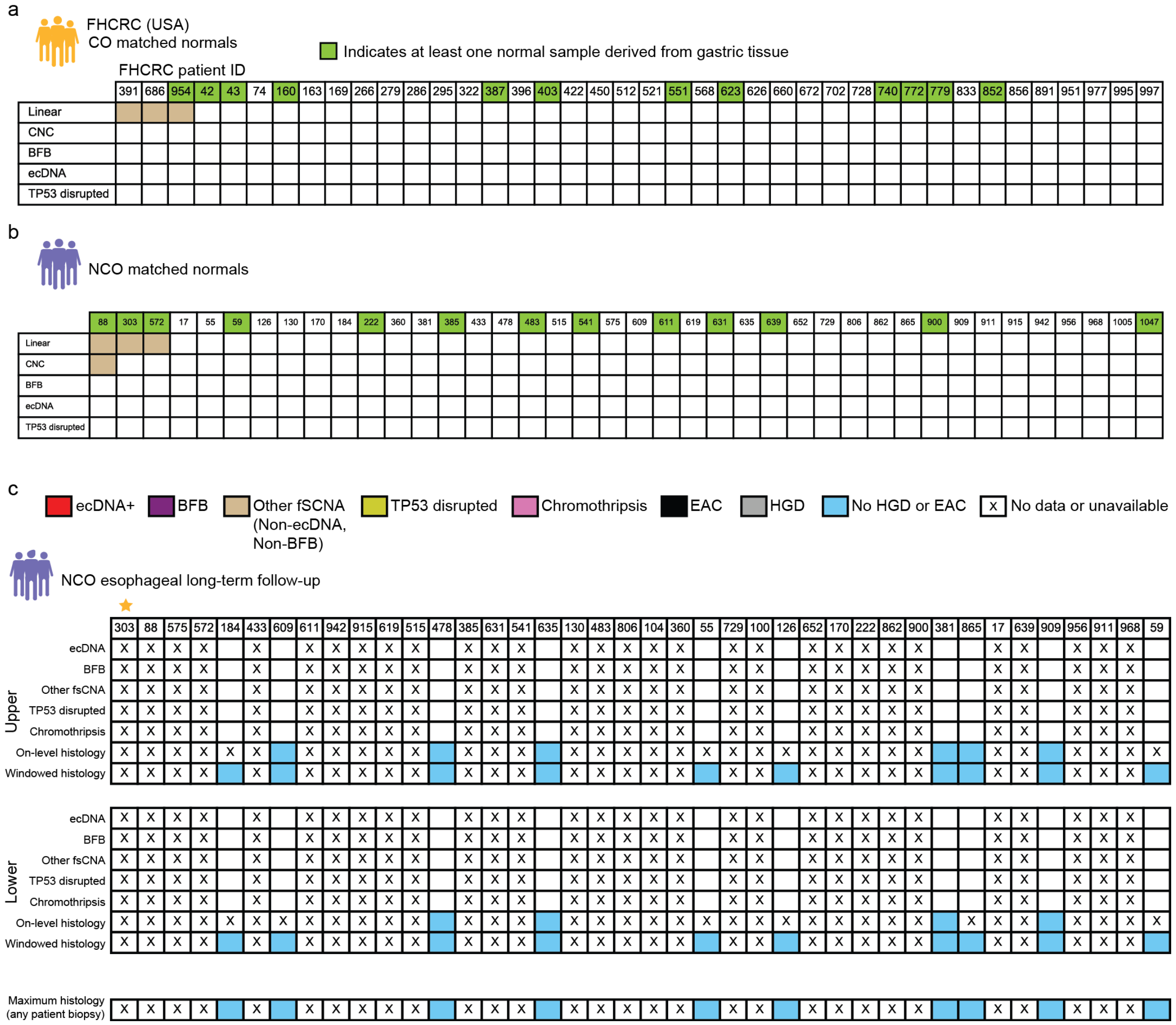


Figure 4.2: The distribution of time differences between T2 and T1 in the FHCRC study, separated by CO status.



a

FHCRC (USA)
CO matched normals

🟩 Indicates at least one normal sample derived from gastric tissue

FHCRC patient ID

	391	686	🟩 954	🟩 42	🟩 43	74	🟩 160	163	169	266	279	286	295	322	🟩 387	396	🟩 403	422	450	512	521	🟩 551	568	🟩 623	626	660	672	702	728	🟩 740	🟩 772	🟩 779	833	🟩 852	856	891	951	977	995	997
Linear	🟫	🟫	🟫																																					
CNC																																								
BFB																																								
ecDNA																																								
TP53 disrupted																																								

b

NCO matched normals

	🟩 88	🟩 303	🟩 572	17	55	🟩 59	126	130	170	184	🟩 222	360	381	🟩 385	433	478	🟩 483	515	🟩 541	575	609	🟩 611	619	🟩 631	635	🟩 639	652	729	806	862	865	🟩 900	909	911	915	942	956	968	1005	🟩 1047
Linear	🟫	🟫	🟫																																					
CNC	🟫																																							
BFB																																								
ecDNA																																								
TP53 disrupted																																								

c

🟥 ecDNA+ 🟪 BFB 🟫 Other fSCNA (Non-ecDNA, Non-BFB) 🟨 TP53 disrupted 🟪 Chromothripsis ⬛ EAC ⬜ HGD 🟦 No HGD or EAC ☒ No data or unavailable

NCO esophageal long-term follow-up

		★ 303	88	575	572	184	433	609	611	942	915	619	515	478	385	631	541	635	130	483	806	104	360	55	729	100	126	652	170	222	862	900	381	865	17	639	909	956	911	968	59
Upper	ecDNA	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	BFB	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	Other fsCNA	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	TP53 disrupted	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	Chromothripsis	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	On-level histology	X	X	X	X	X	X	🟦	X	X	X	X	X	🟦	X	X	X	🟦	X	X	X	X	X	X	X	X	X	X	X	X	X	X	🟦	🟦	X	X	🟦	X	X	X	X
	Windowed histology	X	X	X	X	🟦	X	🟦	X	X	X	X	X	🟦	X	X	X	🟦	X	X	X	X	X	🟦	X	X	🟦	X	X	X	X	X	🟦	🟦	X	X	🟦	X	X	X	🟦
Lower	ecDNA	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	BFB	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	Other fsCNA	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	TP53 disrupted	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	Chromothripsis	X	X	X	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X		X	X	X	X	X			X	X		X	X	X	
	On-level histology	X	X	X	X	X	X	X	X	X	X	X	X	🟦	X	X	X	🟦	X	X	X	X	X	X	X	X	X	X	X	X	X	X	🟦	X	X	X	🟦	X	X	X	X
	Windowed histology	X	X	X	X	🟦	X	🟦	X	X	X	X	X	🟦	X	X	X	🟦	X	X	X	X	X	🟦	X	X	🟦	X	X	X	X	X	🟦	🟦	X	X	🟦	X	X	X	🟦
	Maximum histology (any patient biopsy)	X	X	X	X	🟦	X	🟦	X	X	X	X	X	🟦	X	X	X	🟦	X	X	X	X	X	🟦	X	X	🟦	X	X	X	X	X	🟦	🟦	X	X	🟦	X	X	X	🟦

Figure 4.3: Oncoprint tables for **a)** FHCRC CO and **b)** NCO patient normal blood and/or normal gastric samples showing the classification of focal amplifications and TP53 disruption detected in any normal sample from the patient. **c)** Oncoprint table for NCO long-term follow-up samples esophageal. Table encodes ecDNA status, BFB status, other fsCNA (non-BFB, non-ecDNA) status, TP53 disruption, chromothripsis status, as well as on-level and windowed histology for each time point and both upper and lower WGS samples. Maximum histology of any sample from that time-point is shown at the bottom.

