

# Computational methods for regulating transcription and translation

*Sanjit Batra*



Electrical Engineering and Computer Sciences  
University of California, Berkeley

Technical Report No. UCB/EECS-2023-58

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-58.html>

May 1, 2023

Copyright © 2023, by the author(s).  
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Computational methods for regulating transcription and translation

by

Sanjit Singh Batra

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair

Professor Nilah Ioannidis

Professor Dirk Hockemeyer

Summer 2022

Computational methods for regulating transcription and translation

Copyright 2022  
by  
Sanjit Singh Batra

## Abstract

Computational methods for regulating transcription and translation

by

Sanjit Singh Batra

Doctor of Philosophy in Computer Science

and the Designated Emphasis in

Computational and Genomic Biology

University of California, Berkeley

Professor Yun S. Song, Chair

The central dogma describes the transformation of DNA into mRNA and consequently into a protein. Any of these three stages could be dysregulated in a disease. In this work, we develop three computational tools aimed at better understanding how diseases such as cancer might affect these different stages. The first method describes an approach to detect and track changes in DNA caused by cancer, such as large-scale structural variants. The second method investigates whether the latest advancements in CRISPR can be leveraged to restore balance to gene regulation that might have been disturbed by disease. Finally, the third method provides a way to detect if mutations affect the abundance of a protein thereby causing disease. Together, these methods span the impact of diseases such as cancer on the central dogma of biology and pave the way for a better understanding of underlying mechanisms and future therapies.

To Mama and Papa for everything.

# Contents

<b>Contents</b>	ii
<b>List of Figures</b>	iv
<b>List of Tables</b>	ix
<b>1 Introduction</b>	1
1.1 Overview of the Thesis . . . . .	2
<b>2 Tracing cancer evolution using Hi-C</b>	3
2.1 Introduction . . . . .	3
2.2 Results . . . . .	5
2.3 Discussion . . . . .	20
2.4 Online Methods . . . . .	22
<b>3 Predicting gene expression</b>	28
3.1 Introduction . . . . .	28
3.2 Results . . . . .	29
3.3 Discussion . . . . .	35
3.4 Methods . . . . .	36
<b>4 Predicting translation initiation</b>	39
4.1 Introduction . . . . .	39
4.2 Results . . . . .	40
4.3 Discussion . . . . .	45
4.4 Methods . . . . .	46
<b>5 Conclusions</b>	48
5.1 Future Work . . . . .	48
<b>Bibliography</b>	49
<b>A HiDENSEC Supplementary Information</b>	60

A.1 Notation & Setup . . . . .	60
A.2 Covariate correction . . . . .	61
A.3 Inference of copy numbers & mixture proportions . . . . .	63
A.4 Inference of large-scale structural variants . . . . .	69
A.5 Proof of Theorem . . . . .	73
A.6 Supplementary Figures . . . . .	76
A.7 Supplementary Tables . . . . .	91



# List of Figures

2.1	HiDENSEC pipeline	5
2.2	Validation of HiDENSEC using mixtures of samples	7
2.3	Benchmarking HiDENSEC's LSSV identification	9
2.4	HiDENSEC analysis of Patient 1	11
2.5	HiDENSEC analysis of Patient 2	13
2.6	HiDENSEC analysis of Patient 3	15
2.7	Evolution of the melanoma genome in Patient 3	17
3.1	Metagene plots for different cell types for unnormalized H3K27ac ChIP-seq data across gene expression quantiles (blue is the highest and red is the lowest gene expression quantile). * represents data from HEK293 instead and (A) represents Avocado imputed data.	30
3.2	Metagene plots for the six different histone marks in HEK293T, after S3norm-based normalization, across gene expression quantiles (blue is the highest and red is the lowest gene expression quantile). * represents data from HEK293 instead and (A) represents Avocado imputed data for HEK293T.	31
3.3	Spearman correlation on genes in the test dataset for different input context lengths. Blue curve is the mean across 10 computational replicates of CNNs and the red is the mean across 10 computational replicates of ridge regression. Shaded aread represents standard deviation in the spearman correlation across the 10 computational replicates.	32
3.4	Spearman correlation on genes of cell types held out during training. The barplots represent the mean across 10 computational replicates and the error bars represent their standard deviation.	33
3.5	Distribution of spearman correlation across cell types, computed for each gene in the test data. The different curves represent 10 computational replicates for each model type.	34
3.6	Each point on the X-axis corresponds to <i>in silico</i> perturbation of that assay at that position and the Y-axis measures the predicted fold-change. The five different lines correspond to five different genes in the HEK293T cell type.	35

4.1	Schematic for predicting translation initiation from mature mRNA sequence	40
4.2	Empirical cumulative distribution of uAUG properties can be used to define negative dataset for training the model	41
4.3	Accuracy metrics as uORF length cut-off varies (for determining the negative dataset)	42
4.4	Model predictions as an AUG is scanned across the input context, aggregated across 1000 genes	43
4.5	Probability of observing an uAUG in 5'UTRs of the human genome. X-axis at $x = 0$ denotes the mAUG of each gene.	43
4.6	Model predictions as a <i>strong</i> Kozak sequence, corresponding to AC-CATGGCG is scanned across the input context, whose mAUG exists in a <i>weak</i> Kozak context of TTAATGATT	44
4.7	Variants with high model scores tend to be more rare in gnomAD. Stars represent significant odds ratios. If we assume that <i>all</i> uAUG variants are pathogenic, the corresponding odds ratio obtained in gnomAD is reflected by the dashed line.	45
A.1	Covariate dependence of $H^j = \sum_k H_{jk}$ in GM12878 <i>in situ</i> Hi-C data. The impact of the four covariates compartment structure, GC-content, number of cut-sites and mappability on row sums of Hi-C intensity matrices is displayed. <b>A:</b> $H^j$ conditioned on compartment structure, <b>B:</b> $H^j$ as a function of remaining three covariates; points are coloured by compartment, <b>C:</b> $H^j$ conditioned on quartiles of the corresponding column statistic in B, as a function of the two remaining covariates.	76
A.2	Covariate dependence of $H^j = \sum_k H_{jk}$ in Sample 1-I <i>in vivo</i> Fix-C data. Plots are as described in <b>Supplementary Figure A.1.</b>	77
A.3	Predictions of tumor purity correlate well with known tumor purity. (a) The x-axis represents true tumor purity for in vitro and in silico samples; while the y-axis represents the HiDENSEC inferred tumor purities. There is high concordance between the two and the error bars represent 95% confidence intervals for the HiDENSEC inferred tumor purity. (b) The x-axis represents different samples used in the analysis (1-II represents Sample 1-II, for instance). The y-axis represents the inferred tumor purity (the fraction of cells that are cancerous). The error bars represent 95% confidence intervals for the inferred tumor purity.	78

<b>A.4 Off-diagonal intensities of LSSVs aren't reliable estimators of absolute copy number.</b> Each panel represents the Hi-C intensities corresponding to a particular off-diagonal event in the HCC1187C cell line (for instance the top-left panel represents a translocation between chromosome 1 and chromosome 6). The horizontal axis represents a measure of how large a window around the translocation (which manifests itself as an off-diagonal event on the Hi-C map) was considered to compute the Hi-C intensity. The Y-axis represents the resulting Hi-C intensity. The true ploidies of inter-chromosomal translocations are denoted by the horizontal dotted lines while the colored curves represent the measured Hi-C intensities. For balanced translocations there are two colored lines corresponding to the two fusion events. The fact that the ratio of the true ploidy represented by the horizontal dotted lines and the colored curves is not consistent across the various LSSVs within the same sample, suggests that the off-diagonal intensities are confounded by covariates in addition to absolute copy number and hence HiDENSEC does not use off-diagonal Hi-C intensities to infer absolute copy numbers. . . . .	79
<b>A.5 HiDENSEC absolute copy number predictions correlate well with genome-wide copy numbers inferred from next-generation sequencing.</b> Each panel represents HiDENSEC absolute copy number predictions for a particular sample (whose identity is indicated inside each panel; e.g., 1-II in the top-right represents Sample 1 - II), compared to UCSF500 or Exome sequencing based (relative) copy number calls. Copy number profiles inferred by HiDENSEC from Hi-C data use color codes consistent with the main text (that is, blue curves correspond to Samples 1 - I, 2 - I, 3 - I, beige curves to Samples 1 - II, 2 - II, 3 - II, and the green curve depicts Sample 3 - III), while red and pink profiles represent relative copy numbers (transformed by $x \rightarrow 2 \times 2^x$ ) from CNVkit using UCSF500 or Exome sequencing. For Sample 2 - I, the discordance between the levels of the HiDENSEC absolute copy numbers and the CNVkit scaled relative copy numbers inferred using UCSF500 data is likely due to differences in samples for the Hi-C data and for the UCSF500 data, since the UCSF500 based tumor purity lies outside the 95% confidence intervals of the HiDENSEC inferred tumor purity ( <b>Supplementary Figure A.3, Supplementary Table 4</b> ). For Sample 3 - III, two UCSF500 based curves are displayed, as UCSF500 data from two different metastases corresponding to this sample exists; both of these are concordant with the HiDENSEC absolute copy number inferred using Hi-C data from the metastasis sample, Sample 3 - III. . . . .	80

<b>A.6 Somatic mutant allele frequencies for Patient 1 and Patient 2, based on somatic variant calls derived from Exome sequencing data and UCSF500 data, respectively. (a)</b> Somatic variant calls were obtained using Mutect2 from exome sequencing data in samples derived from Patient 1 and <b>(b)</b> UCSF500 data in samples derived from Patient 2. These somatic variant calls were then filtered out for false positives and the intersection of mutations observed in both samples within a patient was considered. The x-axis and the y-axis in each of the two panels represents somatic mutant allele frequencies. Each individual data point is a particular somatic variant call, from among the intersection of filtered somatic variant calls within the two samples of a patient. The red point denotes the well-known BRAF V600E somatic variant while the dashed lines are drawn at exactly half of its mutant allele frequencies. The resulting quadrants are intended to denote somatic variants common to both samples (corresponding to the top-right quadrant), somatic variants present in only one of the samples (corresponding to the top-left and the bottom-right quadrants), and somatic variants which are likely false positives (bottom-left quadrant). . . . .	81
<b>A.7 Schematics demonstrating two particularly complex structural variants.</b> The translocations involving chromosome 2, 5 and 10 in both samples of Patient 2 (Sample 2 - I and Sample 2 - II) and the inter-chromosomal translocation between chromosome 5 and 7 in Sample - I have been depicted here. In order to infer the contacts constituting these complex structural variants, the Hi-C maps of the relevant chromosomes were carefully analyzed in conjunction with the HiDENSEC inferred copy numbers, as shown in <b>Supplementary Figure A.9, Supplementary Figure A.10</b> . Arrows indicate duplications and inversions of genomic segments of chromosome 2. . . . .	82
<b>A.8 Alternatives to the phylogenetic relationship between the three cell types in Patient 3 .</b> Two alternative phylogenies that are consistent with the LSSVs in Patient 3. The phylogeny in <b>Figure 7b</b> was chosen over these two alternatives following the principle of parsimony since the number of convergent events, denoted in red, are higher in these two phylogenies, than the one described in <b>Figure 7b</b> . . . . .	83
<b>A.9 Zoomed-in Hi-C maps of two translocations observed in Sample 3 - I.</b> Sample 3 - I contains two structural variants that are relatively smaller in size than chromosome arms. The first of which is a complex structural variant between chromosome 5 and chromosome 7, a schematic of which is depicted in <b>Supplementary Figure A.7b</b> . The second structural variant is a short balanced translocation between chromosome 17 and chromosome 19, depicted in the bottom right panel and its inset. . . . .	84

<b>A.10 Hi-C contact maps used to infer the inter-chromosomal translocations involving chromosomes 2, 5 and 10 in Patient 2.</b> This figure shows the zoomed in Hi-C maps for chromosome pairs involved in the complex structural event depicted in <b>Supplementary Figure A.7a</b> . The three line plots adjacent to the Hi-C maps represent HiDENSEC inferred copy numbers for the three chromosomes, which were used to determine the contacts constituting this complex structural variant. . . . .	85
<b>A.11 Covariate correction is protocol-dependent.</b> A Hi-C sample of the reference genome as well as the Fix-C Sample 3-II illustrate the necessity for both covariate correction in general, as well as its protocol-specific nature. . . . .	86
<b>A.12 Generative model of signal &amp; noise.</b> Observed on-diagonal contact intensities are modeled as an underlying effective copy number profile comprised of a convex combination of individual, cell-population-specific absolute copy number profiles ( <b>A</b> ), which is perturbed by heteroskedastic noise ( <b>B</b> ) and scaled by a generally unknown constant $C_0N$ . Under $\mathcal{H}_0$ of $\pi \equiv 2$ , $p$ -values associated with HiDENSEC's test statistics behave super-uniformly or close to uniform ( <b>C</b> ). . . . .	87
<b>A.13 Inference of effective copy number profiles &amp; interpretation of excursions.</b> Inferred $\hat{\pi}$ (solid black line) for each of the non-diploid samples discussed in the main text as well as for in-silico and in-vitro mixtures (with $f = 0.7$ and $f = 0.5$ , respectively) are shown against $\Pi$ (blue line). Each excursion $e$ is associated with a $p$ -value reflecting its biological significance, with greener colors mirroring higher significance. . . . .	88
<b>A.14 Hi-C intensity patterns and associated large-scale structural variants.</b> HiDENSEC detects Hi-C sub-matrices of six distinct patterns ( <b>A</b> ) associated with six types of large-scale structural variants ( <b>B</b> ) (note: non-fusing segments may interact with chromosomes other than $\chi_a$ and $\chi_b$ or be deleted without qualitatively affecting the local Hi-C patterns of ( <b>A</b> )). . . . .	89
<b>A.15 HiDENSEC reliably detects off-diagonal exchange patterns.</b> In those samples that do contain patterns in $\mathcal{P}_2$ , HiDENSEC correctly recovers them at zero false-positive rate ( <b>A</b> ), and identifies their precise locations accurately ( <b>B</b> , blue highlights indicate fusion sites inferred by HiDENSEC). Calibration of HiDENSEC is primarily a result of computed $p$ -values behaving super-uniformly ( <b>C</b> , empirical distributions based on all samples analyzed in the main text). Of the three events, HiNT only detected the $\chi_4 \sim \chi_8$ fusion, locating it, however, $\approx 42$ and $\approx 25$ MB away from the true signal on $\chi_4$ and $\chi_8$ , respectively. . . . .	90
<b>A.16 Comparison of HiDENSEC's (black) and HiNT's (red) top-<math>k</math> recall on the samples analyzed in the main text.</b> As in the corresponding main figure, filled regions indicate rearrangements deemed significant by either method. . . . .	91

## List of Tables

3.1	ChIP-seq $-\log_{10}$ (p-values) were obtained from the ENCODE Imputation Challenge where the ground truth data were available (corresponding to entries labeled <b>T</b> in the table). Avocado imputations were downloaded from the ENCODE data portal , where ground truth data were not available (corresponding to entries labeled <b>A</b> in the table). Entries labeled with an * are obtained from the HEK293 cell line because data for the HEK293T cell line was not available . . . . .	30
-----	---	----

## Acknowledgments

To start at the beginning, my first real research project was at Baylor College of Medicine with Professor Erez Aiden and Olga Dudchenko. This was my first exposure to systematic scientific research and genomics. Thanks to both of you, and to others in the Aiden Lab for inspiring the next 8 years of my research.

Upon my return to my alma mater, Indian Institute of Technology Delhi for my Masters degree, I began working on machine learning methods in computational biology, where I worked with Professor Jayadeva. This work introduced me to interdisciplinary research at the intersection of computer science and biology, allowing me to closely collaborate with experimentalists and iterate, not only on the computational method development but also on the data generation process. Thanks to Professor Munishwar Nath Gupta for playing a pivotal role in providing guidance to my early research projects.

I want to thank my advisor, Professor Yun S. Song, for his advice and incredible support throughout my PhD. When I started working with Yun, I barely had any exposure to deep learning or the vast majority of the field of computational biology. It took me a year of fumbling around and visiting conferences to finally converge on variant effect prediction as one of the more exciting problems in the field for me. Yun encouraged me to explore the latest research and was hugely supportive and patient through this process. Over the course of the last five years, there have been innumerable occasions where Yun's insights have catapulted my research into novel and fascinating directions and for this I am grateful to Yun. Thank you Yun for being an ideal advisor and I hope that we will continue to have a wonderful working relationship moving forward.

When I decided I wanted to work on variant effect prediction for my PhD, I struggled for several months to find the right problem. It wasn't until Kyle Farh from the Illumina AI Lab reached out to Yun seeking collaborations on the exact same problem that I got my first research internship during my PhD. Kyle insights in scientific research are tremendous and combined with Kishore's generous guidance, I learnt a great deal over the last four years working at Illumina AI Lab. Thank you both for helping me get started on this journey of variant effect prediction, and for making research so much fun.

In order to leverage my experience working with Hi-C data, I collaborated with Daniel Rokhsar and Dirk Hockemeyer in the Biology department. Together with melanoma specialists at UCSF and experimentalists at Dovetail genomics, we generated a fascinating dataset for analyzing how melanoma affects DNA folding. I am grateful to all my collaborators for the learnings throughout this project.

While working on this project, I was co-living with my dear friend, Dan, which incidentally also coincided with the advent of COVID. Over the course of this year, Dan began working with me on this Hi-C project and this gave me the fortunate opportunity to learn from an incredible, researcher, friend and athlete. I learnt about running, biking, cooking, advanced probability, reasoning about political issues and numerous other things from Dan; learnings that shaped my world view in a significant way and helped me develop habits that will

continue to help me for long after graduate school. I am grateful to Dan for this time together and I hope that we stay in touch and remain good friends for long.

To explore deep learning and its applications in genomics, I began working with Jeffrey Spence, a senior graduate student in the Song Lab and found a mentor, friend and a model researcher. Jeff is, in my eyes, the epitome of kindness and intelligence. He has been instrumental in my scientific learning and exposure to deep learning while being a pillar of emotional support through hard times. Thank you for being you, Jeff. I hope that one day your advisees also get to experience your mentorship.

While working with Jeff, we serendipitously began collaborating with Alan Cabrera and Professor Isaac Hilton at Rice University. The Hilton Lab works at the cutting edge of experimental CRISPR techniques and it has been an eye-opening experience to work with competent experimentalists and have the opportunity to interact with them and devise data generation together as an interdisciplinary team. Hopefully, we will get to share our fascinating learnings with the broader scientific community soon.

I have been extremely fortunate to encounter some of the most intelligent yet humble people I have ever had the good fortune of meeting, over the course of graduate school. Nick, who has helped brainstorm many ideas has taught me the importance of meditation. Yutong, who makes amazing pineapple cookies is a fun presence to be around. Alan, who is a talented statistician always inspires me and reminds me that physical health and mental health complement each other. Milind, who has recently joined the lab, is one of the most sharp researchers I have met and I'm sure will do extremely well wherever he goes. Jessen, who I met in the Biology department, has helped me learn practical skills to become a more efficient programmer. Jessen has a calm aura and is a great mentor and a friend.

Over the course of the last two years in graduate school, to expose myself to research in industry, I serendipitously connected with Brad, who was leading a team at Google X working on some of the most exciting problems in plant genomics. I spent the better half of 2021 working with Brad and Mathias and learning a great deal from them about research in industry as well as how to start a company while building cutting-edge technology.

Another member of the Song Lab who coincidentally overlapped with me at Google X, Neil Thomas, became one of my best friends in graduate school. Neil is perhaps the single most authentic person I have met. He always has positive energy and his birthday parties have been remarkably unique and refreshing experiences. Neil has taught me the value of physical health, taught me climbing and inspired me to learn cooking and in general, the power of carving out time outside of graduate school to take care of my physical health. Over the last six months in graduate school, arguably the most intense period for me, Neil spent every morning with me co-working over Google Meet. The structure and discipline this regular check-in provided to me has been invaluable. I am sure that wherever Neil goes after graduating, and whoever is fortunate to spend time with him, will cherish him just as I do.

Nilah, with whom I had the pleasure of teaching *Machine Learning in Genomics*, has been a role model for me. Nilah, despite being a professor at Berkeley, is one of the kindest, most humble and polite individuals I have ever encountered. She is an amazing researcher while being forever responsive and supportive of new ideas. I hope to learn much more from



Nilah and I hope that we stay in touch in the future.

Upon my return from Google X, inspired by the world of plant genomics, I actively brainstormed with Gonzalo, another talented member of the Song Lab, who at the time was also seeking interesting new research directions. This culminated with what I can only describe as the most successful project I have had the opportunity to witness from its inception to publication. Gonzalo successfully navigated a really hard problem with his exceptional programming skills and humble perspective on life.

Last but not the least, all through my life, and through graduate school, my friends and family back home in India have been my pillars. Parth, who I consider my strongest support system outside of my immediate family, has kept me on track in many ways. We have traveled together, he has exposed me to some of my favorite music, been there when I broke up with my partners, and has been a constant voice of reason in my life. Rishabh, who moved to the USA during the course of the pandemic, has helped me learn the power of consistency; that discipline can be used to make life much easier and his organization skills are unmatched. He was also there at a moment's notice to play *FIFA* with me and Parth, if ever we wanted to decompress, which happened more often than I care to admit. *Tayaji*, and my *chacha*, Taranjit and Inderpal, have been my greatest inspirations growing up since their professional careers had a similar trajectory to mine. They have always been there to praise my successes and remind me about them, during my failures. It was a deeply emotional moment filled with immense gratitude to have them by my side on the day I was conferred my PhD. Thank you for always being there.

Mama, Papa, Jasnit, Harleen, Rohit, KD, Gurveer, Veer, and Promila *didi* are the most important people in my life. I have love and respect for them beyond words can describe. I will one day write a more extensive memoir of my gratitude for them but suffice it to say that I would not be where I am without them. Thank you from the bottom of my heart. I look forward to traveling more and spending more time with family after finishing graduate school.

# Chapter 1

## Introduction

Modulating phenotypes has been the holy grail of computational biology for decades. Such modulation can be achieved via multiple avenues, such as optimizing DNA sequence and its topology within the nucleus of a cell; by regulating gene expression, which is the amount of available mRNA, or by moderating protein production directly.

Advances in techniques to probe DNA folding in the nucleus have greatly improved our ability to understand and determine the factors responsible for changes in DNA conformation under various cell states, such as disease. Methods complementing these experimental advances are being developed at a staggering pace and continue to reveal hidden structures in these novel data types, such as Hi-C. However, there are vast scopes for improvement in these computational methods, both, in the detection of DNA conformation changes and in describing the causal mechanisms that lead to such changes due to disease.