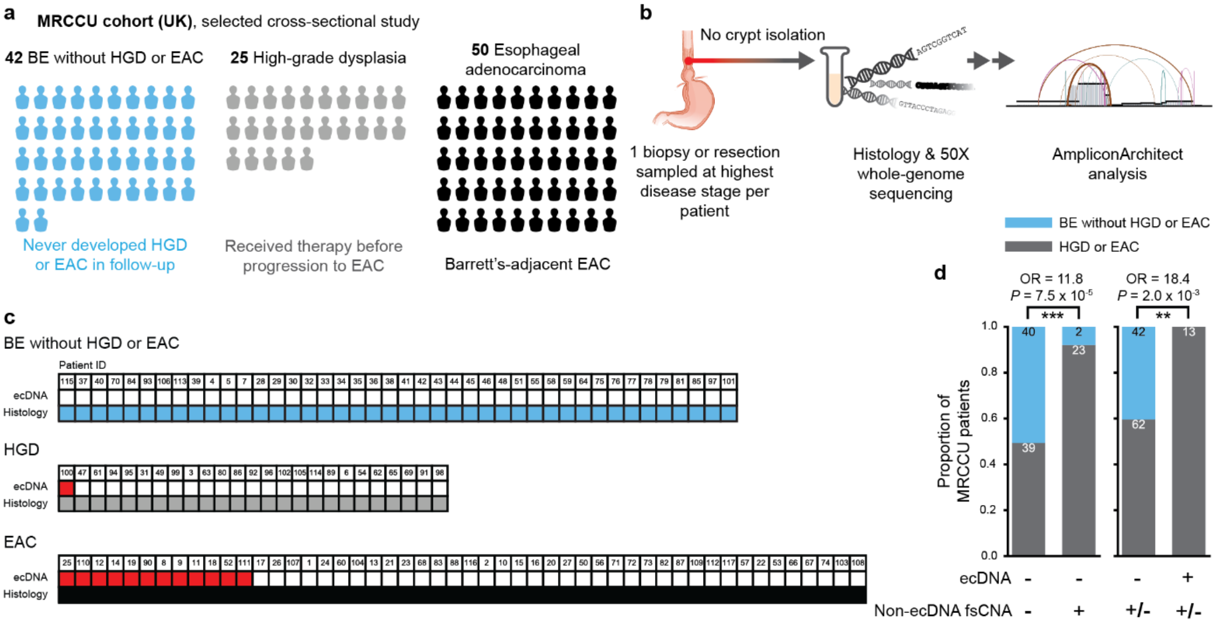
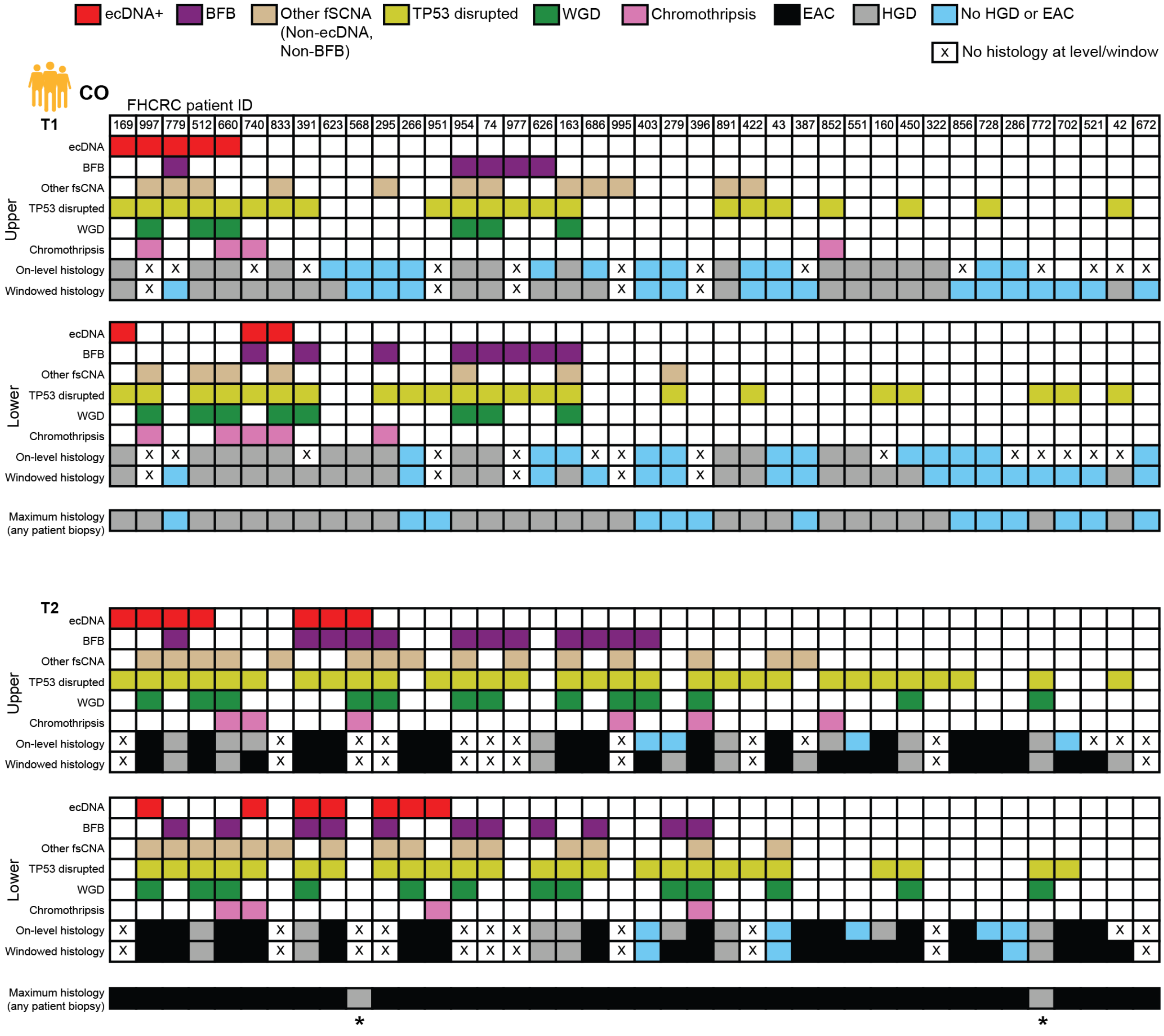


Figure 4.4: MRCCU BE study. **a)** Breakdown of the types of BE patients in the MRCCU selected cross-sectional study by histology of the sample sequenced. **b)** Summarization of the process by which biopsies were selected, sequenced, and characterized by AmpliconArchitect. **c)** MRCCU patients segregated by highest-disease state and ecDNA status from that sample. **d)** Proportion of MRCCU having BE without HGD or EAC versus having HGD or EAC segregated by ecDNA status and non-ecDNA fsCNA status.



* Indicates cancer diagnosis made in follow-up

Figure 4.5: Oncoprint table of FHCRC CO patient WGS samples encoding ecDNA status, BFB status, other fsCNA (non-BFB, non-ecDNA) status, TP53 disruption, whole-genome duplication (WGD) status, chromothripsis status, as well as on-level and windowed histology for each time point and both upper and lower WGS samples. Maximum histology from any histology biopsy is shown at the bottom of each time-point. Asterisk indicates cancer diagnosis was made in a follow-up clinic encounter. However, cancer was believed to be present at T2.

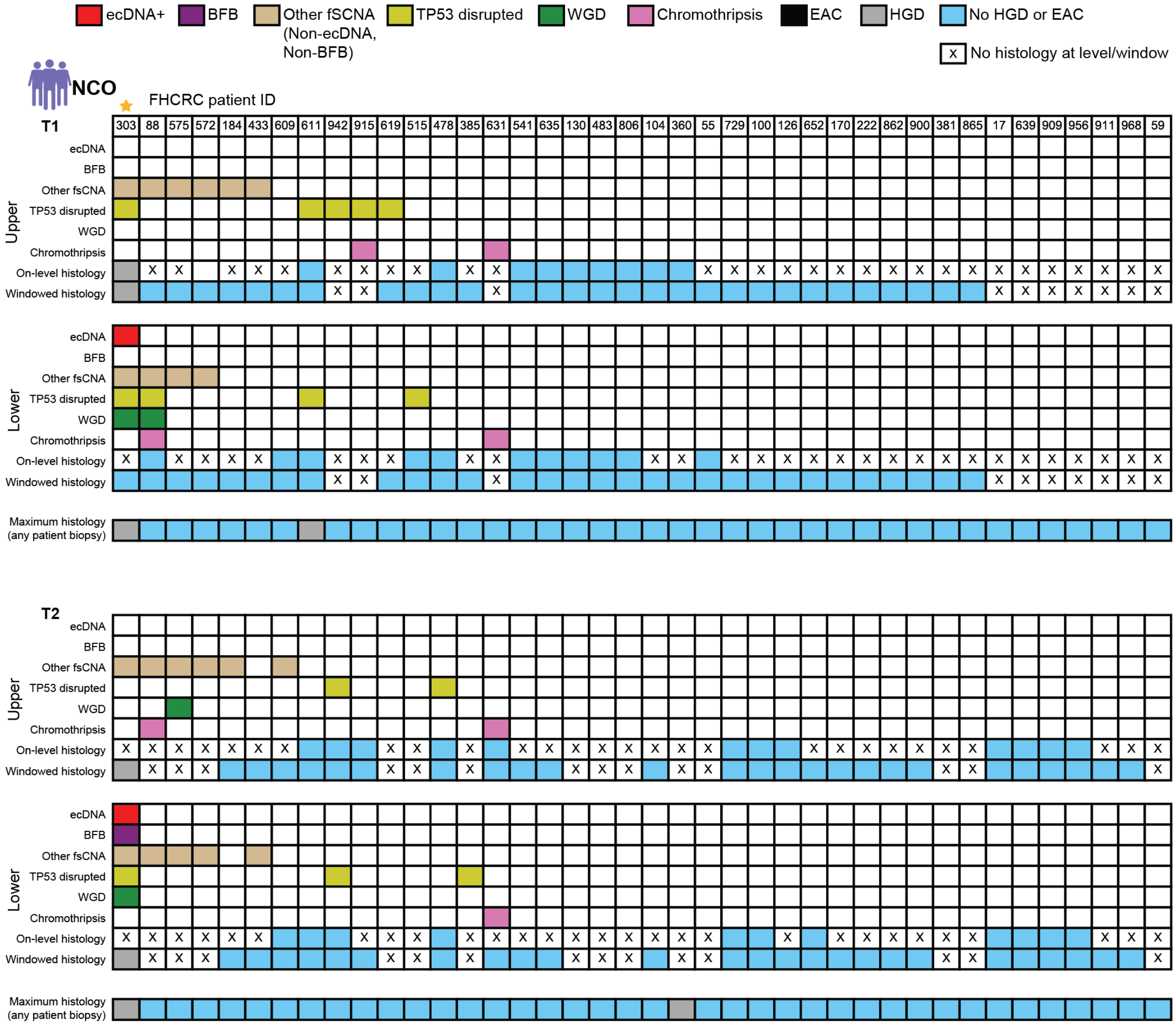


Figure 4.6: Oncoprint table of FHCRC NCO patient WGS samples encoding ecDNA status, BFB status, other fsCNA (non-BFB, non-ecDNA) status, TP53 disruption, whole-genome duplication (WGD) status, chromothripsis status, as well as on-level and windowed histology for each time point and both upper and lower WGS samples. Maximum histology from any histology biopsy is shown at the bottom of each time-point. Gold star over patient indicates patient 303, who died of non-EAC causes 2.5 years after the last endoscopy.

MRCCU cohort (UK)

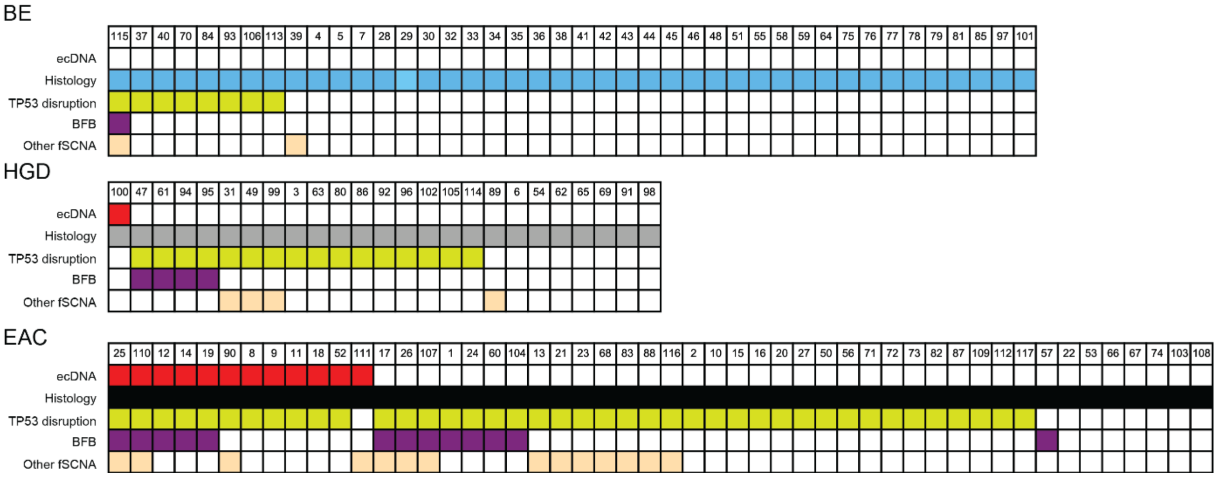


Figure 4.7: Oncoprint table for MRCCU BE and EAC patients segregated by histology type showing ecDNA status, TP53 disruption, BFB status, and other fsCNA status.

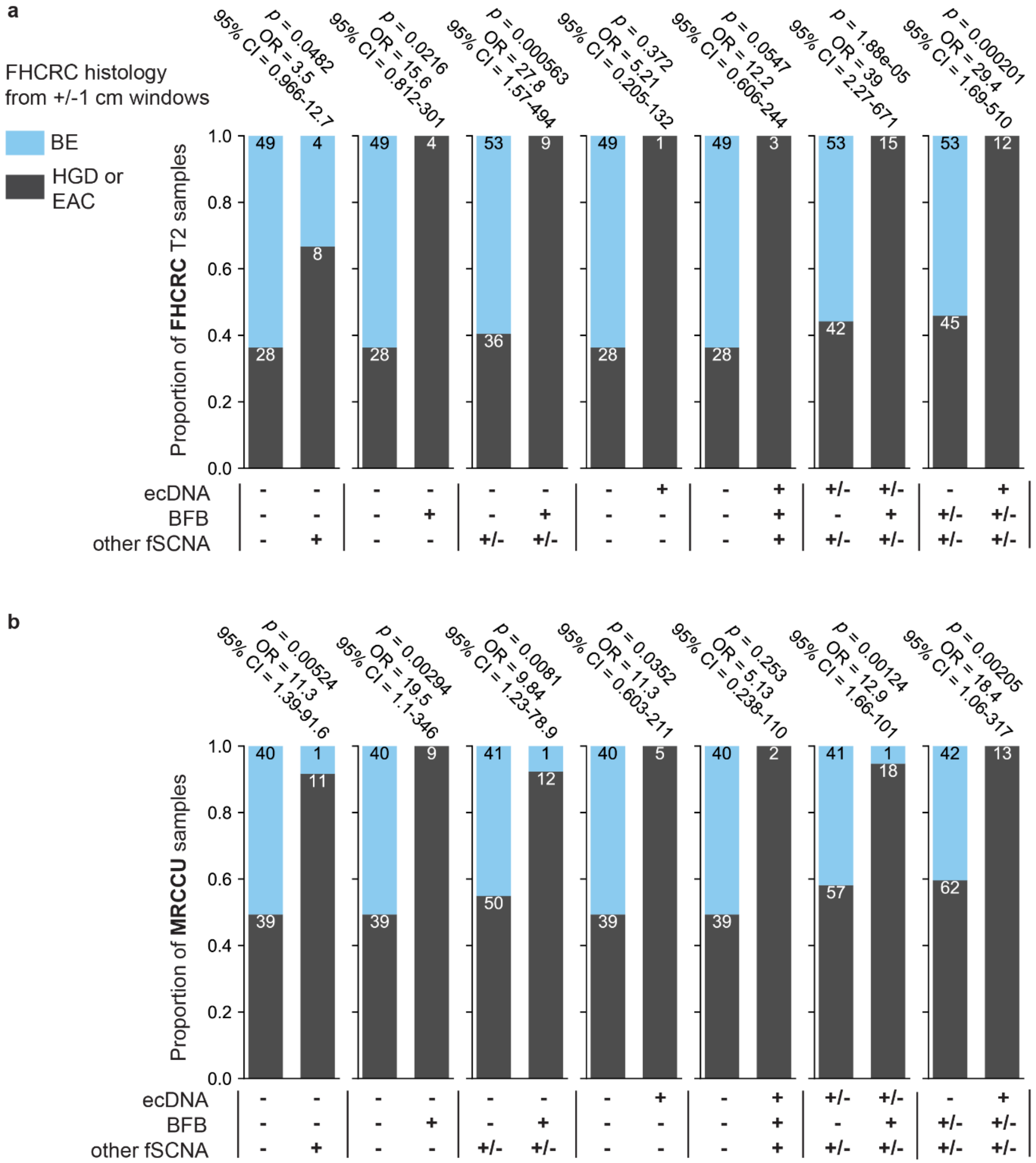


Figure 4.8: a) Odds ratios for FHCRC T2 samples (NCO and CO). The odds ratio here quantifies the strength of association for events (histology – BE versus HGD or EAC) against groups (focal amplification statuses). Shown above the breakdown of histology and focal amplification status are a p-value computed by Fisher exact test, the odds ratio, and a 95% confidence interval for the odds ratio. **b)** Odds ratios for MRCCU samples relating histology (BE versus HGD or EAC) and focal amplification status.

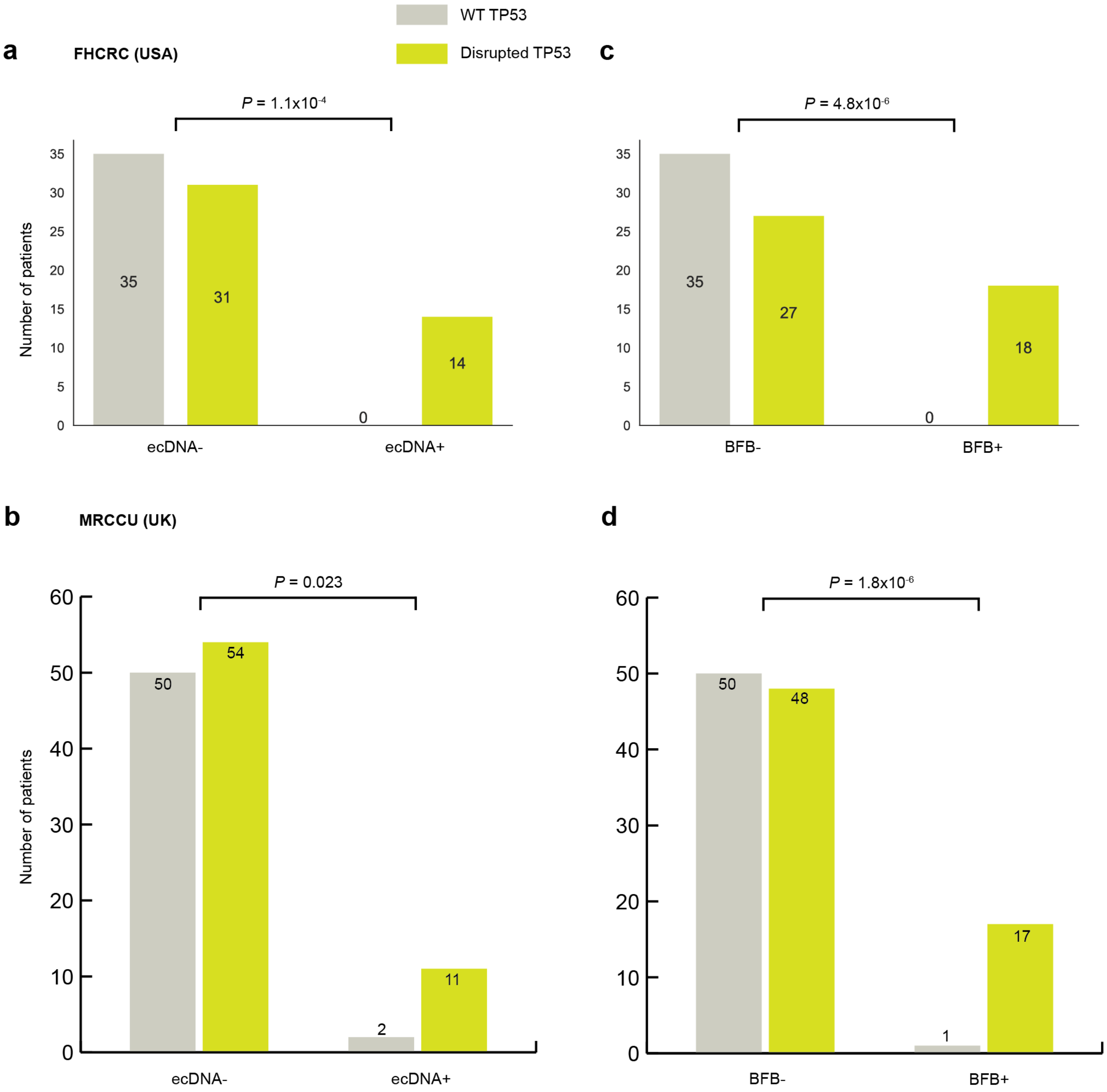


Figure 4.9: TP53 and ecDNA/BFB. **a)** TP53 status of each patient in the FHCRC study (T1 and T2 esophageal samples) segregated by ecDNA status demonstrates that all ecDNA-positive patients in the study have evidence of TP53 disruption. **b)** TP53 status of each patient in the MRCCU study segregated by ecDNA status demonstrates that the majority ecDNA-positive patients in the study have evidence of TP53 disruption. **c)** TP53 status of each patient in the FHCRC study (T1 and T2 esophageal samples) segregated by BFB status demonstrates that all BFB-positive patients in the study have evidence of TP53 disruption. **d)** TP53 status of each patient in the MRCCU study segregated by BFB status demonstrates that the majority of BFB-positive patients in the study have evidence of TP53 disruption.