Alternatively, the Nobel-prize winning discovery, CRISPR, which has revolutionized the field of genome editing, offers new avenues to regulate the amount of mRNA product within a cell and thereby modulate downstream phenotypes. Computational methods guiding experimentalists on where to perform CRISPR experiments in the genome are crucial to reducing the otherwise exponential search space. Furthermore, CRISPR technologies have been adapted to directly allow modulation of gene expression, such as CRISPRa and CRISPRi, instead of indirectly, by modifying the DNA sequence of cis-regulatory elements.

While this approach of modifying gene expression to modulate a downstream phenotype is quite promising, there is room for improvement by regulating protein expression directly. This is because gene expression is not perfectly correlated with protein expression (Spearman correlation of  $\sim 0.5$  in human tissues). Consequently, a better understanding of the translation process which converts mRNA into protein products could lead to more robust control of downstream phenotypes.

The latest developments in the field of deep learning have facilitated huge leaps in our understanding of biology and have enabled novel methodological advances in computational biology. Deep learning has a vast potential to impact the field of biology and it remains to be seen how these advances would dovetail with the latest experimental breakthroughs.

This thesis focuses on developing novel methods across all three stages of the central

dogma of biology, to facilitate a better understanding and design control over phenotypes of interest. As an equally important contribution, it also draws connections between the latest advancements in experimental techniques such as CRISPR and Ribosome Profiling, and deep learning and evaluates whether they can complement each other for improving our ability to modulate phenotypes.

## 1.1 Overview of the Thesis

This dissertation attempts to develop computational tools to answer three key questions related to the central dogma:

1. Can we detect changes in DNA such as chromosomal aberrations like large-scale structural variants using chromatin conformation capture data generated from cancer samples?
2. Can we leverage the latest developments in CRISPR/dCas9-based epigenome editing to regulate gene expression in cancer via post-transcriptional modifications?
3. For any cancer gene, can we predict if a mutation in the 5'UTR of a gene will have an impact on its protein expression?

**Chapter 2** focuses on the first question. We have developed a novel computational method which allows us to detect, quantify and track large-scale structural variants using Hi-C data which can be generated from solid cancer samples.

**Chapter 3** leverages the advances in CRISPR/dCas9-based epigenome editing to develop a computational method to guide where to edit the epigenome in order to achieve desired gene expression levels.

**Chapter 4** develops a novel method to assess the impact of upstream ORFs on protein expression. Such a computational method could facilitate computational prediction of mutations in the 5'UTR of genes that might significantly alter protein expression and thereby cause disease such as cancer.

**Chapter 5** concludes the dissertation, raises new questions, and discusses future work.

## Chapter 2

# Tracing cancer evolution using Hi-C

This is joint work with Dan Daniel Erdmann-Pham, Timothy Turkalo, James Durbin, Marco Blanchette, Iwei Yeh, Hunter Shain, Boris Bastian, Yun S. Song, Daniel Rokhsar and Dirk Hockemeyer and the manuscript is currently under preparation.

### 2.1 Introduction

Cancer progression is driven by ongoing selection for mutations that endow the evolving cancer cell with a proliferative advantage compared to its direct precursor and the surrounding normal tissue. In addition to positive selection for proliferation, neutral mutations can persist as bystanders over time, whereas mutations that reduce the fitness of cells are selected against. Recent cancer genome studies have significantly increased our understanding of how individual mutations drive cancer progression [1], [2], [3]. Since most cancer sequencing efforts rely on short-read sequencing approaches, we currently have a much better catalog of point mutations and small, regional genome alterations on cancer progression compared to LSSVs such as large-scale deletions, inversions, duplications and inter-chromosomal translocations [4], [5], [6], [7], [8].

The realization that the “Philadelphia chromosome” is a driver of cancer progression in chronic myelogenous leukemia resulted from the ability to visualize this recurrent translocation in metaphase spreads [9], [10]. The role of translocations during the initiation and early evolution of solid cancers, however, has been more difficult to study [11] since at early and premalignant stages of cancer development, the incipient cancer lesions are generally small and intermixed with normal cells from the surrounding tissue. To study structural rearrangements in the progression of solid cancer we focus on melanomas, as early-stage tumors and their precursor lesions (melanocytic nevi) are routinely excised from the skin of patients and are therefore available for study [12]. These precursors are typically initiated by activating point mutations in the MAP-kinase pathway [13], [14], [15]. As the melanoma progresses and invades deeper into the skin, genomic alterations are more often driven by LSSVs and copy number changes rather than UV exposure [16].

While LSSVs can be detected and quantified by spectral karyotyping [17], [18], karyotyping requires cell culture and generally can only detect LSSVs larger than 1-10 Mbp. Array CGH is a powerful tool to detect CNV but restricted in detecting copy number neutral changes such as inversions and reciprocal translocations [19], [20] and thus have limited applicability to solid cancer samples [21], [22]. To overcome this limitation, new methods that leverage short-read whole-genome sequencing have been developed to detect and map the breakpoints of chromosomal rearrangements down to a resolution of 100 bp [23], [24], [25] and can infer tumor purity, the percent of cancer cells present in a sample of tumor tissue, by combining detection of copy number alterations with loss of heterozygosity [26], [27], [28]. However, these methods rely on sequencing mate-pairs that span the newly generated fusion points of LSSVs and therefore require relatively high sequencing depth (at least 40x of the cancer genome) to avoid false positives [29]. This approach generally fails to detect break and fusion points in repetitive regions [30] such as centromeres or telomeres, which frequently are involved in LSSVs.

An alternative approach to detecting chromosomal rearrangements takes advantage of high-throughput chromatin conformation capture sequencing, also known as Hi-C [31]. In Hi-C, genomic loci making three-dimensional contact with each other are converted into linked read-pairs by proximity ligation of restriction-digested fixed chromatin [32]. These three-dimensional contacts are readily displayed as a “chromatin contact map” (or matrix) in which the intensity at a point  $(x, y)$  is proportional to the number of read-pairs linking two positions  $x$  and  $y$  on the genome. Hi-C was initially developed to elucidate the three-dimensional folding principles of the human genome, demonstrating that the genome is organized into alternating open “A” and closed “B” chromatin compartments [31]. Compartments and other signatures of organized folding, however, are superimposed on a background of contacts that arise from the polymeric nature of chromatin which leads to an excess of contacts between loci on the same chromosome even if they are distant along the linear sequence. Such local intra-chromosomal contacts appear as a diagonal band in the contact map with a characteristic pattern of decay with increasing separation of the genomic loci that extend out to megabases. These stereotyped patterns of contacts along chromosomes have been used to provide chromosome-scale linkage information for genome assembly [33]. Similarly, Hi-C can be leveraged to detect rearrangements that bring chromosomal segments that are distant on the reference genome together onto the same chromosome in a non-reference sample, since such rearrangements appear as “off-diagonal” signals in the Hi-C contact map. Several computational methods have been developed to analyze Hi-C data to infer copy number variation [34], [35], [36], [37], [38], [39]. Some of these methods developed foundational techniques for identifying and annotating interchromosomal translocations as well [40], [41], [42]. One of these methods, HiNT [42], outperforms the others and can provide translocation breakpoints at single base-pair resolution given sufficiently high Hi-C coverage, and is the current gold-standard for cancer Hi-C data analysis. However, these methods have been developed and tested on pure cancer samples or cancer cell lines while the analysis of clinical cancer samples will often require the analysis of samples comprising genetically heterogeneous cancer cells intermixed with normal cells. This requires inference of tumor purity in order to

estimate absolute copy numbers in cancer cells, which the existing methods, including HiNT, don't.

Here, we describe HiDENSEC (Hi-C based Direct Estimation of Copy Number and Structural rEarrangements in Cancer cells) (**Figure 2.1**). We use Fix-C, an adaptation of Hi-C optimized for formalin-fixed, paraffin-embedded (FFPE) tumor samples; henceforth the terms Fix-C and Hi-C will be used interchangeably [43]. Our computational framework allows us to (1) infer the fraction of cancer cells in mixtures of cancer and normal cells (also termed as tumor purity), (2) estimate absolute copy number across the genome in the cancer cells, and (3) detect and ascribe absolute copy number to large-scale structural variants using Hi-C. We validated our methods by analyzing samples mixed from different genotypes *in silico* and *in vitro*. With HiDENSEC analysis of Hi-C data, we track the emergence and evolution of structural variants during melanoma progression in three patients, demonstrating the utility of HiDENSEC to precisely characterize LSSVs and copy number changes in melanoma progression.

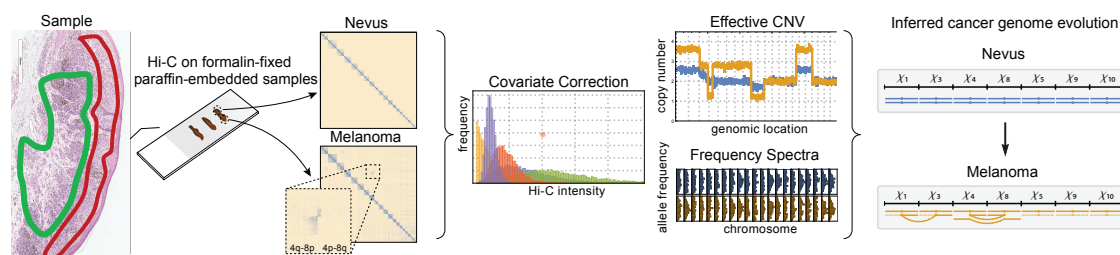


Figure 2.1: **HiDENSEC pipeline**

## 2.2 Results

### Chromatin contacts in cell mixtures are linear superpositions

The application of high-throughput chromatin conformation capture methods to FFPE samples (Fix-C) opens new possibilities for studying chromosome rearrangements in solid cancers [43]. To analyze samples that are mixtures of genetically distinct cells, we assume that the chromatin contact map derived from heterogeneous samples is the weighted superposition of the intracellular contact maps of individual cell genomes, without spurious signals caused by intercellular DNA ligation (Online Methods). We tested this assumption in a synthetic FFPE sample composed of a 1:1 mixture of human and mouse cells and fixing the cell pellet

with formalin, and subsequently embedding it in paraffin before processing it with the Fix-C protocol (Online Methods). We found that the number of Fix-C read pairs connecting human and mouse is less than 0.3%, which confirms that intercellular proximity ligation is negligible. We can therefore interpret Fix-C contact maps as superpositions of the contact maps of the cell populations within the sample.

## Estimating absolute copy number and tumor purity

We estimate the relative copy number along the genome from Hi-C data, specifically, the on-diagonal intensities of the chromatin contact matrix computed in 50 kb windows. Each on-diagonal entry measures the total contact frequency of a genomic window with itself, which is nominally expected to be proportional to the copy number of that genomic window. Raw on-diagonal intensities of Hi-C data derived from cancer cell lines, however, empirically show broad distributions that may not be easily translated to absolute copy numbers, which for each cell type must be simple integers. Since we are specifically interested in describing cancer samples comprising populations with distinct cancer genomes it is important to obtain quantitative measures of copy numbers. We reasoned that the on-diagonal intensities may also be influenced by factors such as the density of restriction enzyme cut sites, short-read mappability, sequencing bias due to GC content, and possible effects of variable chromatin compaction along the genome, such as A/B (open/closed) chromatin compartments [31], [44], [42]. We therefore assessed the impact of these factors on on-diagonal intensities in wild-type cells (without copy number variation), and we found that all of them contributed significantly to overall variation in raw on-diagonal intensity. Introducing covariate corrections for (1) chromatin compartments, (2) restriction enzyme site density, (3) short-read mappability and (4) GC content is able to explain around 80-90% of overall on-diagonal intensity variation in validation Hi-C matrices. We applied this covariate correction to improve copy number inference from Hi-C data.

Absolute copy number profiles facilitate the estimation of tumor purity in cancer samples that include mixtures of cancer and normal cells. Since (1) the contact map of a cell mixture is the weighted superposition of the contact maps of each subpopulation of cells (as shown above), and (2) the copy number of each region of the genome must be an integer, we can jointly infer the cancer cell fraction  $f$  and copy number profile of a mixed sample. While the remainder of the sample is generally assumed to be wild type cells with frequency  $1-f$ , in general, there could be multiple cancer cell populations, with tumor purities  $f_1$ ,  $f_2$  and so on. In order to infer absolute copy numbers and tumor purities from covariate corrected relative copy number profiles, some prior knowledge about absolute copy numbers is necessary. By default, HiDENSEC assumes knowledge of the most common copy number in a given sample (typically, wild type ploidy 2), though alternative specifications can be incorporated as optional inputs to HiDENSEC. We validated our absolute copy number and tumor purity estimation method using (1) *in silico* mixtures of Hi-C data from karyotypically normal GM12878 cells and HCC1187 cancer cells, and (2) Fix-C data derived from *in vitro* (FFPE) pelleted mixtures of karyotypically normal GM12878 cells and HCC1187 cancer cells. The

karyotypes of both cell lines are well-characterized, providing a reliable ground truth [45]. These Hi-C data from in silico and in vitro mixture samples confirmed the accuracy of the absolute copy number estimates of the cancer cell type, and the accuracy of tumor purity estimates (Figure 2.2 a-d).

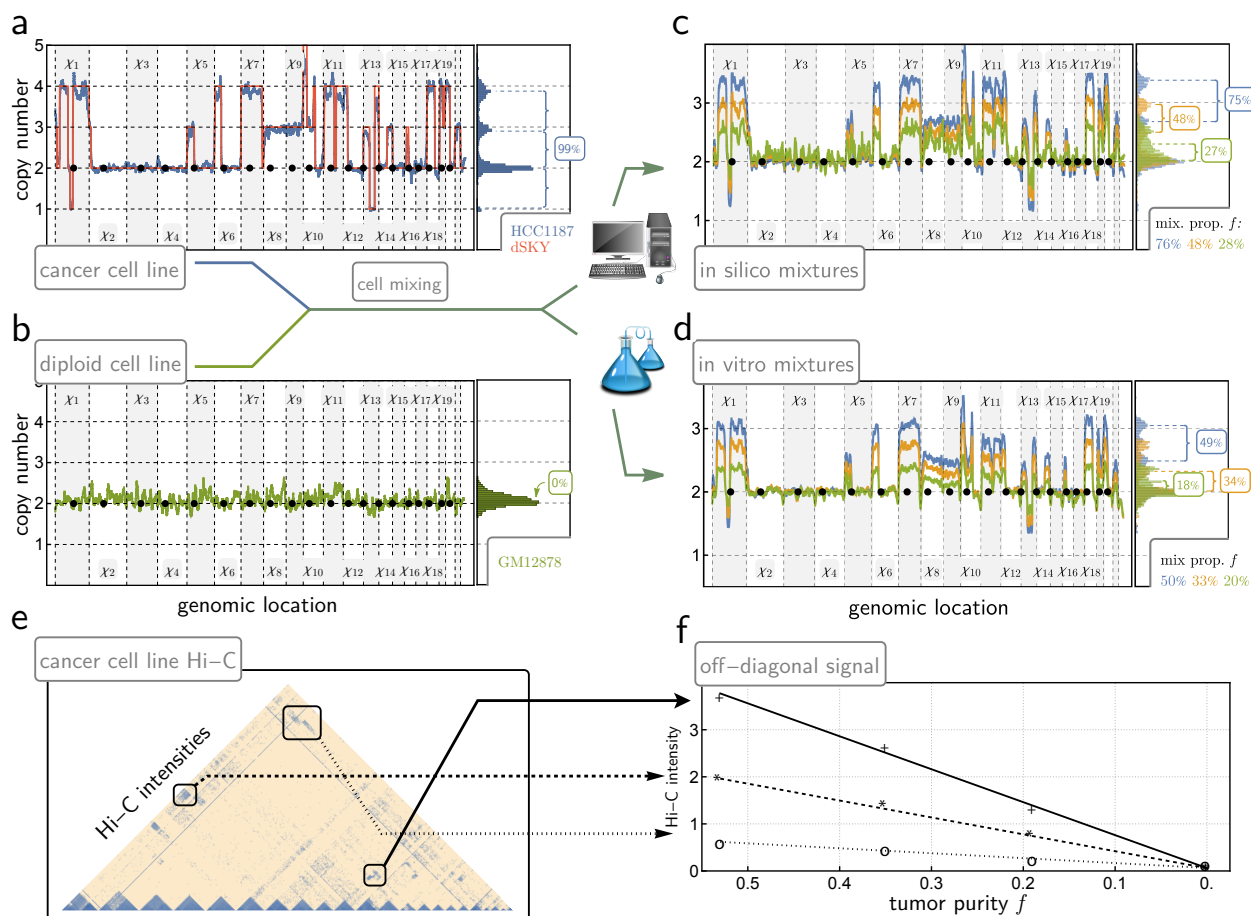


Figure 2.2: Validation of HiDENSEC using mixtures of samples

We initially considered also exploiting the precise magnitudes of off-diagonal contacts (arising from inter-chromosomal fusion events) to quantify tumor purity. Since such contacts are absent in normal cells, these signals should arise only from the cancer cells. As expected, we found that these signals are proportional to the tumor purity in our in silico Hi-C mixtures, following the superposition principle for cell mixtures (Figure 2.2e, f). The absolute read counts arising from such inter-chromosomal contacts, however, vary substantially between translocations. This is likely due to both differences in the chromatin packing of different inter-chromosomal contacts and variance in precise fusion locations within the binning windows that define the resolution of our analyses (by default, 50 kb windows). Given this lack of



meaningful quantitative information beyond presence and absence of off-diagonal contacts, HiDENSEC relies on them solely for detection of large structural variants, with inference of absolute copy numbers and tumor purity primarily based on the on-diagonal intensity analysis described above.

## Detecting reciprocal and copy-number-altering translocations

As part of HiDENSEC we developed automated methods for detecting inter-chromosomal (and long-range intra-chromosomal) rearrangements, which appear as “off-diagonal” features of the Hi-C matrix. These include (1) rearrangements associated with copy number changes of both of the involved chromosomes, whose breakpoints coincide with boundaries of copy number changes, and (2) reciprocal translocations, which show a characteristic “bow-tie” pattern in the Hi-C contact map. For copy-number-associated (type-1) events, we integrated the previously inferred copy number profile and its associated change points with a set of intuitive summary statistics (Online Methods) to identify the most likely translocations and breakpoints. For reciprocal translocations (type-2 events) we exploited the distinctive “bowtie” pattern of the Hi-C contact map in regions around the breakpoints to design summary statistics sensitive to such structured matrices. In both scenarios, each candidate rearrangement is associated with a well-calibrated p-value, allowing assessment of significance through standard multiple testing procedures. Finally, we also found some events that are neither type-1 nor type-2. We found these to be more difficult to detect in an automated fashion, since biological covariates such as compartment structure and chromosome size may produce apparent off-diagonal signals, and identified these by manual curation of off-diagonal Hi-C signals.

We validated HiDENSEC’s performance on the six in vitro mixtures described above (**Figure 2.3a, b**), as well as a manually annotated melanoma Hi-C sample (**Figure 2.3c**). Performance of HiDENSEC relative to HiNT was assessed systematically by recording top-k recall curves for k up to 60 and each validation mixture, with recall measured relative both to all LSSVs present, as well as only those belonging to classes (a) and (b) described above. As indicated in **Figure 2.3a**), HiDENSEC consistently performs favorably, returning fewer total off-diagonal calls at higher recall, without exception. The recall metric here is defined on the level of chromosome pairs; that is, an off-diagonal call is classified as positive if its participating chromosome pair shares a ground-truth rearrangement. To illustrate HiDENSEC’s performance in localizing such fusions, **Figure 2.3b**) shows an example event detected by HiNT at mixture proportions 50%, but not below.

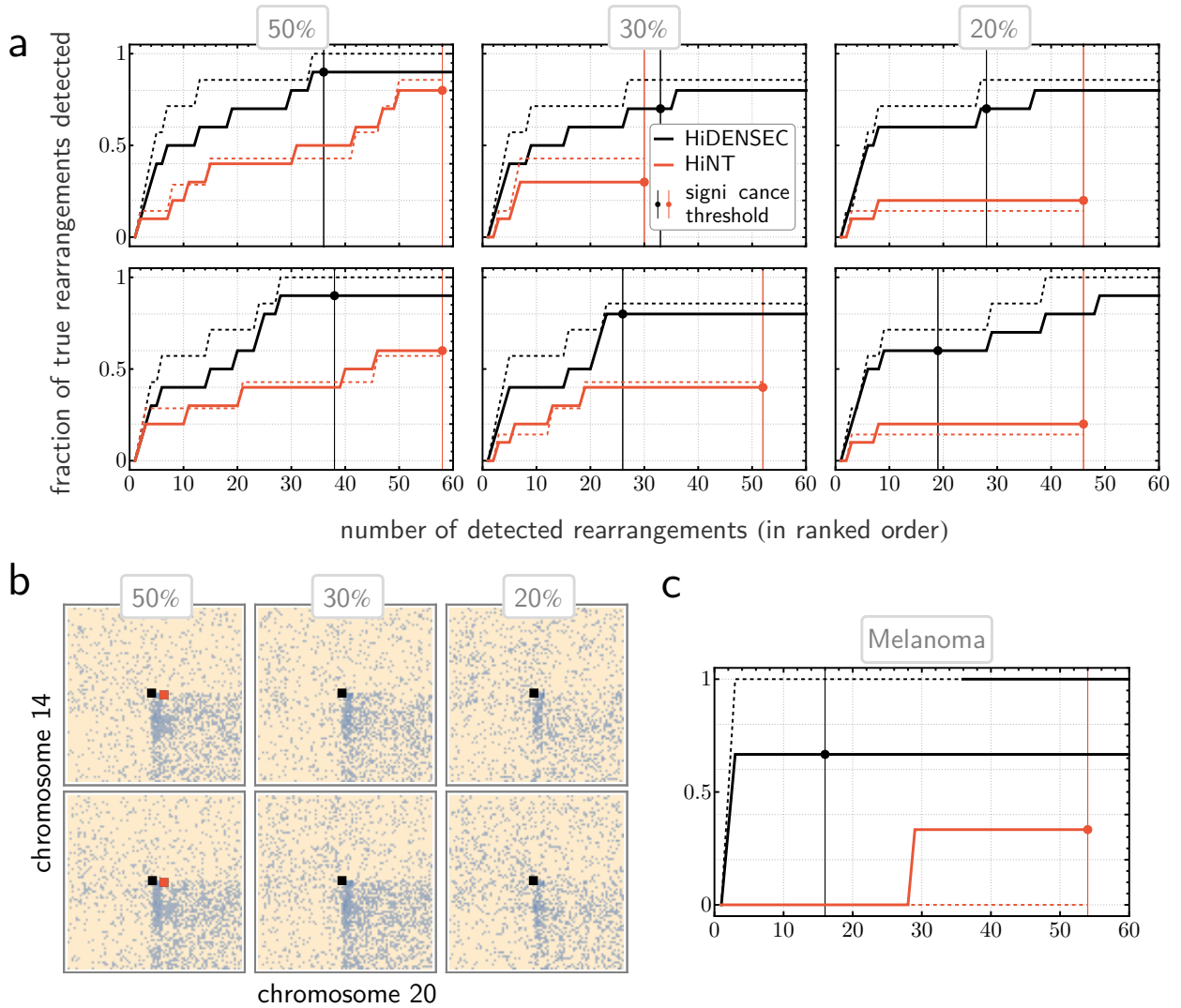


Figure 2.3: Benchmarking HiDENSEC's LSSV identification

## Using HiDENSEC to reveal the the evolution of chromosomal aberrations during melanoma progression

We characterized chromosome evolution during cancer progression using Hi-C data from three patients using HiDENSEC. In all, we generated Hi-C data from nine samples at various stages of melanoma development:

**Patient 1.** A primary cutaneous melanoma (Sample 1 - II) with an adjacent precursor nevus (Sample 1 - I) microdissected from the same FFPE sample.

**Patient 2.** A primary melanoma (Sample 2 - I) and its asynchronous metastasis (sample 2-II).

**Patient 3.** Two histologically distinct regions (1 and 2, Samples 3 - I and 3 - II) of a large primary cutaneous acral melanoma and its asynchronous metastasis (Sample 3 - III).

For each patient, we sequenced and analyzed Fix-C libraries prepared from FFPE sections that were adjacent sections used for either targeted short-read sequencing of a panel of cancer-associated genes called UCSF500 [46] or exome sequencing. These data allowed absolute copy number profiles inferred by HiDENSEC to be compared to phylogenetic relationships between progression stages derived from somatic mutations to develop a comprehensive view of the cancer genome.

## Patient 1

HiDENSEC analysis of Fix-C data from the microdissected nevus (Sample 1 - I) and adjacent melanoma (Sample 1 - II) revealed balanced translocations between chromosome 4 and 8 and chromosome 1q and 3q that were only present in the melanoma (**Figure 2.4a, b**). Chromosome breakpoints for these translocation events did not overlap genes that could be directly linked to melanoma progression. Chromosome arms 1p and 3p showed reduced copy number, with copy number transitions corresponding to the translocation breakpoints, indicating that the reciprocal derivative chromosome was lost in the melanoma. In addition, there were copy number losses of chromosomes 5, 9, and 10 (**Figure 2.4c**) estimated by HiDENSEC to represent monosomies, with a cancer cell fraction  $f = 57\%$ . This estimate is consistent with the allele frequencies determined from the UCSF500 cancer gene panel. Somatic variant calling using exome sequencing of the nevus (Sample 1 - I) and the adjacent melanoma (Sample 1 - II) along with a matched normal sample, identified the BRAF V600E mutation as a driver mutation present in both the nevus and melanoma (**Figure 2.4d**). Together, these combined analyses show that our method can detect chromosomal rearrangements and copy number changes in tumor samples with a considerable contribution of normal cells (**Figure 2.4e**).

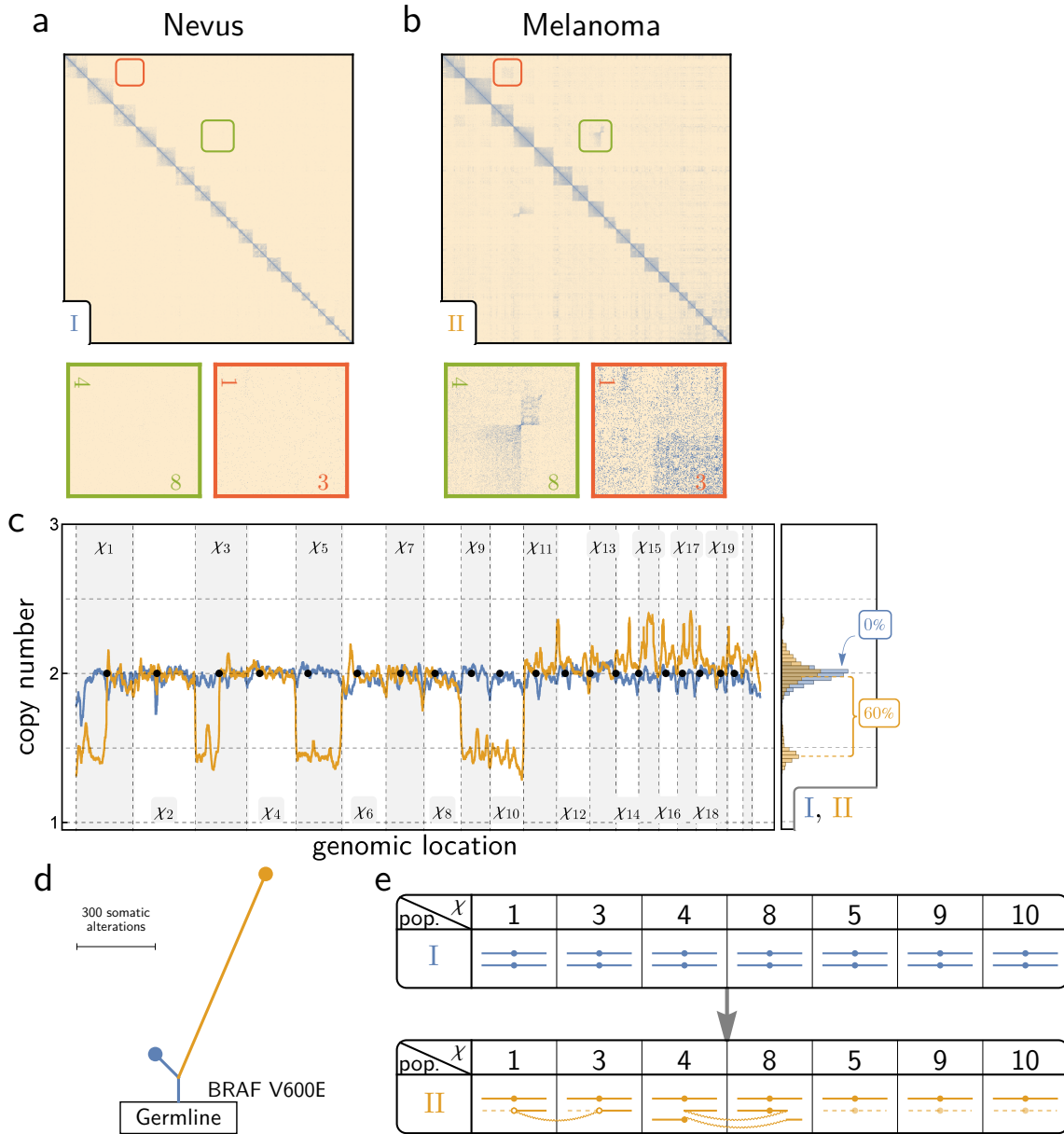


Figure 2.4: HiDENSEC analysis of Patient 1

## Patient 2

For patient 2 we compared a primary melanoma (Sample 2 - I) with a metastasis that was collected from the same patient at a later time (Sample 2 - II). While some translocations and copy number changes were shared by both samples, others were unique to the metastasis

(**Figure 2.5a, b, e**). The existence of shared structural variants between the two samples implies that the metastasis arose from a common ancestor with the primary melanoma (**Figure 2.5e**). Our HiDENSEC analyses are consistent with a model in which each sample has a single dominant (but karyotypically distinct) cancer cell population mixed with normal cells. In the melanoma, this dominant population comprised  $f = 71\%$  of the cells, while the metastasis comprised 59% cancer cells (**Figure 2.5c**). The tumor purity estimate for the primary melanoma sample is consistent with tumor purity estimated using the mutant allele frequency for BRAF V600E, the presumed initiating oncogene (**Figure 2.5d**). Copy number profiles estimated using HiDENSEC are highly concordant with profiles derived from the UCSF500 capture panel from an adjacent tissue section. Differences in tumor purity are expected due to variation across microdissected samples from the same tumor. As with Patient 1, HiDENSEC analysis applied to Fix-C data detected translocations whose genomic breakpoints would be undetectable by standard array CGH and also provides the details of the underlying rearrangements. We used off-diagonal chromatin contact map signals to annotate the associated breakpoints. For example, both the primary melanoma and its metastasis carry a complex translocation event involving chromosome 2, 5 and 10, which is concurrent with loss of 5q, a part of 2p and a part of 10p suggesting that the underlying structural rearrangement occurred early in the primary melanoma, since this is an LSSV shared by both the primary melanoma as well as the metastasis. Moreover, HiDENSEC detects a metastasis-specific translocation between chromosome 11 and 17 that explains the copy number loss of 11q and a gain in copy number of 17q in the metastasis. Similarly, we detected a chromosome translocation between chromosome 1 and 15 that is present in the melanoma and continued to evolve by fusion with chromosome 13. This analysis highlights that HiDENSEC can deconvolve chromosome scale events during cancer evolution (**Figure 2.5e**).

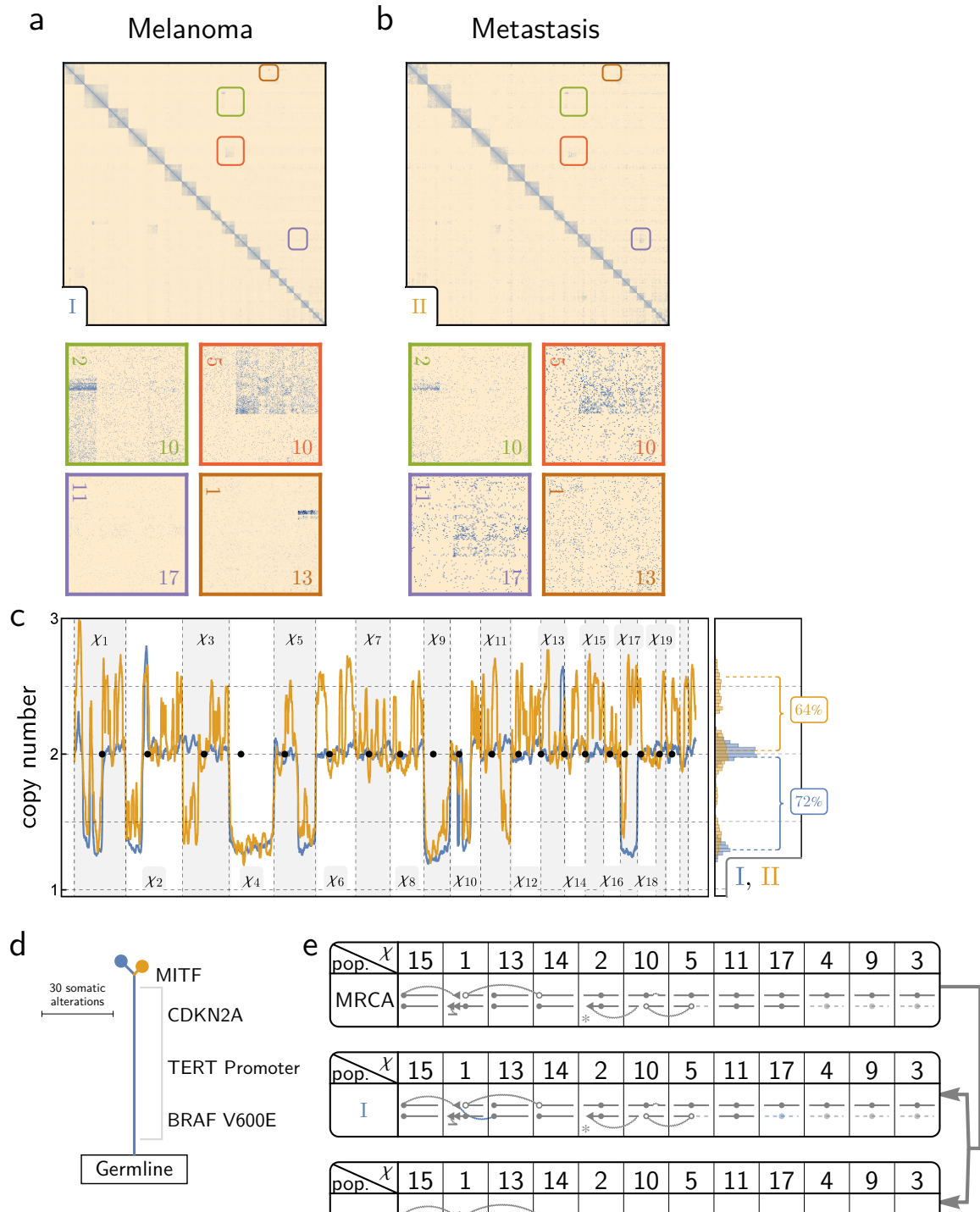


Figure 2.5: HiDENSEC analysis of Patient 2

### Patient 3

The acral melanoma of Patient 3 has been extensively characterized using conventional approaches [47]. Acral melanomas are known to be enriched for structural rearrangements [48]. We analyzed two histopathologically distinct subregions of the primary melanoma of Patient 3 (without prior knowledge of any genetic differences) and one from the metastasis, which arose years later (Figure 2.6). HiDENSEC analysis using Fix-C data generated from these three samples revealed genetic heterogeneity within the primary tumor and the progression of chromosomal alterations during evolution of the melanoma into metastasis. Since the Fix-C samples from Patient 3 were sequenced more deeply we were also able to characterize allele frequencies of inherited variants and trace haplotype copy number (Online Methods, Figure 2.6e). In conjunction with copy number estimates, analysis of these germline variants allowed us to identify (1) homozygous copy number neutral changes arising from the loss of one chromosome and the duplication of its homolog, and (2) infer in triploid cases which haplotype was duplicated (Figure 2.7a). The allele frequency spectra therefore provide independent corroboration of copy number and allow precise inference of lost and/or duplicated haplotypes.

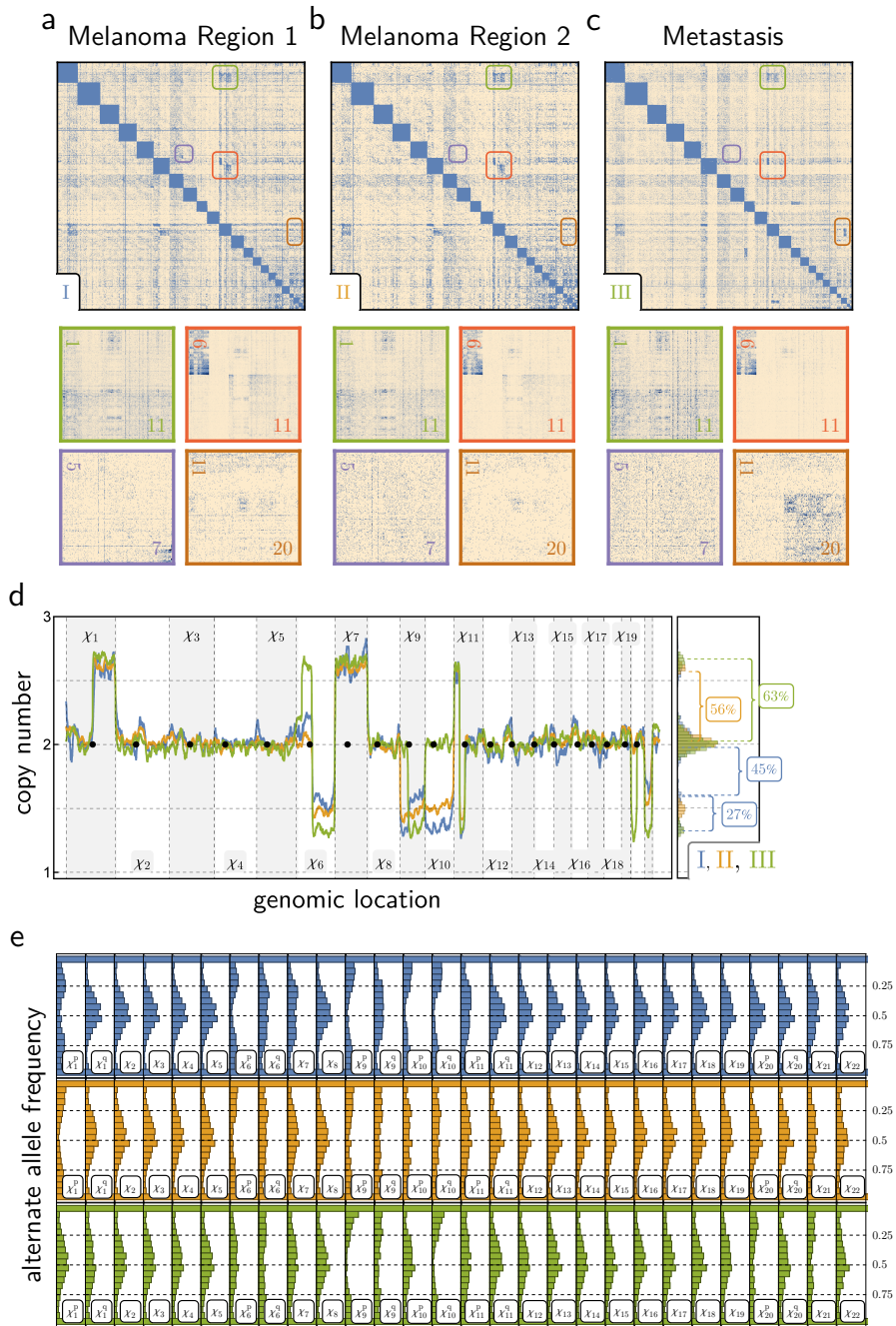


Figure 2.6: **HiDENSEC** analysis of Patient 3

The two regions of the primary melanoma share copy number changes and structural rearrangements (Sample 3 - I and Sample 3 - II), indicating that they are clonally related. HiDENSEC CNV profiles show that, in some samples, several genetically distinct populations of cancer cells are present. In such mixtures, the observed copy number profile is a super-



position of the CNV profiles of each genotype, weighted by cell fraction. Specifically, the absolute copy number increases of chromosomes 1q and 6p, and the copy number decreases of chromosome 6q, 9, 10, 11p, and 21, in Sample 3 - I and Sample 3 - II are inconsistent with a single cancer cell population (**Figure 2.6d**). However, we find that two distinct cancer cell populations can explain both samples I and II of Patient 3. Using HiDENSEC, we estimate that Sample 3 - II comprises  $f_A = 56\%$  cancer cells with genome A and 44% normal cells. While Sample 3 - II could not be described as a mixture of normal cells with a single cancer cell population; knowledge of cancer genome A from Sample 3 - I allowed us to interpret Sample 3 - II as a mixture of normal and cancer cell population with genome A cells with a second cancer cell population with genome B. We inferred that Sample 3 - II comprises  $f_A = 60\%$  cancer cell genotype A,  $f_B = 12\%$  cancer cell B genotype, and  $1-(f_A+f_B) = 28\%$  normal cells. Finally, the metastatic sample can be described as a mixture of normal cells and a third cancer cell population with genotype C, that is most closely related to cancer cell population A of the melanoma, with cancer cell fraction  $f_C = 63\%$ .

The karyotypes of cancer cell genotypes A, B, and C as inferred by HiDENSEC are shown schematically in **Figure 2.7a, b**. These genomes exhibit shared and/or unique copy number gain/loss and large-scale rearrangements to varying degrees that can be mapped onto a phylogeny of the three cancer genomes. Based on a parsimony analysis of multiple shared chromosome-scale features of cancer cells with genotypes A and C, we infer that they share a more recent common ancestor, with melanoma genome B diverging earlier, and parsimonious reconstructions of the AC and ABC ancestors are also shown in (**Figure 2.7b**). This phylogeny of the three cancer cell genomes provides a framework for understanding changes in karyotype through cancer progression.

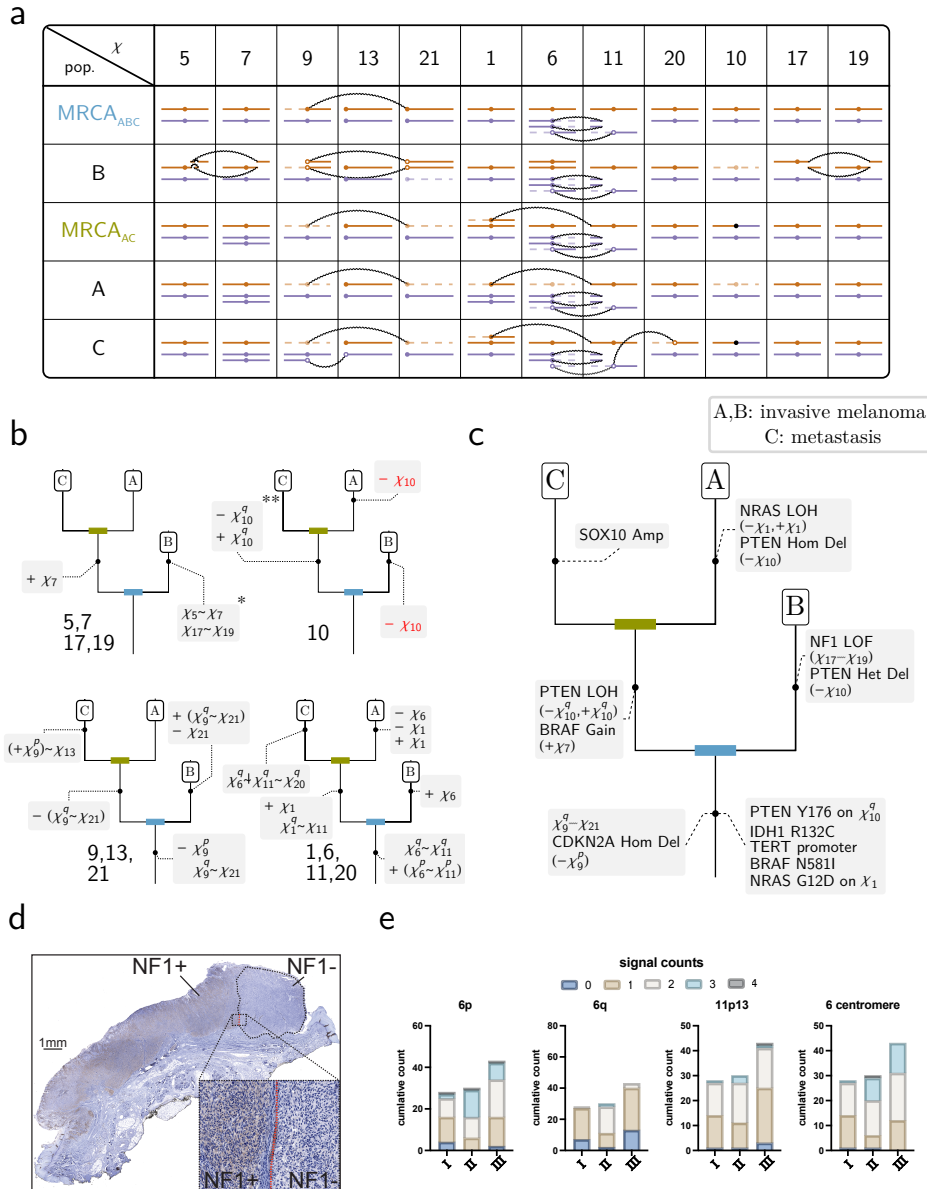


Figure 2.7: Evolution of the melanoma genome in Patient 3

Allele frequency spectra derived from the tumor Fix-C data (Online Methods) are consistent with copy number changes inferred by HIDENSEC and provide additional information about genomic change during cancer progression (Figure 2.6e). While most diploid chromosomes have the expected 1:1 ratio of reference:alternate alleles, in some cancer cells we infer a copy number neutral loss of homozygosity, indicated by the presence of two copies of the same haplotype (depicted as chromosomes with the same color in Figure 2.7a). Chromosomes

or chromosome arms predicted to be haploid or triploid by HiDENSEC show corresponding deviations from 1:1 allele ratios, with consequences for cancer progression.