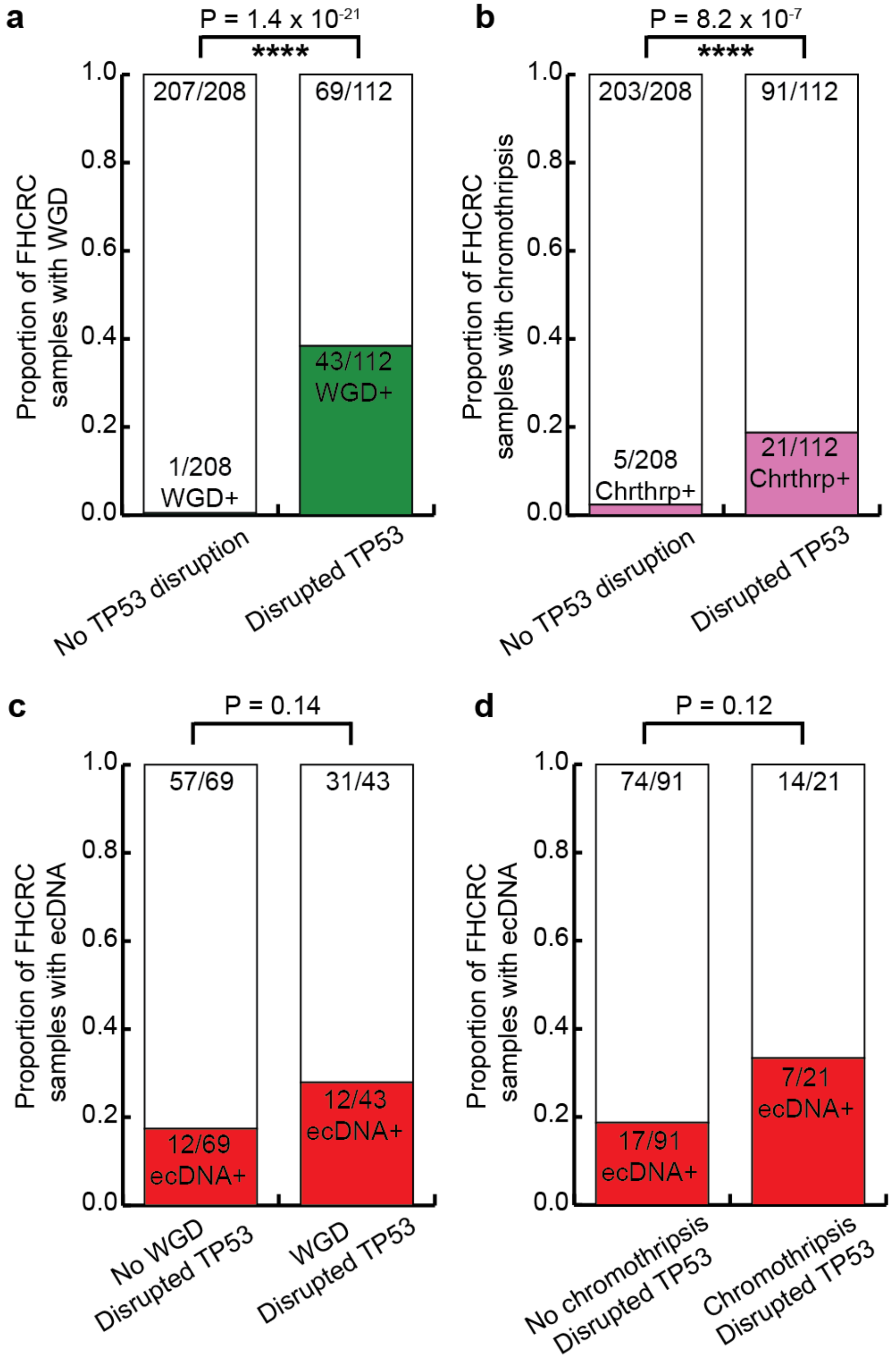


Figure 4.10. TP53 and other genomic features. **a)** Proportion of FHCRC samples having whole-genome duplication (WGD) segregated by TP53 disruption status. **b)** Proportion of FHCRC samples having chromothripsis segregated by TP53 disruption status. **c)** Proportion of FHCRC samples having ecDNA segregated by WGD status where TP53 is also disrupted. **d)** Proportion of FHCRC samples having ecDNA segregated by chromothripsis status where TP53 is also disrupted.

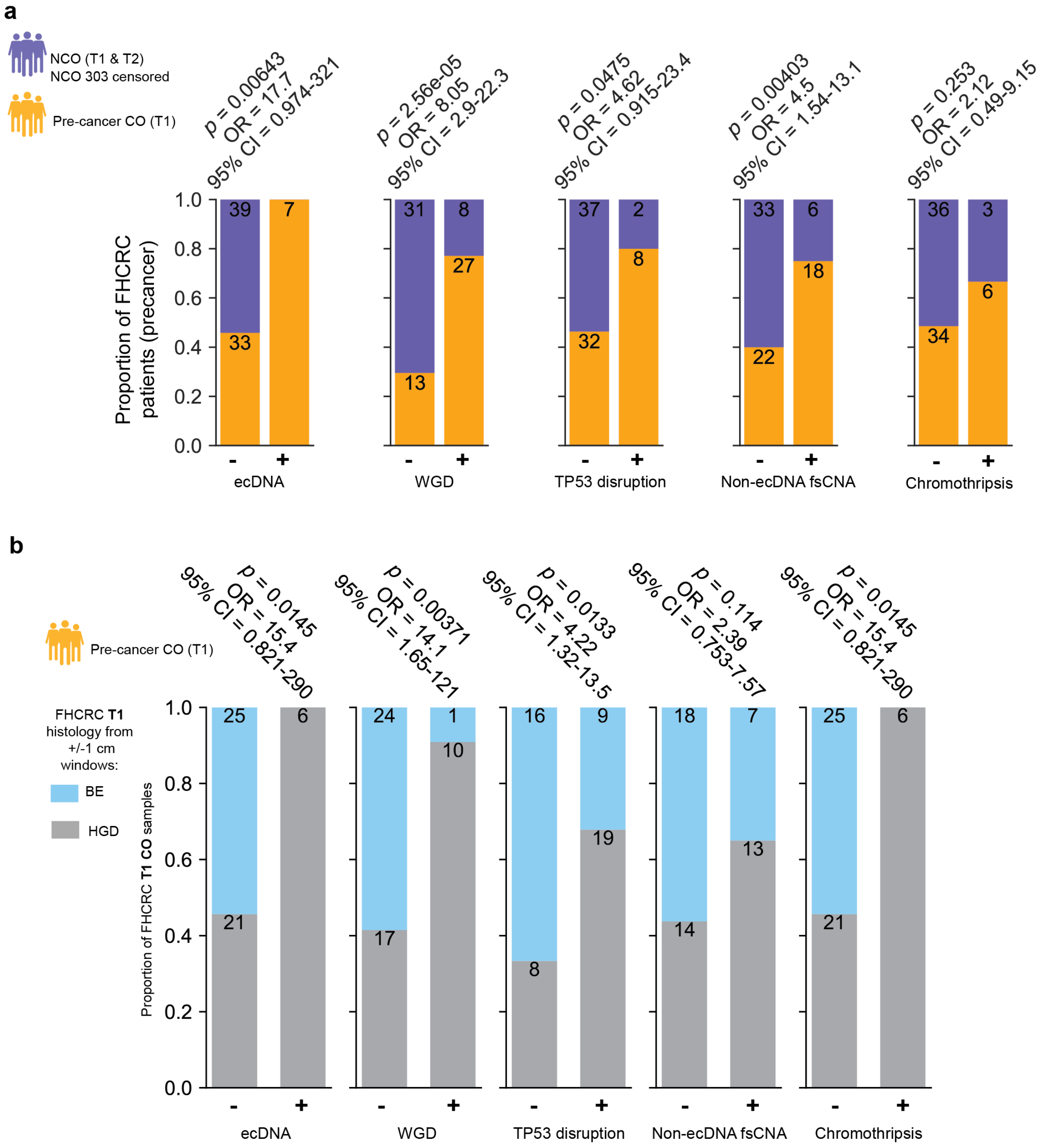


Figure 4.11: FHCRC OR plots. **a)** The relative proportions of FHCRC patients (NCO and CO), all at pre-cancer state, segregated by the status of various genomic lesions (ecDNA, whole-genome duplication TP53 disruption, non-ecDNA fsCNA, chromothripsis). Above the bar chart is annotated a p-value compute by a Fisher exact test, the odds ratio of the association between NCO/CO status and genomic lesion, as well as a 95% confidence interval for the association. Determination of the existence of the genomic lesion was made on the bases of all samples in T1 and T2 for NCO patients, and all samples in T1 for CO patients (pre-cancer). Patient 303 was censored from the NCO group due to poor survival after the last endoscopy (2.5 years). **b)** The relative proportions of FHCRC CO T1 samples in BE versus HGD histology category, segregated by the status of various genomic lesions.