For example, trisomic chromosomes such as chromosome 7 (**Figure 2.6e, f**) show allele frequency spectra consistent with a 1:2 allelic ratio in genotypes A and C (with observed signal diluted by the fraction of wild type cells in the sample; chromosome 7 is diploid in genotype B). The chromosome 7 haplotype that duplicated in the AC ancestor carries an oncogenic BRAF N581I mutation (**Figure 2.7a, c**). This mutation must have arisen early in cancer progression, because it is found in all three cancer cell genotypes, consistent with its appearance in targeted sequencing [47]. Our HiDENSEC analysis shows that the copy number increase in this mutation in the AC progenitor was associated with duplication of the entire mutant chromosome (**Figure 2.7a**). Our analysis also establishes the genetic background on which the mutation occurred, since alleles occurring at higher frequency along chromosome 7 must all lie on the mutant haplotype. Lastly, we find that genotype B carries a unique 5-7 translocation – heterozygously disrupting the LMBR1 gene in chromosome 7 – that is not found in the AC lineage (**Figure 2.7c**).

Chromosome 10, which encodes the tumor suppressor PTEN on its q arm, provides a more complex case. The PTEN Y176X mutation and loss of the wild type 10q arm was an early event found in both cell types A and B (**Figure 2.7c**) [47]. In contrast, chromosome 10 is diploid in genotype C (**Figure 2.7a**). Surprisingly, in genotype C the 10q arm carrying the PTEN Y176X mutation is homozygous but the 10p arm is heterozygous (i.e., wild type) (**Figure 2.7c**). We therefore infer that the two copies of chromosome 10 in genotype C did not arise by simple chromosome duplication in the C lineage (which would have produced a homozygous chromosome), but must have involved chromosome arm exchange either in an ABC ancestor (with parallel loss of the recombinant chromosome in A and B), or in an AC ancestor (with parallel loss of the original brown chromosome 10 in B and loss of the recombinant chromosome 10 in A). Regardless of the timing of the recombination event, we infer the presence of a previously unrecognized chromosome arm exchange during progression (**Figure 2.7a**).

The coordinated evolution of chromosomes 9 and 21 provides another example in which chromosomal rearrangements observed in Fix-C can be used to explain complex karyotypic changes. All three cancer cell genotypes have compound heterozygous deletions of the CDKN2A locus encoding the tumor suppressors p14 and p16. In genotypes A and C one copy of chromosome 9 is completely lost, while genotype B retained 9q. All three genotypes carry a small deletion of CDKN2A on the other (blue) copy of 9p (**Figure 2.7a, c**). Genotypes A and C are missing one copy of chromosome 21 while genome B contains a duplicated homozygous 9q-21 fusion chromosome (making B triploid for 9q overall) but is missing the alternate copy of chromosome 21 relative to A and C (**Figure 2.7a**).

The most parsimonious explanation of these changes is that the 9q-21 translocation occurred in the ABC ancestor, with concomitant loss of a copy of 9p (and a copy of CDKN2A) early in cancer evolution, likely because the 9p fragment liberated by the 9q-21 translocation lacked a centromere (**Figure 2.7c**). The 9q-21 fusion chromosome doubled in the B lineage with loss of the other copy of 21 (**Figure 2.7a**). Conversely, in the AC lineage the 9q-21

fusion was lost, explaining mechanistically (1) the coordinated loss of the same 9q and 21 haplotypes in both A and C and (2) the loss of 9p in A, B, and C. Finally, in the metastatic lineage C a duplicated copy of the remaining intact chromosome 9 experienced a 9p-13 translocation (with loss of 9q) (**Figure 2.7b, c**).

Chromosomes 1, 6, and 11 are the nexus of a complex series of copy number changes and rearrangements in the three cancer cell genotypes, notably involving an early NRAS G12D mutation on 1p (**Figure 2.7c**). Strikingly, as found for chromosomes 9-21-13, the observed sub-chromosomal copy number changes can be explained by translocations followed by consecutive mis-segregation of the resulting chromosome fusions, rather than by direct deletion of an arm (or part of an arm) (**Figure 2.6b**). An initial reciprocal translocation between chromosomes 6 and 11 occurred in a cell ancestral to cancer cell resulting in a 6p-11p fusion chromosome, whose copy number subsequently doubled and a 6q-11q fusion chromosome with a 20 Mbase deletion of the 11p chromosome at the fusion point (**Figure 2.5b**). Subsequently, cancer genome B gained one copy while cancer genome A lost one copy of chromosome 6. We infer that chromosome 6 loss occurred after the establishment of the metastasis genome, since the metastasis – which is also haploid for 6q – retains the alternate homolog of 6q. In the metastasis genome 6q was likely lost when the 6q-11q fusion chromosome passed the 11q chromosome arm to chromosome 20q by reciprocal translocation, followed by the loss of 6q and 20p. Imbalances between the p and q arms of chromosome 6 are very common in melanoma. Gain of 6p and loss of 6q occurs 50% of cases and can be used to diagnostically distinguish melanoma from nevi [16] using fluorescent in situ hybridization. We used this approach as an independent validation for the CNVs changes detected by HiDENSEC (**Figure 2.7e**).

Cancer genotypes A and C both share the loss of 1p and the fusion of the centromere of 1q with the p-telomere of chromosome 11. This coupled loss/fusion must have occurred before the most recent common ancestor of A and C (AC), and can be explained by a single event (as above, the loss of the 1p arm is presumably due to its lack of a centromere, which remained with 1q). C also retains an intact copy of the same haplotype of chromosome 1, but this is lost in A, which carries a duplicated copy of the homologous chromosome 1. The presence of an intact chromosome 1 in C implies its presence in the AC ancestor and both chromosome 1 homologous must have to have duplicated independently throughout the cancer evolution (**Figure 2.7a, b, c**).

Finally, cancer genotype B harbors lineage-specific balanced reciprocal translocations of (1) the distal portions of 17q and 19q, and (2) the 5q and 7q arms, resulting in a (24 Mb) deletion in the proximal 5q arm (**Figure 2.7a**). These changes are not present in cancer genomes A and C. Notably, the breakpoint of the translocation on chromosome 17p occurs within the NF1 locus, producing a loss of function mutation in the neurofibromin 1 tumor suppressor gene. While we did not detect a mutation on the second allele of NF1, immuno-staining for NF1 in an FFPE sample adjacent to Sample 3 - I showed a discrete NF1-negative region suggesting loss of NF1 function in this region (**Figure 2.7d**). Thus despite cells of genotype B not having progressed to a (detected) metastasis, it appears to have continued to evolve under selective pressure to eliminate the NF1 tumor suppressor and

thereby further upregulating MAP-kinase signaling.

## 2.3 Discussion

### A new method

Here we present, HiDENSEC, a new analytical method for investigating cancer genome evolution in patient samples. We determined the chromatin contacts in formalin fixed, paraffin embedded (FFPE) samples from three melanoma patients, using the Fix-C chromatin conformation capture protocol [43]. We show that these data can be used to identify both copy-number altering and copy-number neutral, and track the evolution of chromosome-arm-scale changes in cancer genomes. Patient samples are mixtures of normal and cancer cells and we show using *in silico* and *in vitro* generated controls that observed Fix-C signals are linear superpositions of the signals from normal and cancer cells weighted by cell frequency. This observation allows us to jointly estimate tumor purity as well as genome-wide absolute copy numbers. Hi-C data also provides information about allele frequencies, which can be used to identify copy-number neutral losses of heterozygosity or to identify the more common haplotype in triploid situations.

Since chromatin contacts probed by Hi-C extend over hundreds to thousands of kilobases, the method allows us to capture large-scale (>10Mb) rearrangements. Other short-read sequencing approaches rely on mapping read pairs across rearrangement breakpoints, and have lower sensitivity due to the difficulty in accurately mapping reads to repetitive sequences that often flank breakpoints. Benchmarking experiments for the detection of LSSVs in Hi-C data sets showed that HiDENSEC was more sensitive and accurate in identifying LSSVs than the current gold standard. Since our method can be used with formalin-fixed sections or micro-dissected fixed tissue, it can be applied retrospectively to samples collected over the course of a patient's disease progression. Finally, we show that the sensitivity of our method allows us to detect rearrangements that occur in melanoma development and to define the genetic changes that occurred specifically in minor subpopulations of melanoma cells.

### Biology of cancer progression

Although cancer sequencing approaches based on capture panels (e.g., UCSF500) can reliably estimate copy number profiles across the genome, and even estimate tumor purity, they cannot explain karyotype change in detail, because such methods have low-sensitivity for discovering rearrangements. Even deeper short-read sequencing is not able to reveal breakpoints in repetitive sequences. Thus, the centromeric breakpoints or telomere fusions repeatedly found in our melanoma analyses would be impossible to detect by conventional sequencing. Specifically, in the three patients' samples analyzed, HiDENSEC was able to annotate a total of two reciprocal translocations, three fusions involving telomeres (one of which is present in all three samples of Patient - 3), and five chromosome arm exchanges with breakpoint

in centromeric regions. Our approach therefore provides an integrated picture of cancer heterogeneity and karyotype evolution. A common process in our melanoma progression cases is whole chromosome arm rearrangement followed by loss and/or copy number change by mis-segregation of recombinant chromosomes. The most complex changes were found in the primary acral melanoma and metastasis from patient 3. For example, a chromosome 6 to 11 translocation was followed by a subsequent translocation, so that the 11q of the derivative chromosome becomes fused to 20q with the concurrent loss of 20p.

By sampling several stages of cancer progression, or even different regions within the same melanoma, we show that HiDENSEC can be used to infer the genome organization of multiple subpopulations of cancer cells. Comparing these subpopulations and applying the principle of maximum parsimony along with the known temporal relationships among the samples, we can infer unsampled intermediates and possibly transient states in cancer progression (**Figure 2.7a**). Thus in patient 3, we show that of the three subpopulations detected, melanoma genotype A is more closely related to the metastatic genotype C subpopulation, and that melanoma genotype B diverged prior to the A-C divergence. This, in turn, allows us to characterize the changes that occurred on this cellular phylogeny. We find that the most recent common ancestor of genotype A and genotype C is linked to large structural events that resulted in a gene conversion event of large parts of the q arm of chromosome 10 (**Figure 2.7b**).

Our analysis of two regions of the melanoma from patient 3 highlights the karyotypic heterogeneity in this tumor. In Sample 3 - I, an area of primary melanoma, two genetic subclones were identified. Previous analysis of this patient identified several consecutive mutations that lead to upregulated MAPK signaling by different mutations including, including copy number gain of BRAF in the AC precursor and loss of heterozygosity for the NRASG12D mutation in melanoma genotype A (**Figure 2.7c**) [47]. The identification of structural rearrangement that disrupts the NF1 locus in subpopulation B illustrates that HiDENSEC can uncover novel LSSVs that drive cancer cell evolution, even when only present in small cancer cell populations. While NRAS mutant cancers typically do not have NF1 or BRAF mutations, the G12D mutation likely still has some residual GTPase activity, explaining why NF1 loss and BRAF mutation provides a selective advantage for this branch of the melanoma evolution in patient 3.

Together our analyses of samples from three cancer patients demonstrates that HiDENSEC analysis of Hi-C data can characterize cancer cell genome evolution from the earliest stages of cancer development using microdissected tissue. Notably, this approach allows us to deconvolve heterogeneous mixtures of cancer cells with distinct genotypes, and follow the genomic changes through time by analyzing samples through cancer progressions. Applying this approach at a larger scale to investigate will significantly enhance our understanding of cancer cell genome evolution by revealing common patterns of chromosomal change that can be used both for diagnostic purposes and to further decipher the underlying causal genetic changes during cancer progression.

## 2.4 Online Methods

### Source and characterization of melanoma samples

Formalin-fixed paraffin-embedded tissues were retrieved from the archives of the Dermatopathology Section of the Departments of Dermatology and Pathology. Tumor bearing areas were microdissected from 10  $\mu\text{m}$  thick unstained sections, using HE-stained sections as guidance. Archival formalin-fixed, paraffin-embedded (FFPE) melanoma samples were retrieved from the archives of the UCSF Dermatopathology service, under an IRB approved protocol (11-07951). Routinely stained sections were evaluated and tumor areas were marked by a dermatopathologist. FISH was performed with locus-specific probes for chromosomes 6p (RREB1), 6q (MYB), 11q13 (CCND1, and 6 centromere as previously described [49](#))

### Fix-C methodology and sequencing

FFPE samples were processed using Fix-C<sup>®</sup> kits from Dovetail Genomics. The sample preparation and Fix-C protocol have been described in Troll et al. 2019. Briefly, paraffin embedded tissue was dissolved in xylene followed by centrifugation. The tissue sample was hydrated with a series of ethanol washes (100%, 70%, 20%) and water followed by centrifugation. The tissue sample was digested with proteinase K at 37C for 1h. The digested sample was centrifuged and the supernatant was saved to capture chromatin on beads. The chromatin was digested with a restriction enzyme at 37C for 1h followed by wash. The digested ends were repaired and subjected to proximity ligation at 16C for 1h. Post ligation sample was crosslink reversed and DNA was purified on AMPure XP beads. The purified DNA was sheared and end-repaired for Illumina adaptor ligation. The proximity-ligated DNA is enriched with capture on streptavidin beads. The captured DNA is then PCR amplified on beads for 13 cycles, purified using AMPure XP beads, quantified, and sequenced.

### Cell culture and formalin fixation of cell line mixtures embedded into paraffin blocks

Suspension human normal cells (GM12878) were cultured in RPMI-1640 medium [ATCC] supplemented with 15% FB Essence [Seradigm] and 100 U/mL Penicillin-Streptomycin [Gibco]. Mouse embryo fibroblasts (MEFs) were cultured in DMEM [Gibco] supplemented with 15% FB Essence and 100 U/mL Penicillin-Streptomycin. The human and mouse cell lines were dissociated by Trypsin-EDTA (0.25%) [Gibco] for single cell suspension, then quantified by Trypan blue staining with a Countess cell counter [Invitrogen]. 200  $\mu\text{L}$  of 2% agarose in PBS solution was pipetted into a 1.7 mL microfuge tube and allowed to solidify. Equal numbers human and mouse cells (15 million total) were mixed and pelleted in a 15 mL conical tube, then resuspended in a small volume of neutral-buffered 10% formalin, then finally re-pelleted in the microfuge tube with agarose plug. Supernatant was then aspirated and fresh neutral-buffered 10% formalin was gently pipetted onto the cell pellet.

The microcentrifuge tube was then placed in buffered formalin at room temperature for 24 hours. The bottom of the microcentrifuge tube was then cut off with a razor blade and the plug gently extruded into a tissue cassette immersed in PBS using a pipette tip. The tissue cassette with mixed cell line plug was then embedded in paraffin and sectioned using standard protocols [50].

## Creation of in vitro normal-cancer mixtures

Adherent human cancer cells (HCC1187) were cultured in RPMI-1640 medium supplemented with 10% FB Essence and 100 U/mL Penicillin-Streptomycin. Human wildtype (GM12878) and HCC1187 cancer cells were dissociated by Trypsin-EDTA and mixed in ratios of 1:1, 2:1, and 4:1 WT:cancer cells before pelleting, fixation, paraffin-embedding, and sectioning as described above.

## Allele Frequency Spectrum

To infer haplotype copy number, we computed the regional frequency of nominally inherited variants in 500 kb windows. A single copy of each haplotype would then have a peak at 50% frequency; two copies of one haplotype and one copy of the other would appear as peaks at 33% and 67%; loss of heterozygosity would appear as a peak at 0/100%. Since we did not have matched patient normal samples, we considered variants as nominally inherited if they occurred with alternate allele frequency between 40% and 60% in the 1000 Genomes Project. Note that this allele frequency spectrum (as shown, e.g., in **Figure 2.6e**) does not include somatic mutations, which are considered separately.

In order to arrive at the copy number profiles, mixture proportions and off-diagonal events, HiDENSEC proceeds in broadly three steps: (i) Covariate correction, (ii) joint inference of absolute copy number profile and mixture proportion, (iii) detection of large-scale chromosomal rearrangement events. (ii) and (iii) occur partially in tandem in order to facilitate sharing of information that may improve statistical inference. Despite the large number of cells contributing to any single Hi-C experiment, read counts in general do not tend to follow parametric distributions typically associated with increasing sample sizes, and so HiDENSEC remains fully nonparametric throughout all these steps.

## Running HiNT and hicbreakfinder

HiNT was run with version: 2.2.7. hicbreakfinder was built from the master branch source downloaded from github, <https://github.com/dixonlab/hicbreakfinder>, using commit 30a0dcc6d01859797d7c263df7335fd2f52df7b8 (last updated in 2018). For hicbreakfinder the inter and intra chromosomal break files were provided by the Dixon lab as detailed on the github site. NextFlow pipelines to run both HiNT and hicbreakfinder can be found on the HiDENSEC github.



## HiDENSEC pipeline

Hi-C paired-end (PE150) sequencing reads were aligned to the hg38 reference genome using bwa [51] and then converted to Hi-C maps using Juicer [52] and visualized in Juicebox [53]. These were then processed with the HiDENSEC pipeline which comprises custom pre-processing scripts as well as a Mathematica Notebook reproducing all presented results. The HiDENSEC code is available at <https://github.com/sanjitsbatra/HiDENSEC>. HiDENSEC aims to interpret the Hi-C contact map of a cancer sample as a mixture of cells with distinct genomic types. Each genome has a discrete set of copy number changes and rearrangements relative to the diploid genome, and occurs in a fraction of the cells in the sample. Both the copy numbers, rearrangements, and cell fractions will be inferred from the Hi-C dataset, typically including one normal or wild-type genome and one or two aberrant cancer genomes. In order to arrive at the copy number profiles, rearrangements, and the tumor purity of each genome, HiDENSEC proceeds in broadly three steps: (i) Covariate correction, (ii) joint inference of absolute copy number profile and tumor purity, (iii) detection of large-scale structural variants. (ii) and (iii) occur partially in tandem in order to facilitate sharing of information that may improve statistical inference. Despite the large number of cells contributing to any single Hi-C experiment, read counts in general do not tend to follow parametric distributions typically associated with increasing sample sizes, and so HiDENSEC remains fully non-parametric throughout all these steps. The HiDENSEC pipeline is described in detail in [A](#)

## Covariate Correction

We first account for a variety of biological and experimental factors that are known to affect relative Hi-C read counts. This correction is required for unbiased and stable inference. The most prominent factors affecting Hi-C contact maps include GC (guanine plus cytosine) content, read mappability, cut-site density, and compartment structure. The first three of these covariates were also modeled by HiNT. We also considered compartment structure for the diploid GM12878 cell line. Concretely, covariate correction models observed read counts falling into a bin of length  $w$  around a site  $i$  as

$$\text{reads}_i \propto (\text{absolute copy number})_i \cdot \text{correction}(\text{GC}_i, \text{mappability}_i, \text{cut-sites}_i, \text{compartment}_i), \quad (2.1)$$

where the corrector function is a simple compartment-specific linear model:

$$\text{correction}(\text{GC}_i, \text{mappability}_i, \text{cut-sites}_i, \text{compartment}_i) = \sum_{c \in \text{compartments}} \mathbb{1}_{c=\text{compartment}_i} \times (\beta_{c,1} \cdot \text{GC}_i + \beta_{c,2} \cdot \text{mappability}_i + \beta_{c,3} \cdot \text{cut-sites}_i) \quad (2.2)$$

The linear models are chosen to match observed trends in diploid reference genomes, and reliably account for  $\approx 80\%$  of their variability. The coefficients  $\beta_c$  generally depend on the precise experimental details (e.g., whether Hi-C or Fix-C protocols were used), and we

recommended estimating the covariate corrections using reference maps obtained through the same experimental protocol as the map of interest. In the absence of a reference map, HiDENSEC defaults to performing the correction (2.1)-(2.2) within the map of interest, restricting attention to those genomic sites that are likely to be diploid (see section below). For the samples presented in the main section, we used protocol-matching reference Hi-C maps. We note that the correction procedure in the form given by (2.1)-(2.2) only applies to the diagonal entries of the binned Hi-C matrix, which contains almost all information about copy number profiles. The off-diagonal components, which represent HiC contacts between (binned) site  $i$  and  $j$  are primarily used for detecting fusions and other large-scale structural variations (LSSVs). Since we are primarily interested in the presence or absence of inter-chromosomal fusions, their precise magnitude beyond a broad distinction of large and small is substantially less informative. Thus, even though it would be straightforward to extend (2.1)-(2.2) to correct off-diagonal read counts by regressing against paired covariates, we do not attempt this in HiDENSEC .

## Inference of copy numbers & mixture proportions

Our goal is to estimate both the absolute copy number profiles (i.e., local integer ploidy) and mixture proportion for each of the constituent genomic types. This is, however, an ill-defined problem without additional constraints. First, the HiC contact map cannot distinguish between uniformly diploid and uniformly triploid genomes (although this can be done by measuring allele frequencies which will differ in these two cases). Second, we cannot distinguish between a 50-50 mixture of a wild-type genome with a cancer genome bearing a triploid chromosome 1, vs, a 75-25 mixture of a wild-type with a cancer genome bearing a tetraploid chromosome 1. Both of these kinds of ambiguities arise even in the absence of noise and make the problem under-determined without additional assumptions.

To remove these ambiguities HiDENSEC makes two corresponding assumptions:

1. *The most common absolute copy number (of the mixture) is known.* Knowledge of this copy number mode allows for appropriate rescaling of the Hi-C matrix correcting for the overall unknown constant  $C_0$ . While other statistics of the absolute copy number profile may be used (e.g., mean, median), the mode is particularly appealing since it most often will equal 2, and as it is particularly reliable for estimating  $C_0$ .
2. *Absolute copy number are as close to diploid as consistent with the data.* That is, HiDENSEC returns the biologically most parsimonious estimate.

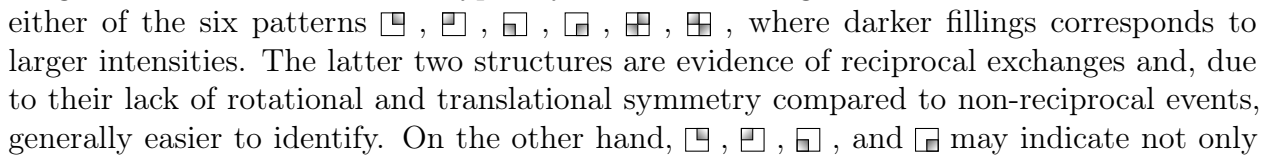
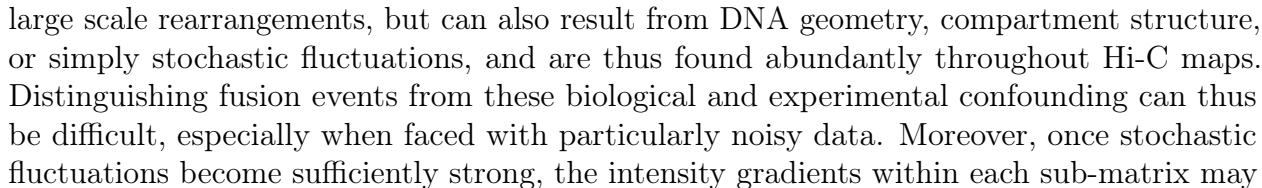
Given these assumptions, HiDENSEC appropriately centers the (covariate corrected) read counts by their largest mode, and infers absolute copy number profiles and mixture proportions jointly. For each cell population at a time, HiDENSEC scans along the genome in overlapping windows of length  $w$ , and identifies for each such window and a fixed choice of mixture proportion the copy number value that minimizes a suitably designed metric between predicted copy number and observed Hi-C intensities. A corresponding global discrepancy

metric is then minimized over all choices of mixture proportions, yielding both an overall mixture proportion estimate as well as local absolute copy number inferences. This estimated copy number profile is then subtracted from the read count data, and the entire procedure repeated in order to detect any potential further sub-populations contributing to the Hi-C matrix.

It can be proven that the inference scheme described above recovers the correct mixture proportions and absolute copy number profiles in the limit of noiseless data and comparatively few distinct cell populations, or in the case cell populations, whose mixture proportions and copy number profiles satisfy certain monotonicity properties. The latter constraint is not surprising, since the general inference task tackled by HiDENSEC is NP-complete in the number of cell populations, while the algorithm described above scales linearly in them. In order to relax the former constraint and accommodate noisy data, HiDENSEC performs a number of additional refinement, model selection and hypothesis testing steps that correct for any copy number changes that may be called purely as a result of random fluctuations or whose precise location may be shifted as a result thereof. In the process, each change point is assigned an interpretable confidence score that indicates to what extent it is likely to reflect actual biological signal, as opposed to being the outcome of noise. After undergoing another round of refinement using detected off-diagonal events (see section below), these estimates are then returned to the user for interpretation.

Likewise, HiDENSEC refines the initial mixture proportion estimate based on similar principles, and additionally equips them with 95% confidence intervals that reflect their associated uncertainty. There are two primary sources that contribute to this estimation uncertainty: Stochastic fluctuations in read counts and uncaptured biological or experimental covariates. While the former is typically well-addressed by classical non-parametric tools like the bootstrap, the latter is more delicate and possibly instance-specific, prompting HiDENSEC to employ bootstrap ideas combined with structured sub-sampling that integrate information within and across individual stretches of copy-number changes. The resulting confidence intervals are conservative, yet not overly so.

## Inference of large-scale structural variants

Large-scale structural variants typically result in off-diagonal sub-matrices structured in either of the six patterns , where darker fillings corresponds to larger intensities. The latter two structures are evidence of reciprocal exchanges and, due to their lack of rotational and translational symmetry compared to non-reciprocal events, generally easier to identify. On the other hand,  may indicate not only large scale rearrangements, but can also result from DNA geometry, compartment structure, or simply stochastic fluctuations, and are thus found abundantly throughout Hi-C maps. Distinguishing fusion events from these biological and experimental confounding can thus be difficult, especially when faced with particularly noisy data. Moreover, once stochastic fluctuations become sufficiently strong, the intensity gradients within each sub-matrix may

wash out, effectively rendering all of them rotationally and translationally equivalent. To address these sources of uncertainty, HiDENSEC resorts to two corrections:

1. *HiDENSEC only aims to detect non-reciprocal fusion events of class (a) as described in the main text.* Due to their effect on local copy numbers,  $\square$ ,  $\square$ ,  $\square$ ,  $\square$  in class (a) allow HiDENSEC to rely on its previously inferred copy number profile to aid in their detection. More concretely, by default HiDENSEC will only consider off-diagonal sub-matrices anchored at coordinates associated with copy number changes deemed significant by the previously outlined analysis. Switching to non-default behavior and scanning points along the whole genome is possible, but care should be taken in interpretation, as confounding by above-mentioned biological and experimental covariates may be present. Additionally, restricting HiDENSEC's search to copy number change points drastically reduces its run-time, with a typical analysis completed in less than twenty minutes on a typical laptop.
2. *Experimental and biological confounders tend to affect rows and columns more globally.* Biological confounders like compartment structure generally elevate read counts of interactions between the region of interest and all other sites in the genome, leading to entire rows and columns in the Hi-C matrix that are enriched. All summary statistics computed by HiDENSEC are thus calibrated by comparing their value at the site-pair of interest against their empirical distribution across the associated row and column.

With these two corrections in hand, HiDENSEC considers two summary statistics that measure the extent to which (a) intensities tend to accumulate in only one of the four quadrants of each sub-matrix, and (b) large- and small-intensity regions are separated by clear boundaries or edges. Under suitable null hypotheses on the Hi-C read count distribution. Since these two summary statistics are normalized against their row- and column-histograms, the corresponding  $p$ -values are readily combined, yielding properly controlled aggregate  $p$ -values based on which HiDENSEC calls significance.

Since sub-matrix patterns  $\boxplus$  and  $\boxminus$  are typically not tied to changes in copy number profiles, detecting potential candidates requires a more global search. Because such potential candidates are generally distinguished by dense patches of large intensities in the contact map, HiDENSEC enumerates the largest connected components of a suitably obtained graph that respects the geometric structure of the Hi-C matrix, and inspects its point of largest intensity, or focal point. Once these candidates are determined, a number of summary statistics aimed at capturing (a) concentration and sharpness properties as with  $\square$ ,  $\square$ ,  $\square$ ,  $\square$ , (b) enrichment near a central focal point, and (c) the presence of a gradual intensity decrease away from the focal point, are computed, and their calibration under suitable null hypotheses again verified

## Chapter 3

# Predicting gene expression from histone modifications using deep learning

This is joint work with Alan Cabrera and Professor Isaac Hilton in the Hilton Lab at Rice University, along with Jeffrey Spence and my advisor, Professor Yun S. Song and the manuscript is currently under preparation.

### 3.1 Introduction

All cells within a multicellular organism have the same genetic sequence up to a minuscule number of somatic mutations. Yet, a menagerie of cell types exist with diverse morphologies and functions. Epigenetics is an important regulator and driver of these differences. The field of Epigenetics was borne from the phenomena of differential yet specific phenotypic expression within a single genotypic system, (eg., differential phenotypes between Myocytes, Cardiomyocytes, T cells among many others). Waddington's "landscape" describes how regulation above the genetic level could explain cell lineages and differential phenotypes within a single genetically identical system [54]. Epigenetic modifications such as post-transcriptional modifications of core histones are involved in a variety of essential regulatory processes in the cell, including transcription control [55], [56], [57]. Consequently, the Histone Code Hypothesis suggests that combinations of different histone modifications specify distinct chromatin states thereby regulating gene expression [58]. Advancement in sequencing technology has allowed us to quantify gene expression and also profile different histone modifications in and around genes. Two large consortia have either performed an extensive number of assays in a small number of cell types (ENCODE [59]) or a small number of assays across many cell types (NIH Roadmap Epigenomics consortium [59, 60]). These include measurements of histone modifications, transcription factor binding, and chromatin accessibility, and are measured in a select set of cellular contexts. This data has enhanced our understanding of transcriptional regulation within these samples and has served to explore general questions in chromatin biology [61], [62], [63].

Studying the function of these epigenetic marks, however, has been largely limited to statistical associations with gene expression [64], [65], [66]. Technologies for targeted direct manipulation of these epigenetic properties are necessary to transform such association-based findings into mechanistic principles of gene regulation. Epigenetic editing allows us to probe the mechanism by which epigenetics affects expression, and promises to be useful for therapeutics by offering more precise targeting of expression levels. Such advances have the potential to benefit human health, as they could lead to gene therapies that modify the epigenetic code at targeted regions of the genome [67]. However, small molecule-based methods globally alter the epigenome and transcriptome, and are not suitable for targeting individual loci [68]. Recently, an epigenome editing strategy for targeted histone acetylation, which is strongly associated with active gene regulatory elements and enhancers, has been described. It leverages the recent emergence of the CRISPR/dCas9 system as a versatile genome engineering platform and presents an easily programmable approach that facilitates robust control of the epigenome and downstream gene expression [69].

In the past decade, deep learning has achieved considerable success in predicting gene expression from epigenetic marks, such as transcription factor binding [70], chromatin accessibility [71], histone marks [66], [72], [73] and DNA methylation [74]. However, whether such computational models gain an implicit understanding of mechanistic, causal relationships between various epigenetic marks and gene expression is an important yet underexplored question. Motivated by this understanding, we developed a model of how chromatin state affects gene expression, by leveraging the data available through ENCODE. We then used the model to predict the effect of epigenetic edits to investigate whether such models do indeed learn a causal understanding of gene regulation and the Histone Code hypothesis. Our model learns a sensible understanding of chromatin structure which is consistent with known patterns of various histone modifications [75].

## 3.2 Results

### Building a gene expression model

We obtained ChIP-seq data from the ENCODE Imputation Challenge [76]. The corresponding histone modifications and cell types used in this study are outlined in **Table 3.1**. Metagene plots describing the various epigenetic marks in different cell types revealed clear batch effects, partly due to inconsistent sequencing depth (**Figure 3.1**). In order to correct for these batch effects, we devised a heuristic technique adapted from S3norm [77], described in detail in the Methods section [3.4]. After correction, the batch effects were considerably reduced (**Figure 3.2**) and these post-processed tracks are what were used for the remainder of the analyses. We also obtained polyA-plus RNA-seq data for each of the 13 cell types from the ENCODE data portal and processed them as described in the Methods section to obtain the gene expression values for each gene [3.4].

Cell Type	polyA Plus RNA-seq	H3K36me3	H3K27me3	H3K27ac	H3K4me1	H3K4me3	H3K9me3
IMR-90	T	T	T	T	T	T	T
H1-hESC	T	T	T	T	T	T	T
trophoblast cell	T	T	T	T	T	T	T
neural stem progenitor cell	T	T	T	T	T	T	T
K562	T	T	T	T	T	T	T
heart left ventricle	T	T	T	T	T	T	T
adrenal gland	T	T	T	T	T	T	T
endocrine pancreas	T	T	T	T	T	T	T
peripheral blood mononuclear cell	T	T	T	T	T	T	T
amnion	T	T	T	T	T	T	T
myoepithelial cell of mammary gland	T	T	T	A	T	T	T
chorion	T	T	T	T	T	T	A
HEK293T	T*	T*	A	T*	T*	T*	T*

Table 3.1: ChIP-seq  $-\log_{10}(\text{p-values})$  were obtained from the ENCODE Imputation Challenge where the ground truth data were available (corresponding to entries labeled **T** in the table). Avocado imputations were downloaded from the ENCODE data portal, where ground truth data were not available (corresponding to entries labeled **A** in the table). Entries labeled with an \* are obtained from the HEK293 cell line because data for the HEK293T cell line was not available

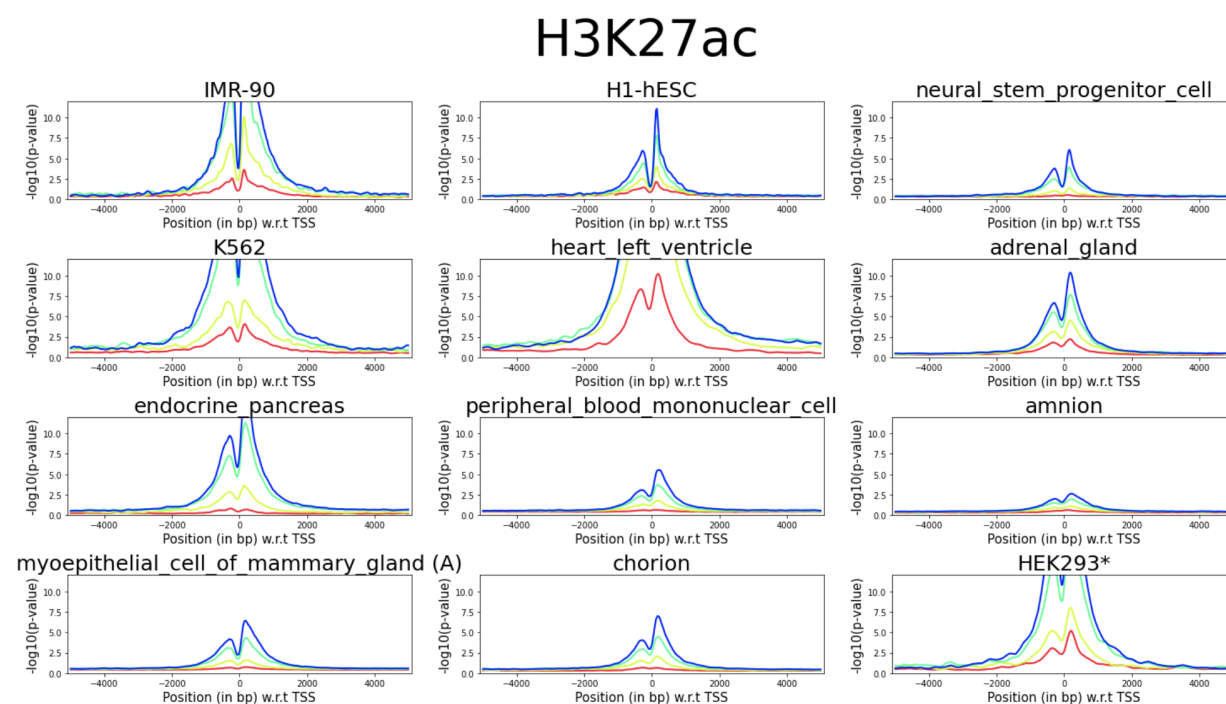


Figure 3.1: Metagene plots for different cell types for unnormalized H3K27ac ChIP-seq data across gene expression quantiles (blue is the highest and red is the lowest gene expression quantile). \* represents data from HEK293 instead and (A) represents Avocado imputed data.

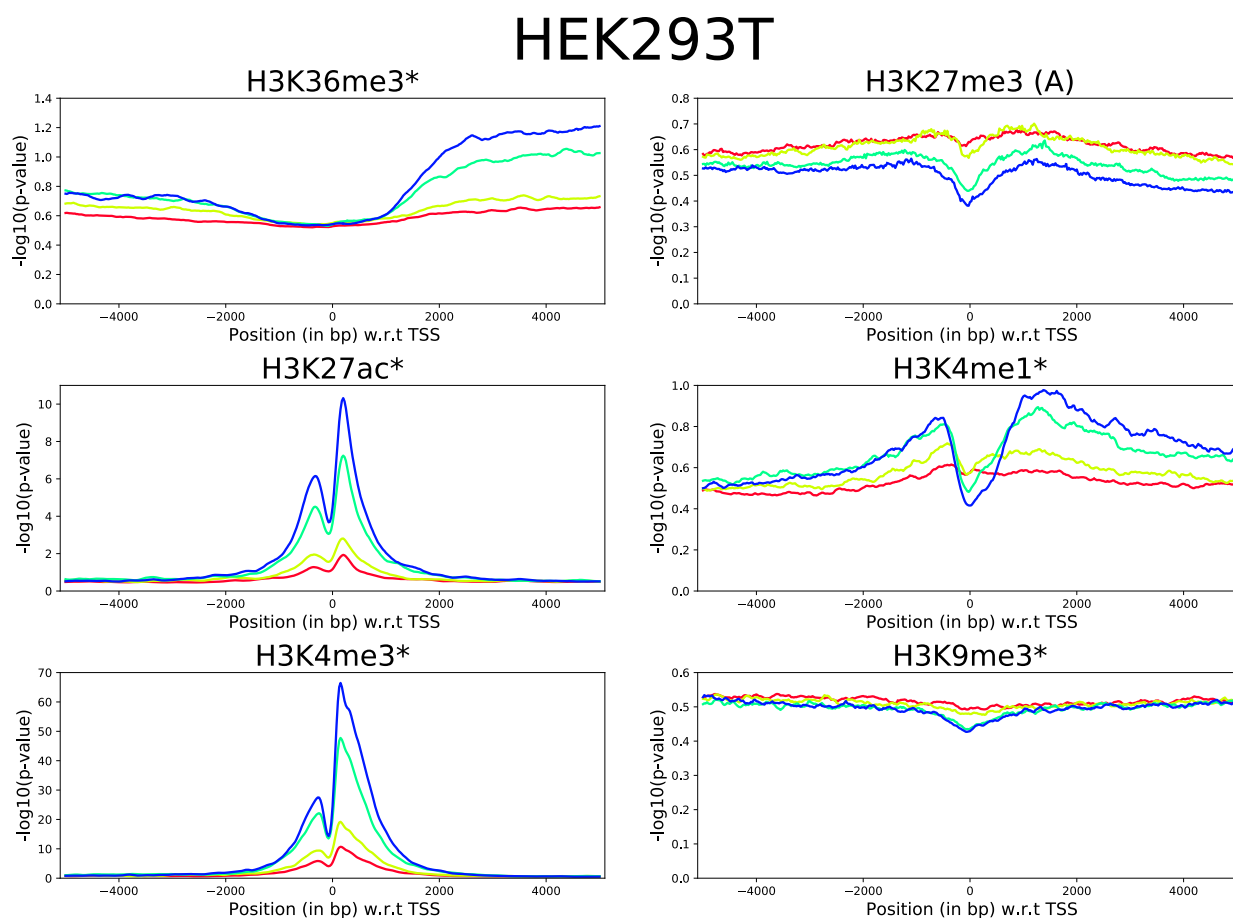


Figure 3.2: Metagene plots for the six different histone marks in HEK293T, after S3norm-based normalization, across gene expression quantiles (blue is the highest and red is the lowest gene expression quantile). \* represents data from HEK293 instead and (A) represents Avocado imputed data for HEK293T.

We trained a convolutional neural network model, described in detail in the Methods section, to predict the gene expression of each gene in each of the 13 cell types using histone modification data centered at the TSS [3.4]. We observe that the Spearman correlation between the true gene expression and the model’s predicted gene expression improves as the input context size increases and at all input context sizes, the convolutional neural network models outperform a ridge regression model trained on the same data (Figure 3.3).



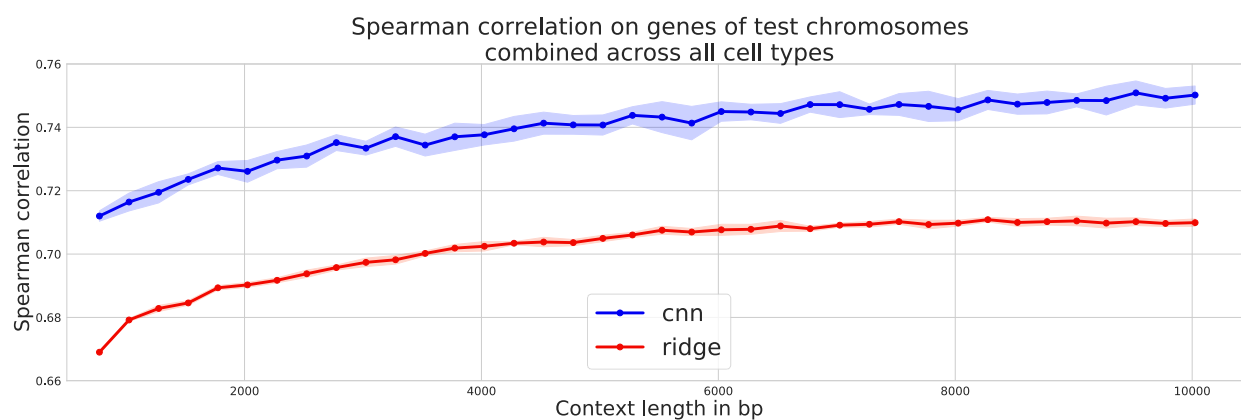


Figure 3.3: Spearman correlation on genes in the test dataset for different input context lengths. Blue curve is the mean across 10 computational replicates of CNNs and the red is the mean across 10 computational replicates of ridge regression. Shaded area represents standard deviation in the Spearman correlation across the 10 computational replicates.

In order to assess the models' ability to generalize to unseen cell types, we trained a set of models for each cell type, while holding it out during training and testing the model's predictive capabilities on genes in this held-out cell type. We observe that the CNNs are able to outperform a ridge regression model on this cross-cell type generalization task and perform quite well across a wide range of cell types; although the Spearman correlation does vary across the different cell types (Figure 3.4).