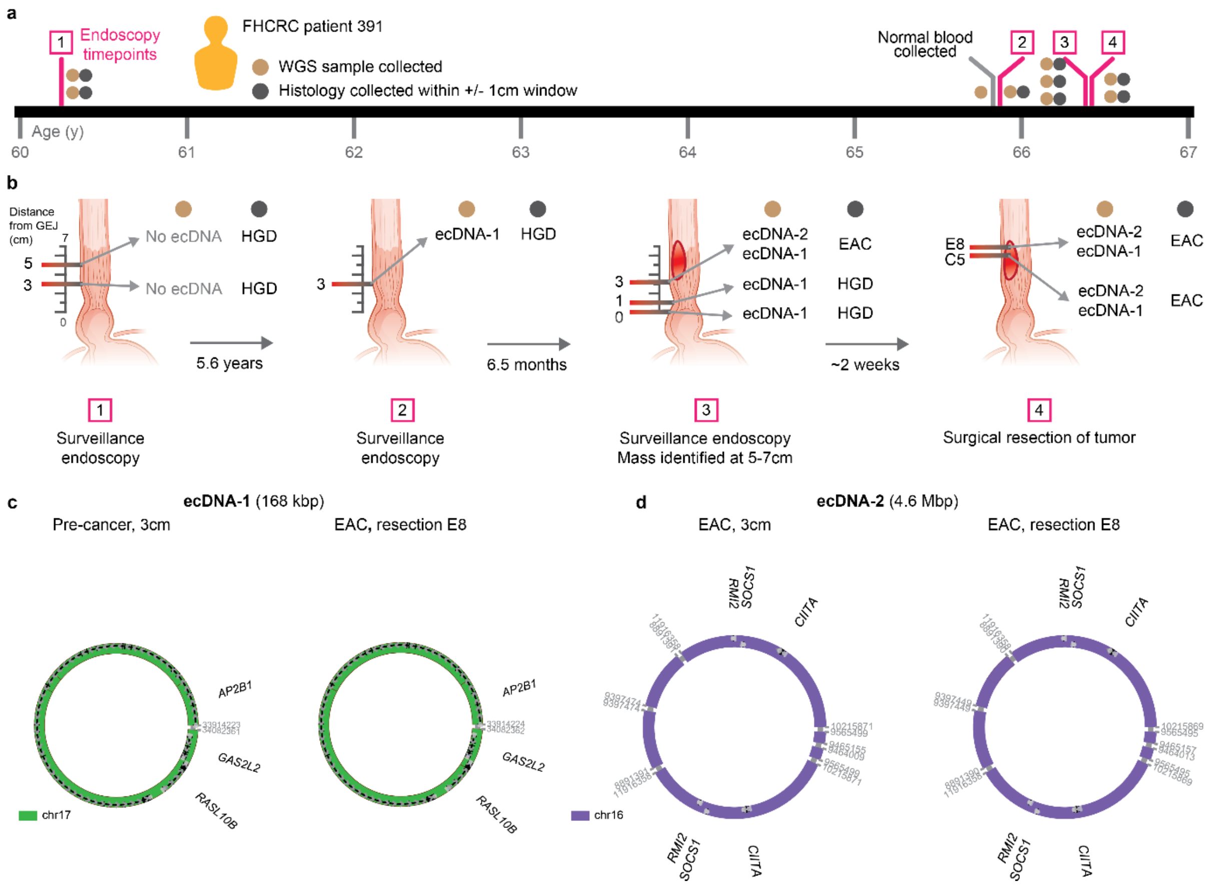
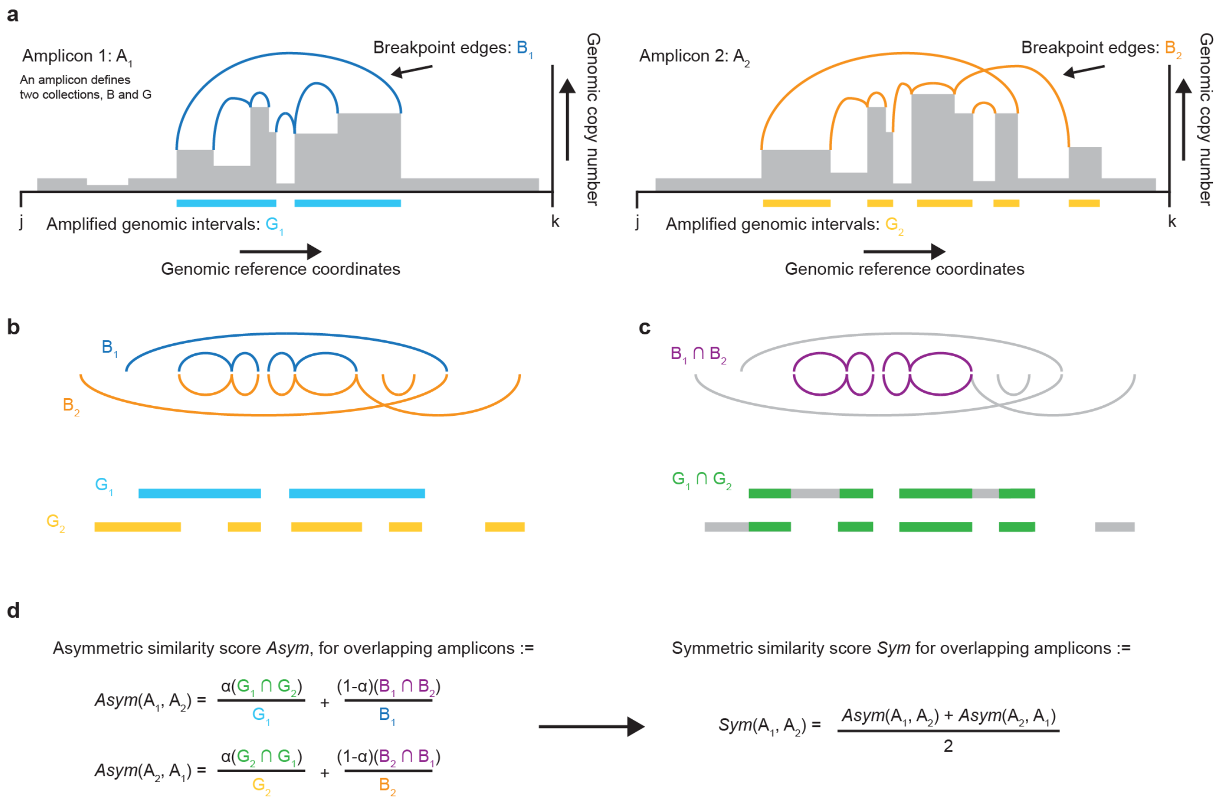


Figure 4.12: Amplicon similarity and ecDNA tracking. **a)** Cartoon representation of two overlapping focal amplifications. **b)** Cartoon representation of the intersection of SV breakpoint junctions and the intersection of genomic segments. **c)** Definitions of the asymmetric and symmetric amplicon similarity scores. **d)** The distribution of maximum asymmetric similarity scores for overlapping amplicons derived from different patients (left) and for overlapping amplicons derived from the same patient (right) in FHCRC NCO and CO patients. **e)** Probability density plot of amplicon similarity scores from a collection of unrelated samples with overlapping focal amplifications (blue), a beta distribution maximum-likelihood estimate of the empirical amplicon similarity score distribution (black), and the similarity scores of overlapping ecDNA amplicons from the same FHCRC patients (red). **f)** Timeline of sample collection in FHCRC CO patient 391 relative to patient age. **g)** Summary of the ecDNA status and windowed histology status for four endoscopies with time interval between each also indicated. Biopsy distances from the gastroesophageal junctions (GEJ) are indicated. Two distinct species of ecDNA are labeled as ecDNA-1 and ecDNA-2. **h)** The structure of ecDNA-1, detected in endoscopy-2 where HGD was in the histology window, and an identical structure derived from the adenocarcinoma in endoscopy 4. **i)** The structure of ecDNA-2, detected in endoscopy-3 where EAC was present in the histology window, and an identical structure derived from the adenocarcinoma in endoscopy-4.


a
Amplicon 1: A1
An amplicon defines two collections, B and G
Breakpoint edges: B1
Genomic copy number
j
k
Amplified genomic intervals: G1
Genomic reference coordinates
Amplicon 2: A2
Breakpoint edges: B2
Genomic copy number
j
k
Amplified genomic intervals: G2
Genomic reference coordinates
b
B1
B2
G1
G2
c
B1 ∩ B2
G1 ∩ G2
d
Asymmetric similarity score Asym, for overlapping amplicons :=
Symmetric similarity score Sym for overlapping amplicons :=

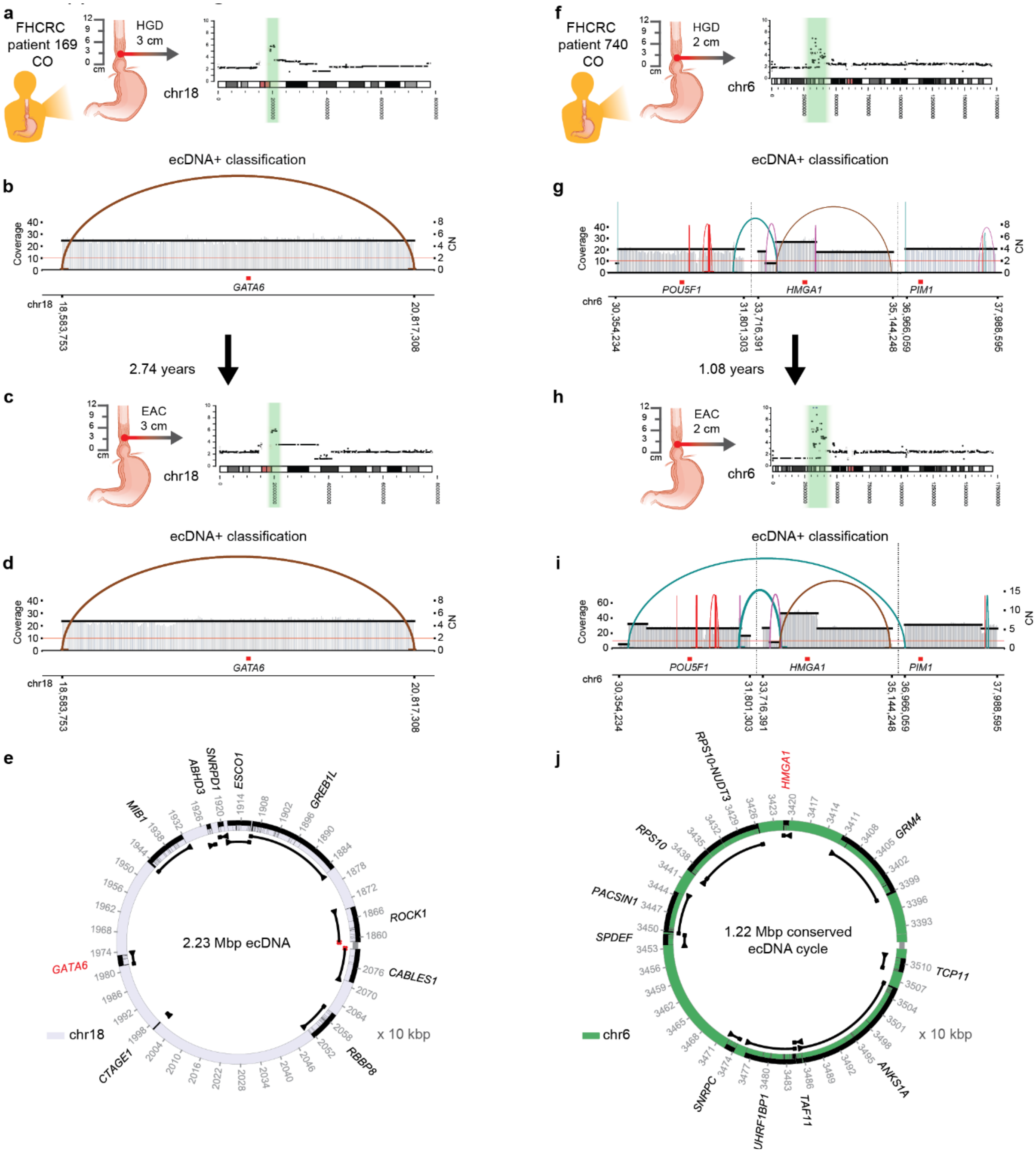
$$Asym(A_1, A_2) = \frac{\alpha(G_1 \cap G_2)}{G_1} + \frac{(1-\alpha)(B_1 \cap B_2)}{B_1}$$