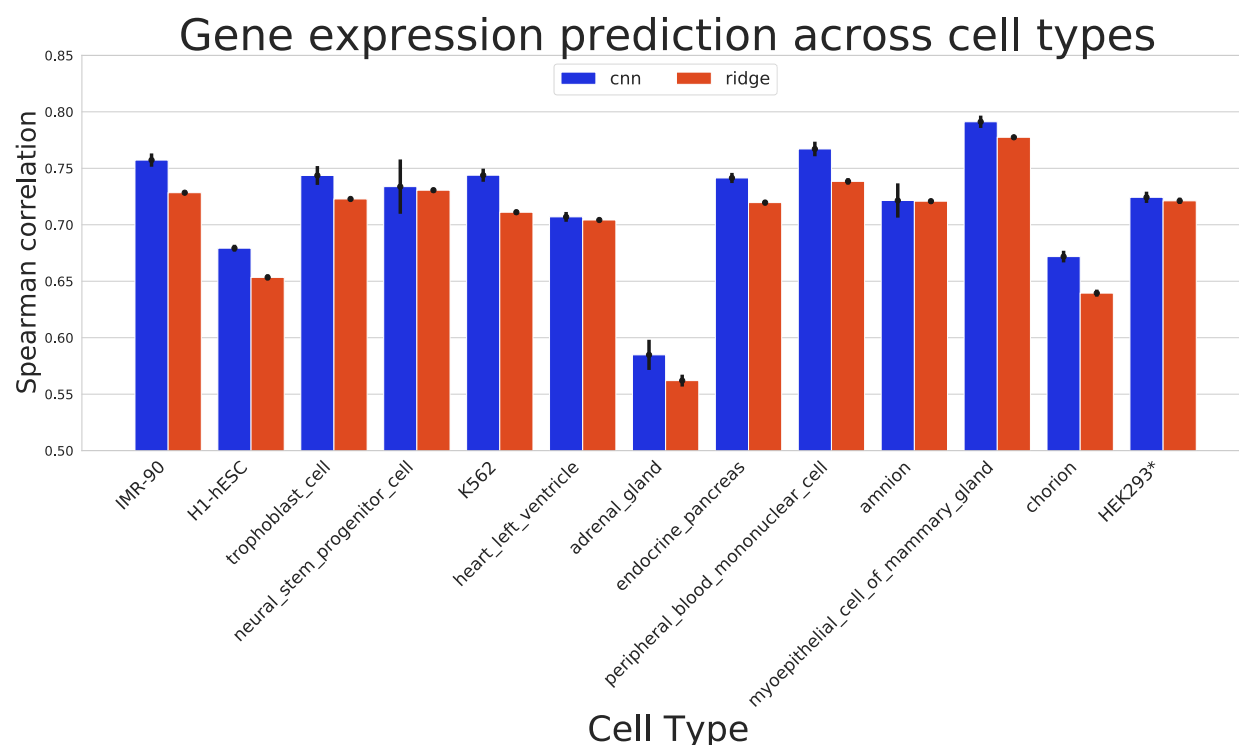


Figure 3.4: Spearman correlation on genes of cell types held out during training. The barplots represent the mean across 10 computational replicates and the error bars represent their standard deviation.

We also assessed the models' predictive capabilities in an orthogonal manner by computing how well the models rank gene expression across cell types within each gene. We did so by computing Spearman correlations between the true gene expression and the predicted gene expression for each gene in the test data. The distribution of the resulting Spearman correlations suggests that the CNNs are able to better rank cell types by gene expression than ridge regression (**Figure 3.5**). In particular, the median gene (corresponding to  $y = 0.5$  in the empirical cumulative density function), achieves a cross-cell type correlation of  $\sim 0.53$  with the CNNs; while only achieving a cross-cell type correlation of  $\sim 0.39$  with ridge regression.

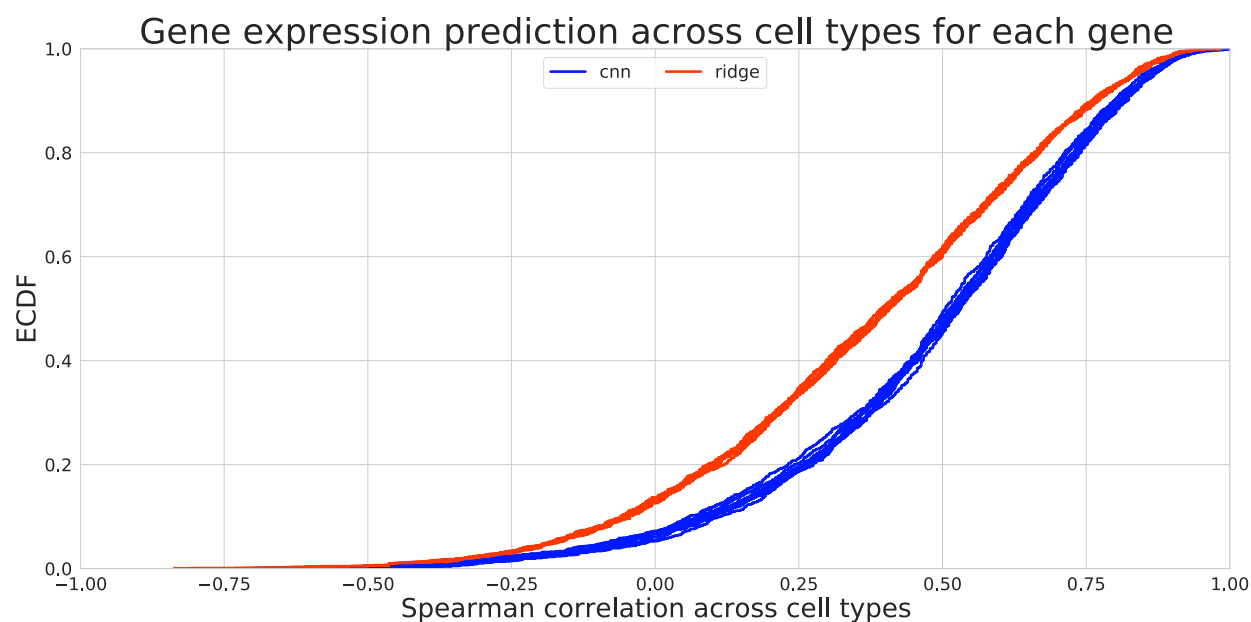


Figure 3.5: Distribution of spearman correlation across cell types, computed for each gene in the test data. The different curves represent 10 computational replicates for each model type.

### *in silico* perturbations of the gene expression model

In order to understand what the models trained to predict gene expression have learnt, we perform *in silico* perturbation of the histone modifications, as described in the Methods section, one-by-one at each position in the input context and measured the predicted fold-change in the gene expression (Figure 3.6 3.4). We observed that the patterns in these perturbation plots closely resemble known patterns of the various epigenetic marks (Figure 3.2) 75.

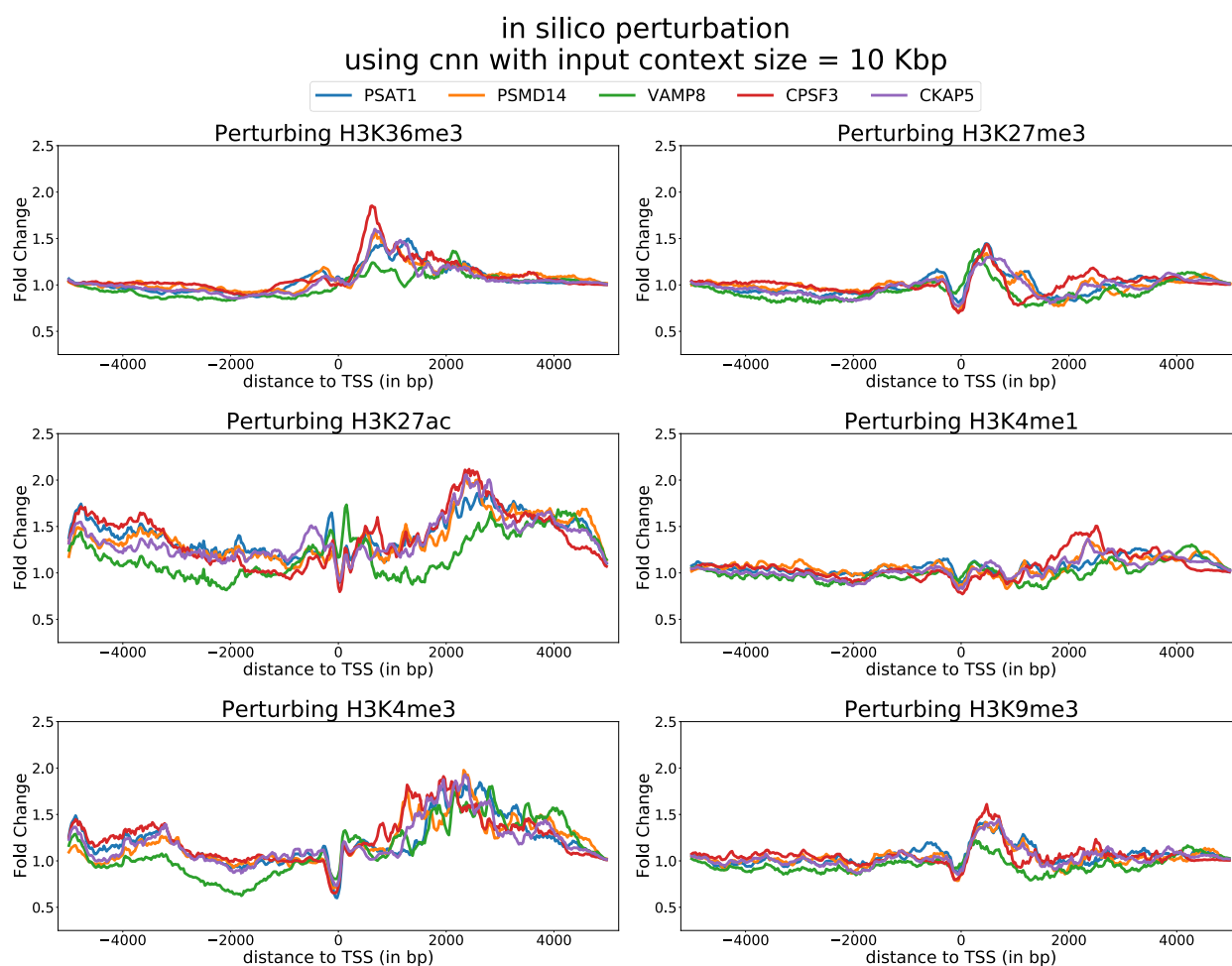


Figure 3.6: Each point on the X-axis corresponds to *in silico* perturbation of that assay at that position and the Y-axis measures the predicted fold-change. The five different lines correspond to five different genes in the HEK293T cell type.

### 3.3 Discussion

We have trained models to predict gene expression using histone modification data which are highly predictive of gene expression of genes on held-out chromosomes. These models are also able to generalize to genes in unseen cell types. We developed a novel metric to assess the models' ability to rank various cell types by gene expression, within each gene, and demonstrate that the trained models perform quite well at this task as well. These models have good predictive power because they seemed to have learnt known patterns of histone modifications that are correlated with gene expression.

These models are different from existing models such as Enformer [78] that predict gene

expression using DNA sequence as the input. The benefit of models such as Enformer is being able to assess the impact of mutations in DNA sequence. In contrast, our trained models could be used to assess the impact of perturbations in the histone modifications which can now be achieved using CRISPR-dCas9 based epigenome editing [69]. In the future, these models can be useful for designing sgRNAs for epigenome editing to achieve desired gene expression fold-changes.

## 3.4 Methods

### Data preparation

We obtained  $-\log_{10}(\text{p-value})$  ChIP-seq tracks created by running the MACS2 peak-caller [79] on read count data, from the ENCODE Imputation Challenge [76]. For three tracks where data were not available, we downloaded Avocado [80] imputations from the ENCODE data portal [59]. We binned each epigenetic track at 25 base pair resolution and pre-processed them with an additional log operation before inputting them into the network.

We downloaded polyA-plus RNA-seq gene expression TPM values for each of the 12 cell types in Table 3.1, from the ENCODE data portal [59] and preprocessed them with a log operation.

### Normalizing $p$ -values by adapting S3norm

We assigned IMR-90 to be a reference cell type, for each of the six epigenetic assays and kept its  $p$ -values unchanged. We then performed a transformation for each of the remaining cell types adapted from the core technique developed by S3norm [77], in order to normalize each epigenetic assay in each of these remaining cell types, with respect to the corresponding epigenetic assay in IMR-90.

First, we computed *peaks* in both, the reference as well as the target cell type. *Peaks* were defined as the 25 base pair bins corresponding to FDR-adjusted  $p$ -values less than 0.05 [81]. For assays that were obtained from Avocado imputations (due to lack of availability of experimental data), *peaks* were defined to be the 1000 bins containing the smallest Avocado imputed  $p$ -values, based on suggestions from the authors of Avocado [80]. All the remaining bins were defined to be *background*, for both, the reference as well as the target cell types.

We then computed the list of *peaks* that were common to both the reference and the target cell types. These were termed, *common peaks*. Similarly, we defined *common background* as the list of bins that were assigned to be *background* in both, the reference as well as the target cell types.

The S3norm method was designed to work with count data, which is always  $\geq 1$ . However, the epigenetic tracks, which are represented as  $-\log_{10}(p\text{-values})$ , are not guaranteed to always be  $\geq 1$ , hence, we transformed all epigenetic tracks by adding 1 to the  $-\log_{10}(p\text{-values})$ , in both the reference as well as the target cell types.

Additionally, since the epigenetic tracks obtained from imputations performed by Avocado were not guaranteed to be distributed similar to experimental  $-\log_{10}(p\text{-values})$ , we scaled all the epigenetic tracks (both experimental as well as Avocado imputations) by dividing them by the minimum observed value in *common peaks* and common background, in order to bring experimental data and Avocado imputations onto a similar footing. In particular, before applying the S3norm normalization, we transformed  $-\log_{10}(p\text{-values})$  in *common peaks* and *common background* for both the reference as well as the target cell type as following:

$$\text{TransformedCommonPeaks}_{i,\text{reference}} = \frac{1 + \text{CommonPeaks}_{i,\text{reference}}}{\min_i(\text{CommonPeaks}_{i,\text{reference}})} \quad (3.1)$$

$$\text{TransformedCommonPeaks}_{i,\text{target}} = \frac{1 + \text{CommonPeaks}_{i,\text{target}}}{\min_i(\text{CommonPeaks}_{i,\text{target}})} \quad (3.2)$$

$$\text{TransformedCommonBackground}_{i,\text{reference}} = \max\left(\frac{1 + \text{CommonBackground}_{i,\text{reference}}}{\min_i(\text{CommonBackground}_{i,\text{reference}})}, 0\right) \quad (3.3)$$

$$\text{TransformedCommonBackground}_{i,\text{target}} = \max\left(\frac{1 + \text{CommonBackground}_{i,\text{target}}}{\min_i(\text{CommonBackground}_{i,\text{target}})}, 0\right) \quad (3.4)$$

The normalization procedure of S3norm then wishes to find two positive parameters,  $\alpha$  and  $\beta$  that are to be learned from the data such that both the following equations are satisfied:

$$\text{mean}(\text{TransformedCommonPeaks}_{\text{reference}}) = \text{mean}(\alpha \times \text{TransformedCommonPeaks}_{\text{target}}^\beta) \quad (3.5)$$

$$\text{mean}(\text{TransformedCommonBackground}_{\text{reference}}) = \text{mean}(\alpha \times \text{TransformedCommonBackground}_{\text{target}}^\beta) \quad (3.6)$$

Specifically,  $\alpha$  is a scale factor that shifts the transformed  $-\log_{10}(p\text{-values})$  of the target data set in log scale, and  $\beta$  is a power transformation parameter that rotates the transformed  $-\log_{10}(p\text{-values})$  of the target data set in log scale. There is one and only one set of values for  $\alpha$  and  $\beta$  that can simultaneously satisfy both the above equations for *common peaks* and the *common background*.

The values of  $\alpha$  and  $\beta$  are estimated by the Powell minimization method implemented in scipy [82], [83]. The resulting normalized  $-\log_{10}(p\text{-values})$  are used for all downstream analyses in this work (Figure 3.2).

## Training a model

We implemented a convolutional neural network to predict gene expression using epigenetic features centered at the Transcription Start Site (TSS) of each gene. The epigenetic features, obtained from the ENCODE Imputation Challenge (as described above), correspond to  $-\log_{10}(p\text{-values})$  for each 25 base pair bin in the genome, are first transformed with a  $\log_e(x + 1)$  transformation. We then use an input context size of 10,000 base pairs centered at the TSS of each gene (**Figure 3.3**). This corresponds to 200 25 base pair bins on either side of the TSS of each gene (including the bin containing the TSS).

The transformed epigenetic features for each gene, are processed with successive convolutional blocks. Each convolutional block consisted of a batch-normalization layer, rectified linear units (ReLU), a convolutional layer consisting of 32 convolutional kernels, each of width 5, followed by a dropout with 0.1 probability and finally a pooling layer to gradually reduce the dimension of the features. After being processed with 5 such convolutional blocks, the output is flattened and passed through a fully connected layer consisting of 16 neurons and a ReLU activation. This is finally processed with a fully connected layer with a single output and a linear activation, since this is a regression task. The model is trained with a mean squared error loss and the Adam optimizer with learning rate of 0.001 for the first 50 epochs and 0.0005 for the remaining 50 epochs. Training took about 0.5 hours on 1 NVIDIA A100 Tensor Core GPU. All associated code is available at <https://github.com/sanjitsbatra/deepENCODE>.

## *in silico* perturbation of epigenetic data

The epigenetic features of gene  $g$  are represented by,

$$E_i^{CT,g} = \ln(1 - \log_{10} p)$$

where,  $CT$  stands for cell type  $CT$ ,  $g$  stands for the gene,  $i$  corresponds to one position within the epigenetic features  $E^{CT,g} \in \mathbb{R}^W$  and  $p$  corresponds to the  $p$ -value corresponding to the ChIP-seq track at that position.  $W$  is chosen to be 10,000 base pairs divided by the resolution of the epigenetic features, which is 25 base pairs, which gives us  $W = 400$ .

We define a perturbation function  $F : \mathbb{R}^{W+3} \rightarrow \mathbb{R}^W$ , which is defined as:

$$F(E^{CT,g}, j, \Delta, d) = \begin{cases} E_i^{CT,g} + \Delta & \text{if } |i - j| \leq d \\ E_i^{CT,g} & \text{if } |i - j| > d \end{cases}$$

We then denote a trained model, as  $T : \mathbb{R}^W \rightarrow \mathbb{R}$ , which converts the epigenetic features  $E^{CT,g}$  to gene expression  $\log_{10}(TPM + 1)$ . The predicted fold-change after perturbing position  $j$  is then computed as:

$$\frac{10^{T(F(E^{CT,g}, j, \Delta, 6))} - 1}{10^{T(E^{CT,g})} - 1}$$

## Chapter 4

# Predicting translation initiation from DNA sequence

This is the product of a multi-year collaboration with Kishore Jaganathan and Kyle Kai-How Farh at the Illumina AI Lab, along with Dan Daniel Erdmann-Pham and my advisor Professor Yun S. Song and the manuscript is currently under preparation.

### 4.1 Introduction

Translation is the generation of proteins from messenger RNA (mRNA). The control of protein expression through translation is a fundamental process of living cells. The nucleotide sequence of mRNA not only determines the protein's sequence but also has an impact on the efficiency of the translation process [84]. There are various regulatory elements embedded within mRNA sequences, such as 5'-untranslated regions (5'UTRs), that impact a protein's expression [85]. 5'UTR sequences encode a variety of cis-regulatory elements, including a 5'-cap structure [86], a translation initiation motif, referred to as the Kozak sequence [87], upstream AUGs (uAUGs) and upstream ORFs [88], internal ribosome entry sites [89], G-quadruplexes [90] and secondary structures [91].

Translation initiation plays an important role in mRNA translation, in which the methionyl-tRNA unique for initiation (Met-tRNA<sub>i</sub>) identifies the AUG start codon and triggers the downstream translation process [92]. In addition to the GENCODE annotated translation initiation sites (mAUGs), the translation process may also start at alternative codons, most often these are upstream AUGs (uAUGs), which have been increasingly detected with the advent of several high-throughput sequencing techniques for profiling initiating ribosomes [93]. Translation initiation at uAUGs leads to upstream open reading frames (uORFs) that can be translated into short peptides or affect the protein expression levels of the main ORFs (mORFs) and cause diseases such as cancer [94].

Consequently, computational methods have been developed to predict where translation initiates [91]. Some of the more recent methods can predict where translation initiation



occurs using mRNA sequence using deep learning [95][96]. However, whether such methods are able to predict the impact of mutations in 5'UTRs remains to be shown.

In this work, we train a deep neural network to predict where translation initiation occurs using mature mRNA sequence alone and demonstrate that such a model is able to predict whether translation initiates at held-out uAUGs. Furthermore, we demonstrate that the trained models are able to accurately capture the effects of mutations in the 5'UTR, thereby paving the way for optimizing protein expression by perturbation of the 5'UTR sequence.

## 4.2 Results

### Building a model to predict translation initiation from sequence

We train a neural network to predict whether a scanning ribosome would initiate translation at a particular position using the surrounding sequence context (Figure 4.1). To train the model on the human genome, we considered all AUGs of ORFs annotated by GENCODE v39 as the positive dataset. In addition to known mORFs in the human genome, we obtained a list of upstream AUGs (uAUGs) from [97] where ribosomes were observed to initiate via ribosome profiling, and added them to the positive dataset.

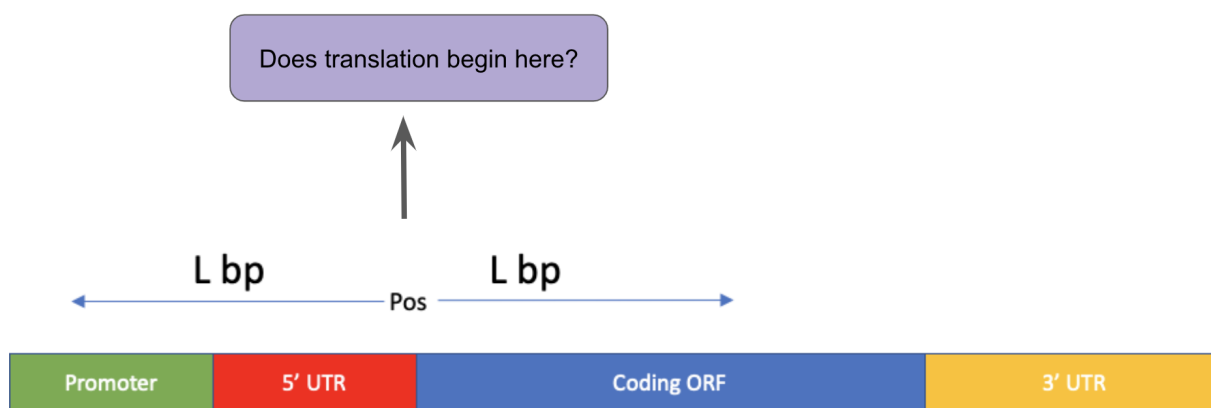


Figure 4.1: **Schematic for predicting translation initiation from mature mRNA sequence**

There are a variety of choices that can be made to construct the negative dataset for training the model, in addition to considering all non-AUG positions in the 5'UTR sequences. In particular, we explore two different strategies by looking at the distribution of distance of each uAUG to either the TSS or the length of the upstream ORF induced by the uAUG (Figure 4.2). Since there are no observed **Ribo-seq certified uAUGs** at the tails of these distributions, we define the negative set to be all those uncertified uAUGs (uAUGs

that were not observed to be initiated at by ribosomes in ribosome profiling experiments) that lie in their right tails.