$$Asym(A_2, A_1) = \frac{\alpha(G_2 \cap G_1)}{G_2} + \frac{(1-\alpha)(B_2 \cap B_1)}{B_2}$$

$$\longrightarrow \quad Sym(A_1, A_2) = \frac{Asym(A_1, A_2) + Asym(A_2, A_1)}{2}$$

Figure 4.13: Overlapping amplicon similarity score. **a)** Cartoon representation of two overlapping focal amplifications (A_1, A_2) consisting of a collection of genomic intervals (G_i) and breakpoints (B_i). Genomic location is shown on the x-axis and copy number on the y-axis. **b)** Representation of the relative locations of B_1, B_2 and G_1, G_2 and the resulting union of those elements. **c)** Representation of the intersection of the elements in B_1, B_2 and G_1, G_2 highlighted in purple and green, respectively. **d)** (Left) Definition of the asymmetric similarity score function *Asym* for two overlapping amplicons. (Right) Definition of the symmetric similarity score, *Sym* for two overlapping amplicons, which is the average of the asymmetric scores.

Figure 4.14: Tracking ecDNA over time. **a)** The time-point T1 sample for FHCRC patient 169 (CO) was found to have HGD within the histology window of the WGS sample taken 3cm from the GEJ. The WGS sample exhibited signs of a focal amplification on chr18. **b)** Subsequent analysis of the focal amplification seed region by AA generated a breakpoint graph encompassing oncogene *GATA6* with a single cycle encompassing it. Colored arcs in the AA diagram indicate SV junctions and are colored by orientation of the junction. **c)** 2.74 years following the T1 sample, the T2 sample for patient 169 was found to have EAC within the histology window of the WGS sample taken 3cm from the GEJ. The T2 3cm WGS sample exhibited similar signs of a focal amplification candidate on chr18. **d)** Subsequent analysis with AA yielded a breakpoint graph having identical structure as the T1 breakpoint graph. **e)** CycleViz visualization of the putative 2.23 Mbp ecDNA structure suggested by the breakpoint graph, containing *GATA6* and other genes. **f)** The time-point T1 sample for FHCRC patient 740 (CO) was found to have HGD within the histology window of the WGS sample taken 2cm from the GEJ. The WGS sample exhibited signs of a focal amplification on chr6. **g)** Subsequent analysis of the focal amplification seed region by AA generated a breakpoint graph encompassing oncogenes *POU5F1*, *HMGA1* and *PIM1*. The breakpoint graph yielded a trivial cycle capturing HMGA1. **h)** 1.08 years following the T1 sample, the T2 sample for patient 740 was found to have EAC within the histology window of the WGS sample taken 2cm from the GEJ. The T2 2cm WGS sample exhibited similar signs of a focal amplification candidate on chr6. **i)** Subsequent analysis with AA yielded a breakpoint graph having additional breakpoint edges but an identical cyclic substructure containing *HMGA1*, however with an increased copy number from T1. **j)** CycleViz visualization of the putative 1.22 Mbp conserved ecDNA structure suggested by the breakpoint graph, containing *GATA6* and other genes.


a
FHCRC
patient 169
CO
HGD
3 cm
cm
chr18
ecDNA+ classification
b
Coverage
CN
GATA6
chr18
18,583,753
20,817,308
2.74 years
c
EAC
3 cm
chr18
ecDNA+ classification
d
GATA6
18,583,753
20,817,308
e
2.23 Mbp ecDNA
chr18
x 10 kbp
MIB1
ABHD3
SNRPD1
ESCO1
GREB1L
ROCK1
CABLES1
RBBP8
CTAGE1
GATA6
f
FHCRC
patient 740
CO
HGD
2 cm
chr6
ecDNA+ classification
g
POU5F1
HMGA1
PIM1
chr6
30,354,234
31,801,303
33,716,391
35,144,248
36,966,059
37,988,595
1.08 years
h
EAC
2 cm
chr6
ecDNA+ classification
i
POU5F1
HMGA1
PIM1
30,354,234
31,801,303
33,716,391
35,144,248
36,966,059
37,988,595
j
1.22 Mbp conserved
ecDNA cycle
chr6
x 10 kbp
RPS10-NUDT3
HMGA1
GRM4
RPS10
PACSIN1
SPDEF
TCP11
ANKS1A
TAF11
UHRF1BP1
SNRPC

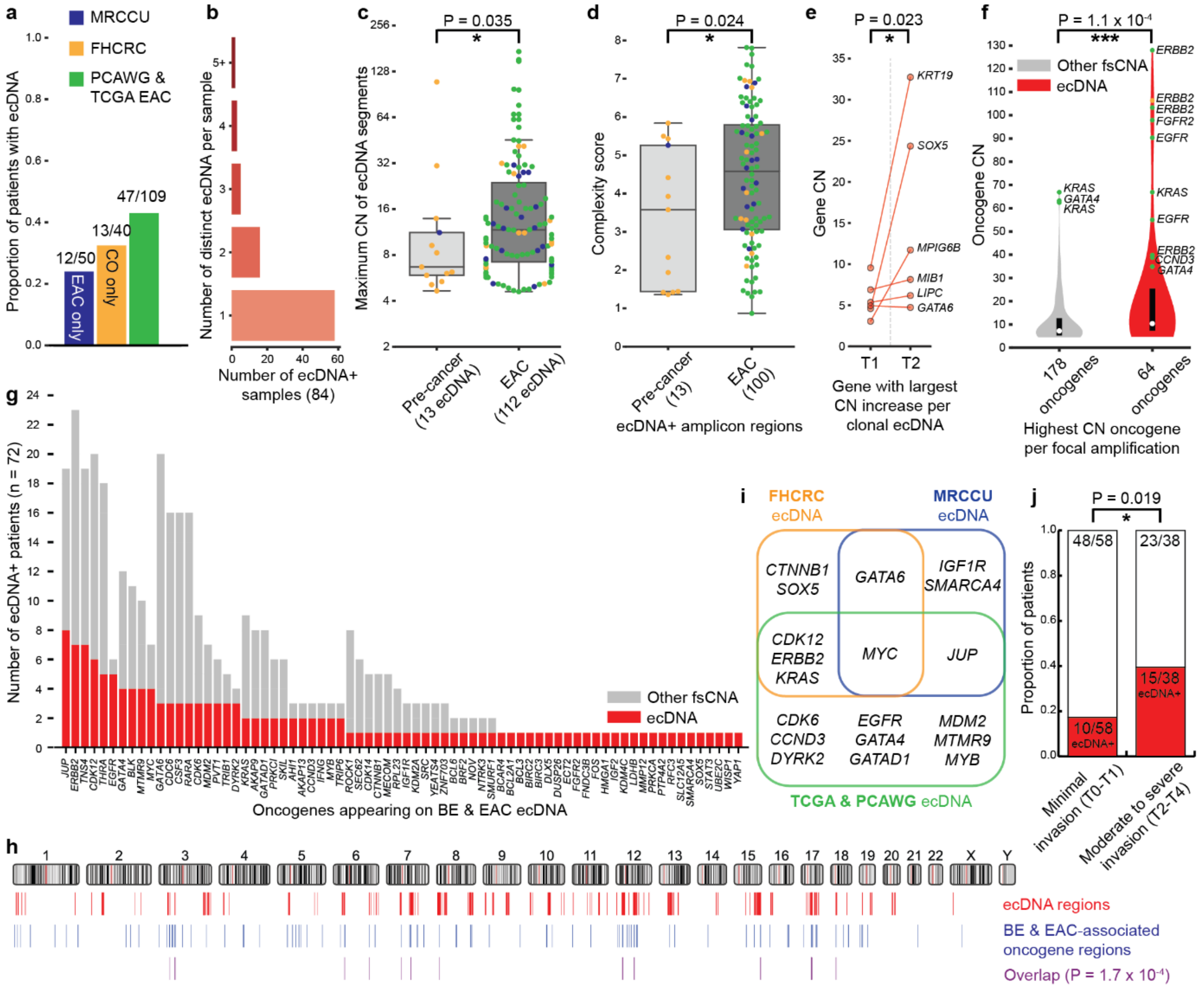


Figure 4.15: Characterization of ecDNA in BE precancer and EAC. **a)** The proportion of patients with ecDNA in the MRCCU study (EAC only shown), FHCRC (CO only shown), and combined ICGC & TCGA EAC samples. **b)** The number of distinct ecDNA per sample identified in ecDNA-positive samples from all combined sources of data. **c)** The maximum genomic copy number of ecDNA segments in pre-cancer samples and EAC samples, colored by sample study source. **d)** The complexity score of focally amplified ecDNA+ genomic regions for pre-cancer and EAC samples. **e)** The largest gene copy number increase for ecDNA which reappeared in samples from the same FHCRC patients taken from T1 and T2. **f)** Oncogene copy number for the highest copy number focally amplified oncogene in each sample having ecDNA or non-ecDNA fsCNA. **g)** The number of ecDNA+ patients of any BE or EAC study source having the oncogene listed at bottom on a focal amplification, where each oncogene appeared on at least one ecDNA. **h)** The locations of BE or EAC ecDNA regions (combined cohorts) in the human genome as well as the locations of canonical BE- and EAC-associated genes. The overlap of regions is shown at bottom with overlap significance computed by ISTAT. **i)** Venn diagram of canonical BE- and EAC-associated genes carried on ecDNA in each study source. **j)** Proportion of EAC patients with ecDNA in any study source where the patient had AJCC T-staging data available, segregated by degree of tumor invasion.