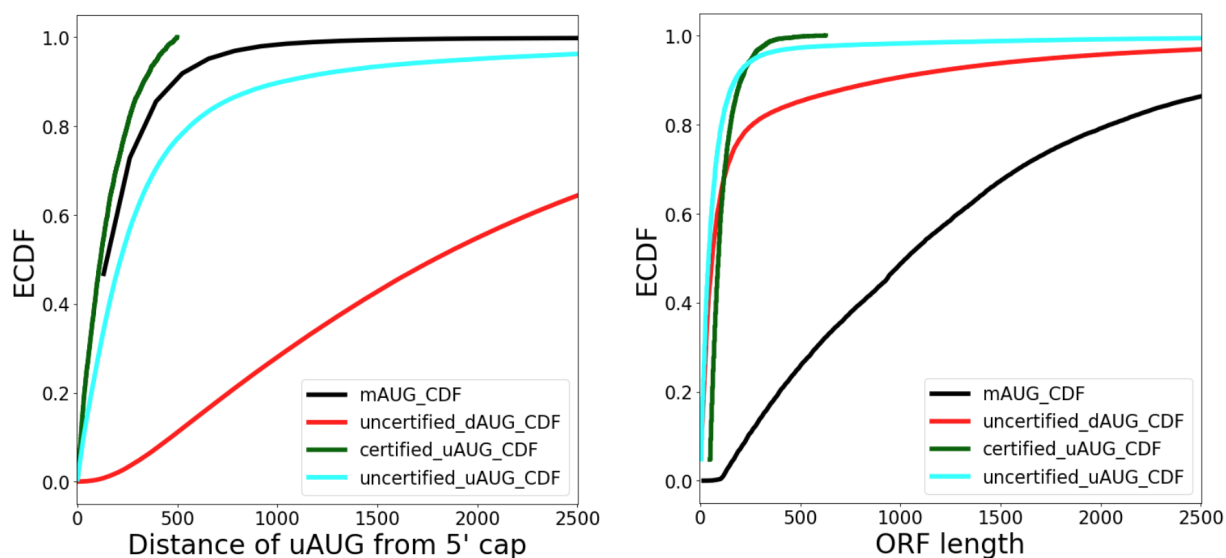


Figure 4.2: **Empirical cumulative distribution of uAUG properties can be used to define negative dataset for training the model**

We train a convolutional neural network, described in detail in the Methods section, using mature mRNA sequence and binary labels to predict where translation initiation occurs [4.4](#). We observe that by using uORF length to determine the negative dataset, the predictive performance of the model increases and reaches a sweet spot around a cut-off of  $\sim 700$  base pairs. With this choice of uORF length cut-off to determine the negative dataset for training, the corresponding model trained with mAUGs from GENCODE v39 and ribosome profiling certified uAUGs as the positive dataset, achieves high predictive accuracy in being able to distinguish ribosome profiling certified uAUGs from other AUGs (**Figure [4.3](#)**).

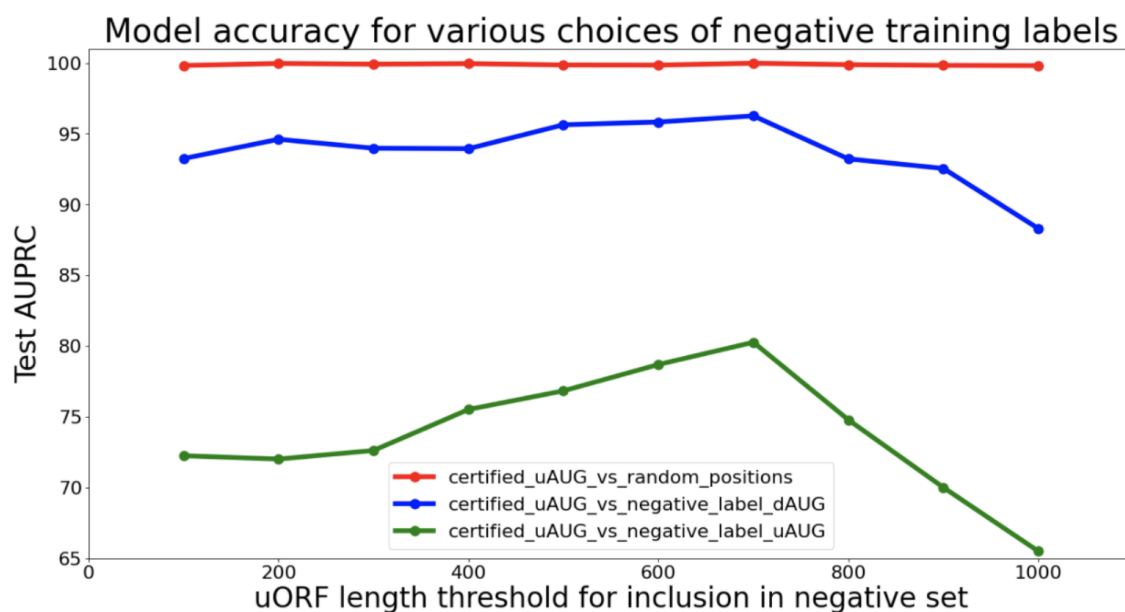


Figure 4.3: Accuracy metrics as uORF length cut-off varies (for determining the negative dataset)

## Interpreting the trained model

To understand the features that the neural network learns, we scan an AUG motif through the input sequence context and measure the model's prediction. We observe that the mean prediction of the model across many genes (**Figure 4.4**) mirrors the observed probability of uAUGs in different frames w.r.t to the mAUG (**Figure 4.5**). This suggests that the model has learnt various statistical properties about uAUGs that might allow it to distinguish certified uAUGs from uncertified uAUGs.

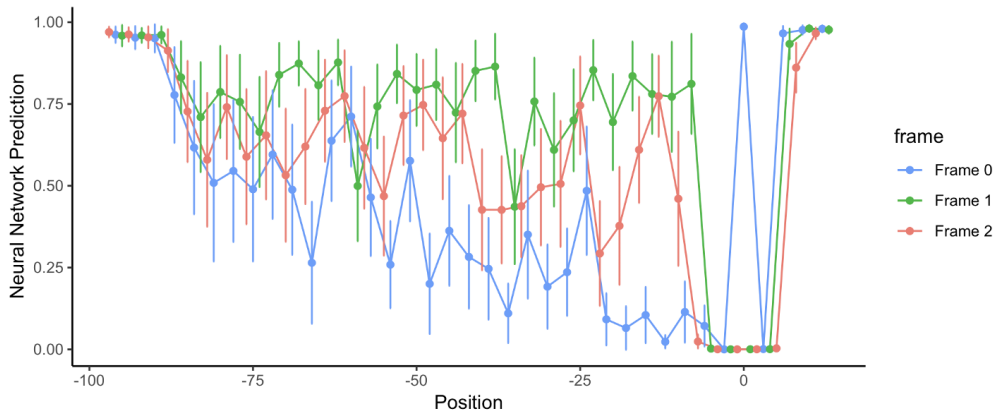


Figure 4.4: Model predictions as an AUG is scanned across the input context, aggregated across 1000 genes

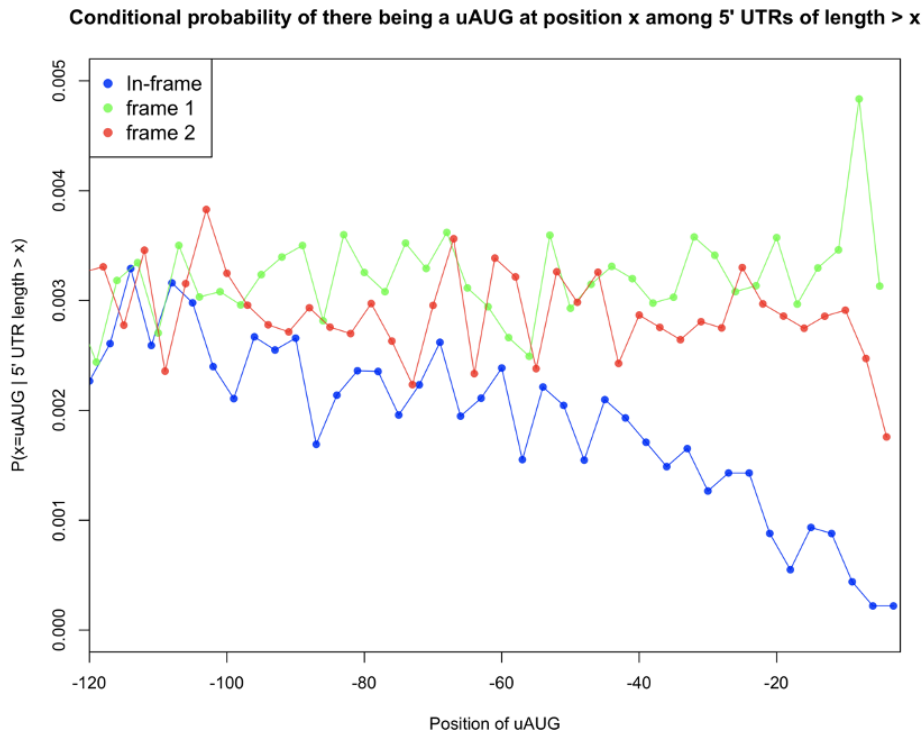


Figure 4.5: Probability of observing an uAUG in 5'UTRs of the human genome. X-axis at  $x = 0$  denotes the mAUG of each gene.

To probe whether the model has built an understanding of Kozak sequences [87], we

scanned a *strong* Kozak sequence across the input context of a gene with a *weak* endogenous Kozak sequence at its mAUG, and observed that the model predictions were lowered over a long range of positions (**Figure 4.6**). Moreover, when the *strong* Kozak sequence was inserted at the mAUG, the model prediction increased. This suggests that the model has learnt an implicit ranking of Kozak sequences.

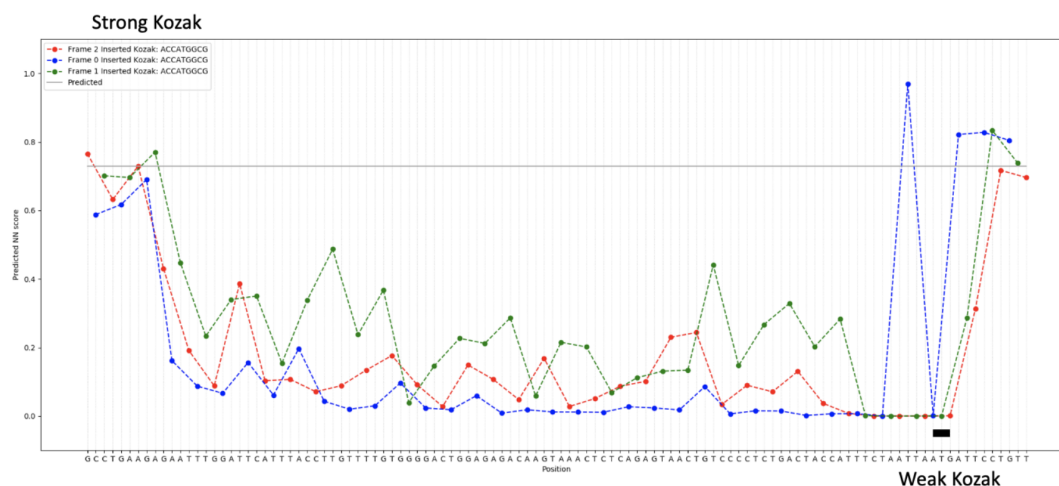


Figure 4.6: Model predictions as a *strong* Kozak sequence, corresponding to ACCATGGCG is scanned across the input context, whose mAUG exists in a *weak* Kozak context of TTAATGATT

## Variants scored highly by the model tend to be under negative selection in gnomAD

We scored all variants observed in gnomAD [98][99], within 100bp of the mAUG on 5'UTRs of transcripts in the human genome with the trained model. We did so by computing the difference between the prediction of the model with the reference allele inserted vs with the alternate allele inserted. We computed the maximum across all such differences centering the input sequence at every possible uAUG and the mAUG.

We assess the model's ability to discern whether a variant is under negative selection and therefore deleterious by performing a Fisher's exact test using the variants observed in gnomAD, as described in detail in the Methods section 4.4. **Figure 4.7** shows the resulting odds ratios from Fisher's exact tests performed on the single nucleotide variants in the gnomAD dataset. The X-axis describes the pathogenicity threshold chosen for the model scores and the Y-axis represents the corresponding odds ratio. If we assume that *all* uAUG variants are pathogenic, the corresponding odds ratio obtained in gnomAD is reflected by the dashed line. We observe that the trained neural network vastly outperforms this baseline and

therefore demonstrates that the model has the ability to pick out pathogenic variants among the uAUG variant class.

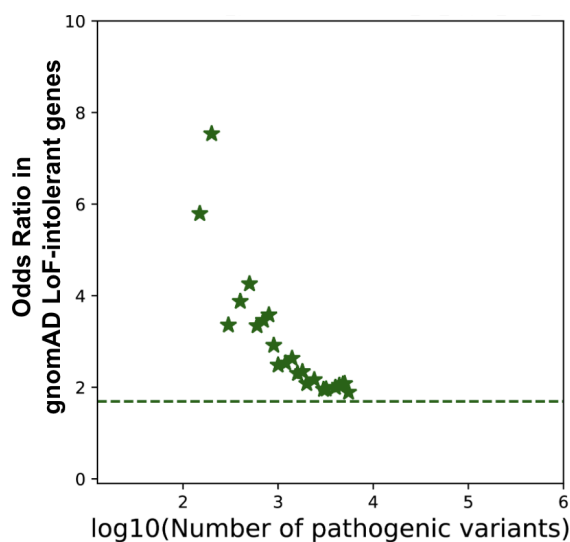


Figure 4.7: Variants with high model scores tend to be more rare in gnomAD. Stars represent significant odds ratios. If we assume that *all* uAUG variants are pathogenic, the corresponding odds ratio obtained in gnomAD is reflected by the dashed line.

### 4.3 Discussion

In this study, we develop a method to predict where a scanning ribosome initiates translation using surround sequence context. We investigate the role of different choices for choosing the negative dataset on model performance and demonstrate that the trained neural network is able to predict held out uORFs with high accuracy. We further show that the model is able to achieve high predictive accuracy by learning the statistical patterns of uAUG incidence in endogenous 5'UTRs of the human genome.

In order to assess the model's ability to capture effect of mutations, we show that mutations scored highly by the trained neural network are under stronger negative selection (i.e. they tend to be more rare in healthy individuals) than those having a low model score, in a large cohort of healthy individuals, namely gnomAD.

This model paves way to a better understanding of translation initiation and could be utilized to optimize protein expression by modifying the sequence of the 5'UTR.

## 4.4 Methods

### Data preparation

To obtain the mAUGs, we downloaded the latest GENCODE v39 annotation and subsetted to lines with the annotation of **start codon**. We then processed ribosome profiling certified uAUGs from [97]. The authors present a list of  $\sim 4,000$  uAUGs and we obtained their coordinates in the hg38 reference genome and added them to the positive dataset.

To obtain the negative dataset, we computed various properties of uAUGs such as their distance from the TSS and the length of the induced uORF and trained models for various choices of cut-offs for these properties. We then chose the model which performed best in its ability to distinguish certified uAUGs from uncertified uAUGs.

### Training the model

We trained a neural network with  $L = 200$  base pairs on each side of a position corresponding to each data point (Figure 4.7). More precisely, the input to the models is a sequence of one-hot encoded nucleotides, where A, C, G, and T (or equivalently U) are encoded as [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1] respectively and the output of the model corresponds to the probability of a scanning initiating translation at that position.

The architecture of the neural network is inspired by SpliceAI [100]. The basic unit of the SpliceAI architecture is a residual block, which consists of batch-normalization layers, rectified linear units (ReLU), and convolutional units. Residual blocks are commonly used when designing deep neural networks. The architecture consists of 12 stacked residual blocks connecting the input layer to the penultimate layer, and a convolutional unit with a sigmoid activation connecting the penultimate layer to the output layer.

Each residual block has three hyper-parameters  $N$ ,  $W$ , and  $D$ , where  $N$  denotes the number of convolutional kernels,  $W$  denotes the window size, and  $D$  denotes the dilation rate of each convolutional kernel. Since a convolutional kernel of window size  $W$  and dilation rate  $D$  extracts features spanning  $(W - 1) \times D$  neighboring positions, a residual block with hyper-parameters  $N$ ,  $W$ , and  $D$  extracts features spanning  $2 \times (W - 1) \times D$  neighboring positions. Hence, the total neighbor span of the SpliceAI architectures is given by  $S = \sum_{i=1}^{12} 2 \times (W_i - 1) \times D_i$ , where  $N_i$ ,  $W_i$  and  $D_i$  are the hyperparameters corresponding to the  $i^{th}$  residual block, which were chosen so that  $S = 400$ .

The model is trained with a binary cross entropy loss and the Adam optimizer with learning rate of 0.001 for the first 50 epochs and 0.0005 for the remaining 50 epochs. Training took about 2 hours on 1 NVIDIA A100 Tensor Core GPU. All associated code used to train the model is available at <https://github.com/sanjitsbatra/5UTR>.

## gnomAD analysis

We obtained gnomAD variant data from gnomAD v3 [98] and processed the variants using the ENSEMBL variant effect predictor [101]. We subsetted to variants that were within 100bp of the mAUG on each transcript and used the UTRannotator [102] to predict whether the mutation was an uAUG mutation or not. To score each variant, we performed a forward pass through the trained model with the input context centered at the variant's position, with the reference and alternate alleles and obtaining the difference in the corresponding model scores.

We then defined rare variants as those with allele count equal to 1, and common variants as those with allele count 1000 or more. Model scores were defined as pathogenic or benign based on a threshold that determined how many variants were classified as pathogenic. We then constructed a  $2 \times 2$  table whose columns were **rare** and **common**. The rows of the table were defined as pathogenic and **benign**. A variant was deemed **pathogenic** if the model score for that variant was higher than the chosen *threshold* and the variant was a uAUG variant. We then calculated the odds ratio and p-value of a Fisher's exact test performed on this contingency table. We carried out this procedure for various choices of the model score *threshold* and report the corresponding odds ratios in **Figure 4.7**



## Chapter 5

### Conclusions

The computational methods developed in this work can pave the way for a better understanding of how the central dogma of biology is affected by diseases such as cancer. These tools can prioritize disease therapies based on their impact on DNA, in the form of large-scale structural variants, gene regulation and protein abundance. This would open the doors for a more efficient search for disease therapies accelerated by computational engines.

#### 5.1 Future Work

Advancements in experimental techniques such as Hi-C, CRISPR/dCas9 and ribosome profiling could lead to improvements in the computational methods developed in this work. There will undoubtedly be vast synergistic potential to leverage the latest breakthroughs in machine learning along with these experimental advancements to improve the presented computational tools.

# Bibliography

- [1] P C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, October 1976.
- [2] Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, Joshua F McMichael, John W Wallis, Charles Lu, Dong Shen, Christopher C Harris, David J Dooling, Robert S Fulton, Lucinda L Fulton, Ken Chen, Heather Schmidt, Joelle Kalicki-Veizer, Vincent J Magrini, Lisa Cook, Sean D McGrath, Tammi L Vickery, Michael C Wendl, Sharon Heath, Mark A Watson, Daniel C Link, Michael H Tomasson, William D Shannon, Jacqueline E Payton, Shashikant Kulkarni, Peter Westervelt, Matthew J Walter, Timothy A Graubert, Elaine R Mardis, Richard K Wilson, and John F DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, January 2012.
- [3] Samuel Aparicio and Carlos Caldas. The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.*, 368(9):842–851, February 2013.
- [4] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219, March 2013.
- [5] J R Naegele, Y Arimatsu, P Schwartz, and C J Barnstable. Selective staining of a subset of GABAergic neurons in cat visual cortex by monoclonal antibody VC1.1. *J. Neurosci.*, 8(1):79–89, January 1988.
- [6] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, 11(4):396–398, April 2014.
- [7] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, 16:35, February 2015.

- [8] Layla Oesper, Ahmad Mahmood, and Benjamin J Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.*, 14(7):R80, July 2013.
- [9] C B Lozzio and B B Lozzio. Human chronic myelogenous leukemia cell-line with positive philadelphia chromosome. *Blood*, 45(3):321–334, March 1975.
- [10] R Kurzrock, J U Gutterman, and M Talpaz. The molecular genetics of philadelphia chromosome-positive leukemias. *N. Engl. J. Med.*, 319(15):990–998, October 1988.
- [11] Richard P Junghans. The challenges of solid tumor for designer CAR-T therapies: a 25-year perspective. *Cancer Gene Ther.*, 24(3):89–99, March 2017.
- [12] Alisa M Goldstein and Margaret A Tucker. Dysplastic nevi and melanoma. *Cancer Epidemiol. Biomarkers Prev.*, 22(4):528–532, April 2013.
- [13] Eran Hodis, Ian R Watson, Gregory V Kryukov, Stefan T Arold, Marcin Imielinski, Jean-Philippe Theurillat, Elizabeth Nickerson, Daniel Auclair, Liren Li, Chelsea Place, Daniel Dicara, Alex H Ramos, Michael S Lawrence, Kristian Cibulskis, Andrey Sivachenko, Douglas Voet, Gordon Saksena, Nicolas Stransky, Robert C Onofrio, Wendy Winckler, Kristin Ardlie, Nikhil Wagle, Jennifer Wargo, Kelly Chong, Donald L Morton, Katherine Stemke-Hale, Guo Chen, Michael Noble, Matthew Meyerson, John E Ladbury, Michael A Davies, Jeffrey E Gershenwald, Stephan N Wagner, Dave S B Hoon, Dirk Schadendorf, Eric S Lander, Stacey B Gabriel, Gad Getz, Levi A Garraway, and Lynda Chin. A landscape of driver mutations in melanoma. *Cell*, 150(2):251–263, July 2012.
- [14] A Hunter Shain, Iwei Yeh, Ivanka Kovalyshyn, Aravindhan Sriharan, Eric Talevich, Alexander Gagnon, Reinhard Dummer, Jeffrey North, Laura Pincus, Beth Ruben, William Rickaby, Corrado D’Arrigo, Alistair Robson, and Boris C Bastian. The genetic evolution of melanoma from precursor lesions. *N. Engl. J. Med.*, 373(20):1926–1936, November 2015.
- [15] Nicholas K Hayward, James S Wilmott, Nicola Waddell, Peter A Johansson, Matthew A Field, Katia Nones, Ann-Marie Patch, Hojabr Kakavand, Ludmil B Alexandrov, Hazel Burke, Valerie Jakrot, Stephen Kazakoff, Oliver Holmes, Conrad Leonard, Radhakrishnan Sabarinathan, Loris Mularoni, Scott Wood, Qinying Xu, Nick Waddell, Varsha Tembe, Gulietta M Pupo, Ricardo De Paoli-Iseppi, Ricardo E Vilain, Ping Shang, Loretta M S Lau, Rebecca A Dagg, Sarah-Jane Schramm, Antonia Pritchard, Ken Dutton-Regester, Felicity Newell, Anna Fitzgerald, Catherine A Shang, Sean M Grimmond, Hilda A Pickett, Jean Y Yang, Jonathan R Stretch, Andreas Behren, Richard F Kefford, Peter Hersey, Georgina V Long, Jonathan Cebon, Mark Shackleton, Andrew J Spillane, Robyn P M Saw, N ria L pez-Bigas, John V Pearson, John F Thompson, Richard A Scolyer, and Graham J Mann. Whole-genome landscapes of major melanoma subtypes. *Nature*, 545(7653):175–180, May 2017.