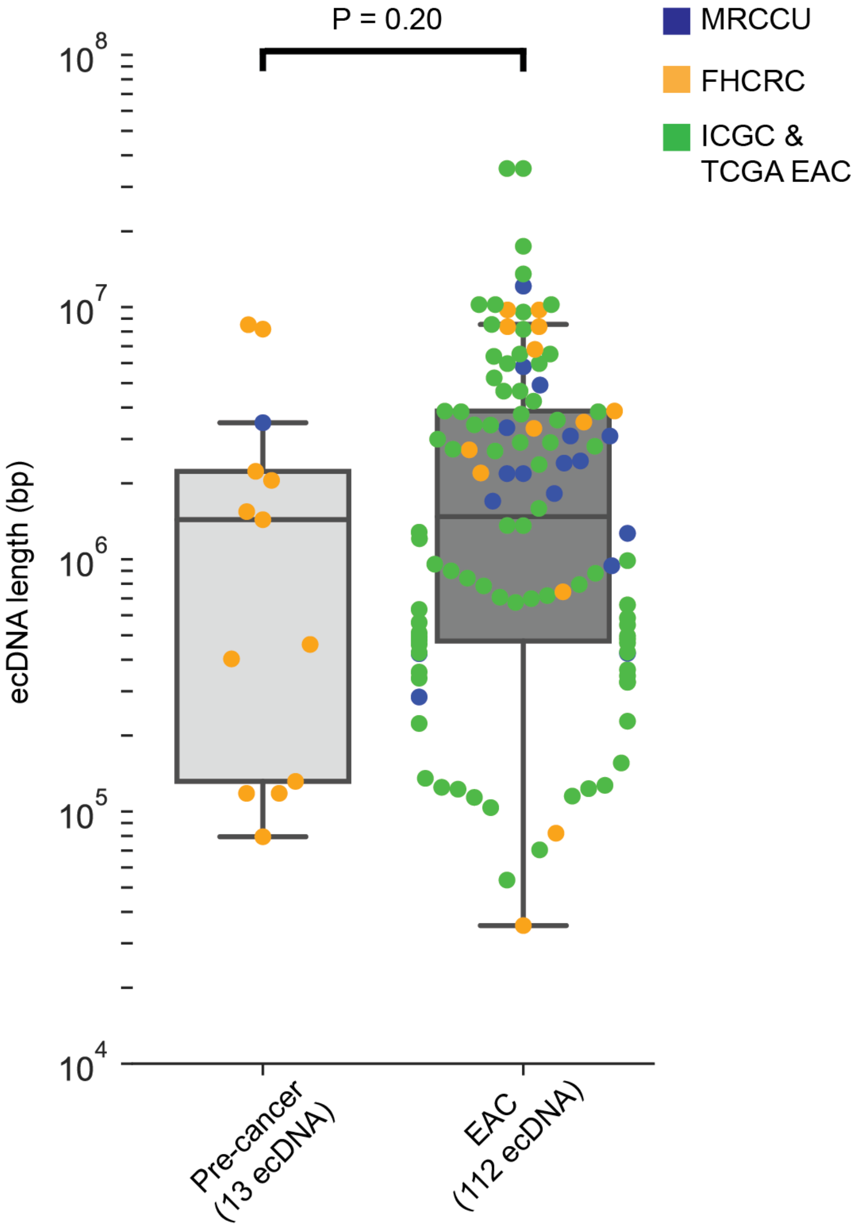


Figure 4.16: The length of putative ecDNA regions, visualized on log10 scale, for each distinct ecDNA in any sample from MRCCU, FHCRC, and ICGC & TCGA EAC, segregated by precancer versus EAC.

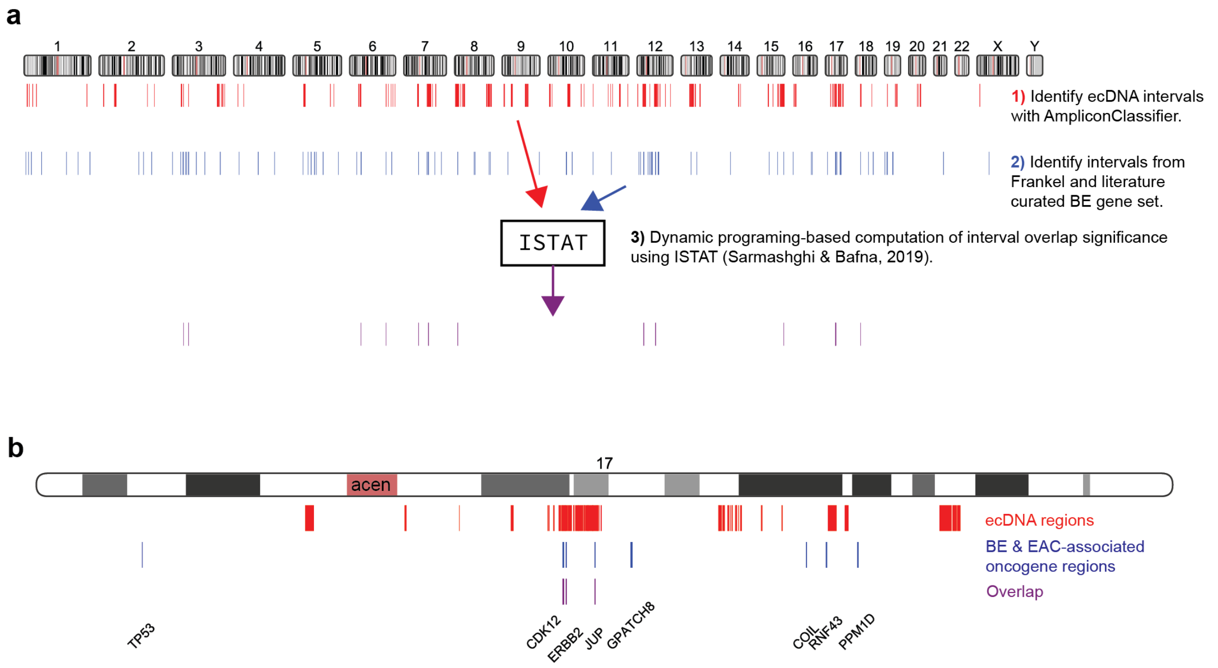

a
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
1) Identify ecDNA intervals with AmpliconClassifier.
2) Identify intervals from Frankel and literature curated BE gene set.
ISTAT
3) Dynamic programing-based computation of interval overlap significance using ISTAT (Sarmashghi & Bafna, 2019).
b
17
acen
ecDNA regions
BE & EAC-associated oncogene regions
Overlap
TP53
CDK12
ERBB2
JUP
GPATCH8
COIL
RNF43
PPM1D

Figure 4.17: EcDNA overlap computation diagrams. **a)** EcDNA and oncogene overlap computation annotated diagram showing how intervals were selected, and the methodology used to compute the overlap statistical significance. EcDNA regions were derived from any ecDNA+ sample identified in our study. **b)** Illustration of the overlap between ecDNA regions and canonical BE- and EAC-associated oncogenes for chr17.

4.7 References

1. R.C. Fitzgerald, Molecular basis of Barrett's oesophagus and oesophageal adenocarcinoma. *Gut*. **55**, 1810–1818 (2006).

2. B. J. Reid, T. G. Paulson, X. Li, Genetic Insights in Barrett's Esophagus and Esophageal Adenocarcinoma. *Gastroenterology*. **149**, 1142 (2015).

3. S. Killcoyne, E. Gregson, D. C. Wedge, D. J. Woodcock, M. D. Eldridge, R. de la Rue, A. Miremadi, S. Abbas, A. Blasko, C. Kosmidou, W. Januszewicz, A. V. Jenkins, M. Gerstung, R. C. Fitzgerald, Genomic copy number predicts esophageal cancer years before transformation. *Nat. Med.* **26**, 1726–1732 (2020).

4. T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhang, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, R. Beroukhim, Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

5. 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. *Nat. Genet.*, 1–7 (2020).

6. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell*. **144** (2011), pp. 646–674.

7. G. R. Bignell, C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, S. Widaa, J. Hinton, C. Fahey, B. Fu, S. Swamy, G. L. Dalgliesh, B. T. Teh, P. Deloukas, F. Yang, P. J. Campbell, P. A. Futreal, M. R. Stratton, Signatures of mutation and selection in the cancer genome. *Nature*. **463**, 893–898 (2010).

8. D. Stuart, W. R. Sellers, Linking somatic genetic alterations in cancer to therapeutics. *Curr. Opin. Cell Biol.* **21** (2009), pp. 304–310.

9. 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, P. S. Mischel, Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*. **543**, 122–125 (2017).

10. P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, P. Roepman, M. Voda, H. J. Bloemendal, V. C. G. Tjan-Heijnen, C. M. L. van Herpen, M. Labots, P. O. Witteveen, E. F. Smit, S. Sleijfer, E. E. Voest, E. Cuppen, Pan-cancer whole-genome analyses of metastatic solid tumours. *Nat. 2019 5757781*. **575**, 210–216 (2019).

11. S. M. Carroll, M. L. DeRose, P. Gaudray, C. M. Moore, D. R. Needham-Vandevanter,

D. D. Von Hoff, G. M. Wahl, Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Mol. Cell. Biol.* **8**, 1525–1533 (1988).

12. D. Cox, C. Yuncken, A. I. Spriggs, Minute chomatin bodies in malignant tumours of childhood. *Lancet (London, England)*. **1**, 55–58 (1965).

13. 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, P. S. Mischel, Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. *Science (80-.).* **343**, 72–76 (2014).

14. B. McClintock, The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics*. **26**, 234–82 (1941).

15. S. Zakov, M. Kinsella, V. Bafna, An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5546–51 (2013).

16. P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, P. J. Campbell, Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*. **144**, 27–40 (2011).

17. P. Ly, S. F. Brunner, O. Shoshani, D. H. Kim, W. Lan, T. Pyntikova, A. M. Flanagan, S. Behjati, D. C. Page, P. J. Campbell, D. W. Cleveland, Chromosome Segregation Errors Generate a Diverse Spectrum of Simple and Complex Genomic Rearrangements. *Nat. Genet.* **51**, 705 (2019).

18. U. NT, Z. CZ, L. LD, B. LJ, C. AM, T. R, S. L, A. HF, J. K, M. TJ, S. A, P. D, Mechanisms generating cancer genome complexity from a single cell division error. *Science*. **368** (2020), doi:10.1126/SCIENCE.ABA0712.

19. Y. Oobatake, N. Shimizu, Double-strand breakage in the extrachromosomal double minutes triggers their aggregation in the nucleus, micronucleation, and morphological transformation. *Genes. Chromosomes Cancer*. **59**, 133–143 (2020).

20. 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* (2019), doi:10.1038/s41586-019-1763-5.

21. A. R. Morton, N. Dogan-Artun, Z. J. Faber, G. MacLeod, C. F. Bartels, M. S. Piazza, K. C. Allan, S. C. Mack, X. Wang, R. C. Gimple, Q. Wu, B. P. Rubin, S. Shetty, S. Angers, P. B. Dirks, R. C. Sallari, M. Lupien, J. N. Rich, P. C. Scacheri, Functional Enhancers

Shape Extrachromosomal Oncogene Amplifications. *Cell* (2019), doi:10.1016/j.cell.2019.10.039.

22. O. Shoshani, S. F. Brunner, R. Yaeger, P. Ly, Y. Nechemia-Arbely, D. H. Kim, R. Fang, G. A. Castillon, M. Yu, J. S. Z. Li, Y. Sun, M. H. Ellisman, B. Ren, P. J. Campbell, D. W. Cleveland, Chromothripsis drives the evolution of gene amplification in cancer. *Nature*, 1–5 (2020).

23. V. Deshpande, J. Luebeck, N. P. D. Nguyen, M. Bakhtiari, K. M. Turner, R. Schwab, H. Carter, P. S. Mischel, V. Bafna, Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10** (2019), doi:10.1038/s41467-018-08200-y.

24. C. T. Miller, J. R. Moy, L. Lin, M. Schipper, D. Normolle, D. E. Brenner, M. D. Iannettoni, M. B. Orringer, D. G. Beer, Gene Amplification in Esophageal Adenocarcinomas and Barrett's with High-Grade Dysplasia. *Clin. Cancer Res.* **9**, 4819–4825 (2003).

25. C. M. Eischen, Genome Stability Requires p53. *Cold Spring Harb. Perspect. Med.* **6** (2016), doi:10.1101/CSHPERSPECT.A026096.

26. H. W, M. UM, Links between mutant p53 and genomic instability. *J. Cell. Biochem.* **113**, 433–439 (2012).

27. K. Curtius, J. H. Rubenstein, A. Chak, J. M. Inadomi, Computational modelling suggests that Barrett's oesophagus may be the precursor of all oesophageal adenocarcinomas. *Gut*. **70**, 1435–1440 (2021).

28. M. Secrier, X. Li, N. De Silva, M. D. Eldridge, G. Contino, J. Bornschein, S. Macrae, N. Grehan, M. O'Donovan, A. Miremadi, T. P. Yang, L. Bower, H. Chettouh, J. Crawte, N. Galeano-Dalmau, A. Grabowska, J. Saunders, T. Underwood, N. Waddell, A. P. Barbour, B. Nutzinger, A. Achilleos, P. A. W. Edwards, A. G. Lynch, S. Tavaré, R. C. Fitzgerald, A. Noorani, R. F. Elliott, J. Weaver, C. Ross-Innes, L. Smith, Z. Abdullahi, R. De La Rue, A. Cluroe, S. Malhotra, R. Hardwick, H. Ford, M. L. Smith, J. Davies, R. Turkington, S. J. Hayes, Y. Ang, S. R. Preston, S. Oakes, I. Bagwan, V. Save, R. J. E. Skipworth, T. R. Hupp, J. R. O'Neill, O. Tucker, P. Taniere, F. Noble, J. Owsley, L. Lovat, R. Haidry, V. Eneh, C. Crichton, H. Barr, N. Shepherd, O. Old, J. Lagergren, J. Gossage, A. Davies, F. Chang, J. Zylstra, G. Sanders, R. Berrisford, C. Harden, D. Bunting, M. Lewis, E. Cheong, B. Kumar, S. L. Parsons, I. Soomro, P. Kaye, P. Collier, L. Igali, I. Welch, M. Scott, S. Sothi, S. Suortamo, S. Lishman, D. Beardsmore, H. E. Francies, M. J. Garnett, J. V. Pearson, K. Nones, A. M. Patch, S. M. Grimmond, Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131 (2016).

29. A. M. Frankell, S. G. Jammula, X. Li, G. Contino, S. Killcoyne, S. Abbas, J. Perner, L. Bower, G. Devonshire, E. Ococks, N. Grehan, J. Mok, M. O'Donovan, S. MacRae, M. D. Eldridge, S. Tavaré, R. C. Fitzgerald, A. Noorani, P. A. W. Edwards, N. Grehan, B. Nutzinger, C. Hughes, E. Fidziukiewicz, S. MacRae, A. Northrop, G. Contino, X. Li, R. de la Rue, A. Katz-Summercorn, S. Abbas, D. Loureda, M. O'Donovan, A. Miremadi, S. Malhotra, M. Tripathi, S. Tavaré, A. G. Lynch, M. Eldridge, M. Secrier, G. Devonshire, J. Perner, S. G. Jammula, J. Davies, C. Crichton, N. Carroll, P. Safranek, A. Hindmarsh, V. Sujendran, S. J. Hayes, Y. Ang, A. Sharrocks, S. R. Preston, S. Oakes, I. Bagwan,

V. Save, R. J. E. Skipworth, T. R. Hupp, J. R. O'Neill, O. Tucker, A. Beggs, P. Taniere, S. Puig, T. J. Underwood, R. C. Walker, B. L. Grace, H. Barr, N. Shepherd, O. Old, J. Lagergren, J. Gossage, A. Davies, F. Chang, J. Zylstra, U. Mahadeva, V. Goh, F. D. Ciccarelli, G. Sanders, R. Berrisford, C. Harden, M. Lewis, E. Cheong, B. Kumar, S. L. Parsons, I. Soomro, P. Kaye, J. Saunders, L. Lovat, R. Haidry, L. Igali, M. Scott, S. Sothi, S. Suortamo, S. Lishman, G. B. Hanna, K. Moorthy, C. J. Peters, A. Grabowska, R. Turkington, D. McManus, H. Coleman, D. Khoo, W. Fickling, R. C. Fitzgerald, The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nat. Genet.* **51**, 506–516 (2019).

30. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013) (available at http://arxiv.org/abs/1303.3997).

31. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

32. K. Hadi, X. Yao, J. M. Behr, A. Deshpande, C. Xanthopoulakis, H. Tian, S. Kudman, J. Rosiene, M. Darmofal, J. DeRose, R. Mortensen, E. M. Adney, A. Shaiber, Z. Gajic, M. Sigouros, K. Eng, J. A. Wala, K. O. Wrzeszczyński, K. Arora, M. Shah, A. K. Emde, V. Felice, M. O. Frank, R. B. Darnell, M. Ghandi, F. Huang, S. Dewhurst, J. Maciejowski, T. de Lange, J. Setton, N. Riaz, J. S. Reis-Filho, S. Powell, D. A. Knowles, E. Reznik, B. Mishra, R. Beroukhim, M. C. Zody, N. Robine, K. M. Oman, C. A. Sanchez, M. K. Kuhner, L. P. Smith, P. C. Galipeau, T. G. Paulson, B. J. Reid, X. Li, D. Wilkes, A. Sboner, J. M. Mosquera, O. Elemento, M. Imielinski, Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. *Cell.* **183**, 197-210.e32 (2020).

33. E. Talevich, A. H. Shain, T. Botton, B. C. Bastian, CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12** (2016), doi:10.1371/JOURNAL.PCBI.1004873.

34. P. Van Loo, G. Nilsen, S. H. Nordgard, H. K. M. Vollan, A. L. Børresen-Dale, V. N. Kristensen, O. C. Lingjærde, Analyzing cancer samples with SNP arrays. *Methods Mol. Biol.* **802**, 57–72 (2012).

35. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. Pietro Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G. L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N.

Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*. **17**, 261–272 (2020).

36. S. Sarmashghi, V. Bafna, Computing the statistical significance of overlap between genome annotations with iStat. *Cell Syst.* **8**, 523 (2019).

37. Y. Liu, J. Sun, M. Zhao, ONGene: A literature-based database for human oncogenes. *J. Genet. Genomics*. **44**, 119–121 (2017).

38. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6**, 80–92 (2012).

39. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

40. S. Kim, K. Scheffler, A. L. Halpern, M. A. Bekritsky, E. Noh, M. Källberg, X. Chen, Y. Kim, D. Beyter, P. Krusche, C. T. Saunders, Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*. **15**, 591–594 (2018).

41. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).

42. A. C. Decarvalho, H. Kim, L. M. Poisson, M. E. Winn, C. Mueller, D. Cherba, J. Koeman, S. Seth, A. Protopopov, M. Felicella, S. Zheng, A. Multani, Y. Jiang, J. Zhang, D. H. Nam, E. F. Petricoin, L. Chin, T. Mikkelsen, R. G. W. Verhaak, Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.* **50**, 708–717 (2018).