- [16] Boris C Bastian, Adam B Olshen, Philip E LeBoit, and Daniel Pinkel. Classifying melanocytic tumors based on DNA copy number changes. *Am. J. Pathol.*, 163(5): 1765–1770, November 2003.
- [17] W M Abdel-Rahman, K Katsura, W Rens, P A Gorman, D Sheer, D Bicknell, W F Bodmer, M J Arends, A H Wyllie, and P A Edwards. Spectral karyotyping suggests additional subsets of colorectal cancers characterized by pattern of chromosome rearrangement. *Proc. Natl. Acad. Sci. U. S. A.*, 98(5):2538–2543, February 2001.
- [18] Marwan Shinawi and Sau Wai Cheung. The array CGH and its clinical applications. *Drug Discov. Today*, 13(17-18):760–770, September 2008.
- [19] Cédric Le Caignec and Richard Redon. Copy number variation goes clinical. *Genome Biol.*, 10(1):301, January 2009.
- [20] Pawandeep Dhama, Alison J Coffey, Stephen Abbs, Joris R Vermeesch, Jan P Dumanski, Karen J Woodward, Robert M Andrews, Cordelia Langford, and David Vetrie. Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am. J. Hum. Genet.*, 76(5):750–762, May 2005.
- [21] Paul N Scriven. The scope, limitations and interpretation of copy number detection in the early embryo using the array CGH technique. *Hum. Reprod.*, 28(1):2–5, January 2013.
- [22] A E Oostlander, G A Meijer, and B Ylstra. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clin. Genet.*, 66(6):488–495, December 2004.
- [23] Ruibin Xi, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, Jianhua Zhang, Mark D Johnson, Donna M Muzny, David A Wheeler, Richard A Gibbs, Raju Kucherlapati, and Peter J Park. Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proc. Natl. Acad. Sci. U. S. A.*, 108(46):E1128–36, November 2011.
- [24] Ruibin Xi, Semin Lee, Yuchao Xia, Tae-Min Kim, and Peter J Park. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.*, 44(13):6274–6286, July 2016.
- [25] Peiyong Guan and Wing-Kin Sung. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods*, 102:36–49, June 2016.
- [26] Yi Li and Xiaohui Xie. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics*, 30(15):2121–2129, August 2014.

- [27] Marek Cmero, Ke Yuan, Cheng Soon Ong, Jan Schröder, PCAWG Evolution and Heterogeneity Working Group, Niall M Corcoran, Tony Papenfuss, Christopher M Hovens, Florian Markowetz, Geoff Macintyre, and PCAWG Consortium. Inferring structural variant cancer cell fraction. *Nat. Commun.*, 11(1):730, February 2020.
- [28] Eric Talevich, A Hunter Shain, Thomas Botton, and Boris C Bastian. CNVkit: Genome-Wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.*, 12(4):e1004873, April 2016.
- [29] Matthew Hayes and Jing Li. Bellerophon: a hybrid method for detecting interchromosomal rearrangements at base pair resolution using next-generation sequencing data. *BMC Bioinformatics*, 14 Suppl 5:S6, April 2013.
- [30] J R Lupski. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.*, 14(10):417–422, October 1998.
- [31] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragozy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.
- [32] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, 14(6):390–403, June 2013.
- [33] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. De novo assembly of the aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, April 2017.
- [34] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3D genome. *Nat. Rev. Genet.*, 19(7):453–467, July 2018.
- [35] Louise Harewood, Kamal Kishore, Matthew D Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V Peter Collins, and Peter Fraser. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.*, 18(1):125, June 2017.
- [36] Nicolas Servant, Nelle Varoquaux, Edith Heard, Emmanuel Barillot, and Jean-Philippe Vert. Effective normalization for copy number variation in Hi-C data. *BMC Bioinformatics*, 19(1):313, September 2018.

- [37] Enrique Vidal, François le Dily, Javier Quilez, Ralph Stadhouders, Yasmina Cuartero, Thomas Graf, Marc A Marti-Renom, Miguel Beato, and Guillaume J Filion. OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res.*, 46(8):e49, May 2018.
- [38] Ahmed Ibrahim Samir Khalil, Siti Rawaidah Binte Mohammad Muzaki, Anupam Chattopadhyay, and Amartya Sanyal. Identification and utilization of copy number information for correcting Hi-C contact map of cancer cell lines. *BMC Bioinformatics*, 21(1):506, November 2020.
- [39] Xiaotao Wang, Jie Xu, Baozhen Zhang, Ye Hou, Fan Song, Huijue Lyu, and Feng Yue. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods*, 18(6):661–668, June 2021.
- [40] Abhijit Chakraborty and Ferhat Ay. Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics*, 34(2):338–345, January 2018.
- [41] Jesse R Dixon, Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T Le, Galip Gürkan Yardımcı, Abhijit Chakraborty, Darrin V Bann, Yanli Wang, Royden Clark, Lijun Zhang, Hongbo Yang, Tingting Liu, Sriranga Iyyanki, Lin An, Christopher Pool, Takayo Sasaki, Juan Carlos Rivera-Mulia, Hakan Ozadam, Bryan R Lajoie, Rajinder Kaul, Michael Buckley, Kristen Lee, Morgan Diegel, Dubravka Pezic, Christina Ernst, Suzana Hadjur, Duncan T Odom, John A Stamatoyannopoulos, James R Broach, Ross C Hardison, Ferhat Ay, William Stafford Noble, Job Dekker, David M Gilbert, and Feng Yue. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.*, 50(10):1388–1398, October 2018.
- [42] Su Wang, Soohyun Lee, Chong Chu, Dhawal Jain, Peter Kerpedjiev, Geoffrey M Nelson, Jennifer M Walsh, Burak H Alver, and Peter J Park. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol.*, 21(1):73, March 2020.
- [43] Christopher J Troll, Nicholas H Putnam, Paul D Hartley, Brandon Rice, Marco Blanchette, Sameed Siddiqui, Javkhlan-Ochir Ganbat, Martin P Powers, Ramesh Ramakrishnan, Christian A Kunder, Carlos D Bustamante, James L Zehnder, Richard E Green, and Helio A Costa. Structural variation detection by proximity ligation from Formalin-Fixed, Paraffin-Embedded tumor tissue. *J. Mol. Diagn.*, 21(3):375–383, May 2019.
- [44] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.

- [45] Wenhan Chen, Alan J Robertson, Devika Ganesamoorthy, and Lachlan J M Coin. sCNPhase: using haplotype resolved read depth to genotype somatic copy number alterations from low cellularity aneuploid tumors. *Nucleic Acids Res.*, 45(5):e34, March 2017.
- [46] Cancer Center Membership. UCSF500 cancer gene panel. <https://cancer.ucsf.edu/research/molecular-oncology/ucsf500> Accessed: 2022-6-1.
- [47] A Hunter Shain, Nancy M Joseph, Richard Yu, Jamal Benhamida, Shanshan Liu, Tarl Prow, Beth Ruben, Jeffrey North, Laura Pincus, Iwei Yeh, Robert Judson, and Boris C Bastian. Genomic and transcriptomic analysis reveals incremental disruption of key signaling pathways during melanoma evolution. *Cancer Cell*, 34(1):45–55.e4, July 2018.
- [48] Kevin Hadi, Xiaotong Yao, Julie M Behr, Aditya Deshpande, Charalampos Xanthopoulos, Huasong Tian, Sarah Kudman, Joel Rosiene, Madison Darmofal, Joseph DeRose, Rick Mortensen, Emily M Adney, Alon Shaiber, Zoran Gajic, Michael Sigouros, Kenneth Eng, Jeremiah A Wala, Kazimierz O Wrzeszczyński, Kanika Arora, Minita Shah, Anne-Katrin Emde, Vanessa Felice, Mayu O Frank, Robert B Darnell, Mahmoud Ghandi, Franklin Huang, Sally Dewhurst, John Maciejowski, Titia de Lange, Jeremy Setton, Nadeem Riaz, Jorge S Reis-Filho, Simon Powell, David A Knowles, Ed Reznik, Bud Mishra, Rameen Beroukhim, Michael C Zody, Nicolas Robine, Kenji M Oman, Carissa A Sanchez, Mary K Kuhner, Lucian P Smith, Patricia C Galipeau, Thomas G Paulson, Brian J Reid, Xiaohong Li, David Wilkes, Andrea Sboner, Juan Miguel Mosquera, Olivier Elemento, and Marcin Imielinski. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell*, 183(1):197–210.e32, October 2020.
- [49] Pedram Gerami, Susan S Jewell, Larry E Morrison, Beth Blondin, John Schulz, Teresa Ruffalo, Paul Matushek, 4th, Mona Legator, Kristine Jacobson, Scott R Dalton, Susan Charzan, Nicholas A Kolaitis, Joan Guitart, Terakeith Lertsbarapa, Susan Boone, Philip E LeBoit, and Boris C Bastian. Fluorescence in situ hybridization (FISH) as an ancillary diagnostic tool in the diagnosis of melanoma. *Am. J. Surg. Pathol.*, 33(8): 1146–1156, August 2009.
- [50] Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Paraffin embedding tissue samples for sectioning. *CSH Protoc.*, 2008:db.prot4989, May 2008.
- [51] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.
- [52] Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas S P Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a One-Click system for analyzing Loop-Resolution Hi-C experiments. *Cell Syst*, 3(1):95–98, July 2016.

- [53] Neva C Durand, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst*, 3(1):99–101, July 2016.
- [54] Atefeh Taherian Fard and Mark A Ragan. Quantitative modelling of the waddington epigenetic landscape. *Methods Mol. Biol.*, 1975:157–171, 2019.
- [55] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.
- [56] Matthew D VerMilyea, Laura P O’Neill, and Bryan M Turner. Transcription-independent heritability of induced histone modifications in the mouse preimplantation embryo. *PLoS One*, 4(6):e6086, June 2009.
- [57] Y Zhang and D Reinberg. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.*, 15(18):2343–2360, September 2001.
- [58] T Jenuwein and C D Allis. Translating the histone code. *Science*, 293(5532):1074–1080, August 2001.
- [59] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.
- [60] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfennig, Xinchun Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shores, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, Ginell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthal, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos,



- Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.
- [61] Albert J Keung, J Keith Joung, Ahmad S Khalil, and James J Collins. Chromatin regulation at the frontier of synthetic biology. *Nat. Rev. Genet.*, 16(3):159–171, March 2015.
- [62] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.
- [63] Daniel Holoch and Danesh Moazed. RNA-mediated epigenetic regulation of gene expression. *Nat. Rev. Genet.*, 16(2):71–84, February 2015.
- [64] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, 107(7):2926–2931, February 2010.
- [65] Bruce Stillman. Histone modifications: Insights into their influence on gene expression. *Cell*, 175(1):6–9, September 2018.
- [66] Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, September 2016.
- [67] Navneet Matharu and Nadav Ahituv. Modulating gene regulation to treat genetic disorders. *Nat. Rev. Drug Discov.*, 19(11):757–775, November 2020.
- [68] V Swaminathan, B A Ashok Reddy, B Ruthrotha Selvi, M S Sukanya, and Tapas K Kundu. Small molecule modulators in epigenetics: implications in gene expression and therapeutics. *Subcell. Biochem.*, 41:397–428, 2007.
- [69] Isaac B Hilton, Anthony M D’Ippolito, Christopher M Vockley, Pratiksha I Thakore, Gregory E Crawford, Timothy E Reddy, and Charles A Gersbach. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.*, 33(5):510–517, May 2015.
- [70] Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K Polansky, Peter Ebert, Karl Nordström, Matthias Barann, Anupam Sinha, Sebastian Fröhler, Jieyi Xiong, Azim Dehghani Amirabad, Fatemeh Behjati Ardakani, Barbara Hutter, Gideon Zipprich, Bärbel Felder, Jürgen Eils, Benedikt Brors, Wei Chen, Jan G Hengstler, Alf Hamann, Thomas Lengauer, Philip Rosenstiel, Jörn Walter, and Marcel H Schulz. Combining transcription factor binding affinities with

- open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.*, 45(1):54–66, January 2017.
- [71] Florian Schmidt, Fabian Kern, and Marcel H Schulz. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics Chromatin*, 13(1):4, February 2020.
- [72] Arshdeep Sekhon, Ritambhara Singh, and Yanjun Qi. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics*, 34(17):i891–i900, September 2018.
- [73] Fabrizio Frasca, Matteo Matteucci, Michele Leone, Marco J Morelli, and Marco Masseroli. Accurate and highly interpretable prediction of gene expression from histone modifications. *BMC Bioinformatics*, 23(1):151, April 2022.
- [74] Huan Zhong, Soyeon Kim, Degui Zhi, and Xiangqin Cui. Predicting gene expression using DNA methylation in three human populations. *PeerJ*, 7:e6757, May 2019.
- [75] Hiroshi Kimura. Histone modifications for human epigenome analysis. *J. Hum. Genet.*, 58(7):439–445, July 2013.
- [76] Jacob Schreiber, Jeffrey Bilmes, and William Stafford Noble. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biol.*, 21(1):82, March 2020.
- [77] Guanjue Xiang, Cheryl A Keller, Belinda Giardine, Lin An, Qunhua Li, Yu Zhang, and Ross C Hardison. S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. *Nucleic Acids Res.*, 48(8):e43, May 2020.
- [78] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, 18(10):1196–1203, October 2021.
- [79] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying chip-seq enrichment using macs. *Nature protocols*, 7(9):1728–1740, 2012.
- [80] Jacob Schreiber, Timothy Durham, Jeffrey Bilmes, and William Stafford Noble. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome biology*, 21(1):1–18, 2020.
- [81] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

- [82] Michael JD Powell. The convergence of variable metric methods for nonlinearly constrained optimization calculations. In *Nonlinear programming 3*, pages 27–63. Elsevier, 1978.
- [83] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.
- [84] Thomas C Evans and Craig P Hunter. Translational control of maternal rnas. *Worm-Book: The Online Review of C. elegans Biology [Internet]*, 2005.
- [85] Lucy W Barrett, Sue Fletcher, and Steve D Wilton. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and molecular life sciences*, 69(21):3613–3634, 2012.
- [86] Sarah F Mitchell, Sarah E Walker, Mikkel A Algire, Eun-Hee Park, Alan G Hinnebusch, and Jon R Lorsch. The 5′-7-methylguanosine cap on eukaryotic mrnas serves both to stimulate canonical translation initiation and to block an alternative pathway. *Molecular cell*, 39(6):950–962, 2010.
- [87] Marilyn Kozak. Point mutations define a sequence flanking the aug initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, 44(2):283–292, 1986.
- [88] Pieter Spealman, Armaghan W Naik, Gemma E May, Scott Kuersten, Lindsay Freeberg, Robert F Murphy, and Joel McManus. Conserved non-aug uorfs revealed by a novel regression analysis of ribosome profiling data. *Genome research*, 28(2):214–222, 2018.
- [89] Alexey A Gritsenko, Shira Weingarten-Gabbay, Shani Elias-Kirma, Ronit Nir, Dick De Ridder, and Eran Segal. Sequence features of viral and human internal ribosome entry sites predictive of their activity. *PLoS computational biology*, 13(9):e1005734, 2017.
- [90] Daniela Rhodes and Hans J Lipps. G-quadruplexes and their regulatory roles in biology. *Nucleic acids research*, 43(18):8627–8637, 2015.
- [91] Shlomi Dvir, Lars Velten, Eilon Sharon, Danny Zeevi, Lucas B Carey, Adina Weinberger, and Eran Segal. Deciphering the rules by which 5′-utr sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences*, 110(30):E2792–E2801, 2013.
- [92] John WB Hershey, Nahum Sonenberg, and Michael B Mathews. Principles of translational control: an overview. *Cold Spring Harbor perspectives in biology*, 4(12):a0111528, 2012.

- [93] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223, 2009.
- [94] Hung-Hsi Chen and Woan-Yuh Tarn. uorf-mediated translational control: recently elucidated mechanisms and implications in cancer. *RNA biology*, 16(10):1327–1338, 2019.
- [95] Sai Zhang, Hailin Hu, Tao Jiang, Lei Zhang, and Jianyang Zeng. Titer: predicting translation initiation sites by deep learning. *Bioinformatics*, 33(14):i234–i242, 2017.
- [96] Jim Clauwaert, Zahra McVey, Ramneek Gupta, and Gerben Menschaert. Tis transformer: Re-annotation of the human proteome using deep learning. *bioRxiv*, 2021.
- [97] Jonathan M Mudge, Jorge Ruiz-Orera, John R Prensner, Marie A Brunet, Ferriol Calvet, Irwin Jungreis, Jose Manuel Gonzalez, Michele Magrane, Thomas F Martinez, Jana Felicitas Schulz, et al. Standardized annotation of translated open reading frames. *Nature Biotechnology*, pages 1–6, 2022.
- [98] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.
- [99] Nicola Whiffin, Konrad J Karczewski, Xiaolei Zhang, Sonia Chothani, Miriam J Smith, D Gareth Evans, Angharad M Roberts, Nicholas M Quaife, Sebastian Schafer, Owen Rackham, et al. Characterising the loss-of-function impact of 5′ untranslated region variants in 15,708 individuals. *Nature communications*, 11(1):1–12, 2020.
- [100] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.
- [101] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):1–14, 2016.
- [102] Xiaolei Zhang, Matthew Wakeling, James Ware, and Nicola Whiffin. Annotating high-impact 5′ untranslated region variants with the utrannotator. *Bioinformatics*, 37(8):1171–1173, 2021.

# Appendix A

## HiDENSEC Supplementary Information

HiDENSEC broadly proceeds by (i) correcting Hi-C counts for confounding covariates, (ii) using these corrected counts to infer mixture proportion and copy number, and (iii) identifying large structural variants based on the resulting profiles. Before describing each of these steps in detail, and in order to make equations easier to understand, we begin by fixing the notation and introducing general conventions that will be used throughout.

### A.1 Notation & Setup

Our starting point is a raw nucleotide-level Hi-C matrix  $H^{\text{raw}}$  obtained by mapping Hi-C or Fix-C reads to a reference genome (**Online Methods**). This matrix is symmetric, i.e.,  $H_{ij}^{\text{raw}} = H_{ji}^{\text{raw}}$ .

In order to balance signal-to-noise ratios and computational burden with interpretability and localization of changes in copy number and/or rearrangement breakpoints/junctions, we first coarse grain this nucleotide-level Hi-C matrix  $H^{\text{raw}}$  to genomic windows of length  $w = 50\text{kb}$ :

$$H_{ij} = \sum_{\substack{m=i, \dots, i+w \\ n=j, \dots, j+w}} H_{mn}^{\text{raw}},$$

The indices of this matrix label genomic windows. All further steps in HiDENSEC work with such a coarse-grained matrix.

For a completely homogeneous cell population, we can model  $H$  as a random matrix of the form  $H = Nh + \varepsilon_N$ , where  $N$  is the total number of cells involved in the Hi-C experiment,  $h$  represents the (deterministic) interactions between genomic loci in cells of a given type (and so in particular,  $h$  may differ across different cell populations), and  $\varepsilon_N$  is an error matrix, on which we do not impose any restrictions other than  $\lim_{N \rightarrow \infty} \varepsilon_N/N = 0$  (in probability). This assumption ensures the identifiability of  $h$ , and is expected under the very plausible assumption that the numbers of reads contributed by different cells in the experiment are independent).

Hi-C experiments are generally conducted on samples containing mixtures of cells with different genomes. We denote the set of genomes by  $\mathcal{G}$ , and represent the cellular mixture fractions by  $(f^G)_{G \in \mathcal{G}}$ . Since chromatin contacts only occur between genomic segments within the same nucleus, the Hi-C contact map of the sample is expected to be a simple superposition of these distinct Hi-C signatures. That is,

$$h_{ij} = \sum_{G \in \mathcal{G}} f^G h_{ij}^G,$$

where  $h^G$  is the  $h$  matrix associated with genome  $G$ . To test the superposition assumption we confirmed the lack of Hi-C contacts between neighboring cells by performing Fix-C on a 1:1 mixture of human and mouse cells, finding negligible ligation products that include both mouse and human sequence (**Supplementary Table 1**).

Each genome  $G$  will have an absolute copy number profile  $p_i^G$  at each site  $i$  representing the local, integer-valued, copy number at that locus. In a superposition of multiple genomes, however, we more directly measure the weighted average of these copy numbers,  $\pi_i = \sum_{G \in \mathcal{G}} f^G p_i^G$ , which we refer to as the effective copy number  $\pi_i$  for the sample. If we knew the contact maps  $h_{ij}^G$  in isolation for each genome  $G \in \mathcal{G}$ , then we should be able to infer the mixture proportions  $f^G$ , for  $G \in \mathcal{G}$ . With these in hand, we can also interpret rearrangements within each genome, which are represented by  $p_{ij}^G$ , the absolute number of site  $i$  copies that are in contact with site  $j$  in genome  $G$  due to translocations and other rearrangements. To keep formulas and descriptions compact, we will occasionally refer to  $p_i$  as  $p_{ii}$ .

In analyzing the Hi-C matrices discussed in this paper, we find that it is typically sufficient to include one or sometimes two cancer genotypes along with the diploid reference genome (which is free of copy number and other structural variation). Thus  $|\mathcal{G}| = 2$  for most samples, except sample 1 from patient 3 which requires  $|\mathcal{G}| = 3$ . When  $|\mathcal{G}| = 2$  there is no risk of confusing distinct genomes, and so for notational simplicity we drop the superscripts and refer to the interaction strengths, mixture proportion and copy numbers of the non-reference genome as  $h$ ,  $f$  and  $p$ , respectively.

The goal of HiDENSEC is then simply stated: Given experimentally observed Hi-C counts  $H$  from a composite sample, we wish to infer the mixture proportions  $f$ , copy number profile  $p = (p_i)$ , and identify the set of non-adjacent structural variants  $\{(i, j) : p_{ij} > 0\}$  together with the precise nature of these structural changes; that is, determine the values of  $p_{ij}$  (whenever identifiable).

## A.2 Covariate correction

As described in the **Online Methods**, there are several biological and technical effects which can substantially affect read counts (see **Supplementary Figure A.1A**, and **Supplementary Figure A.1B**). To begin to account for these covariates we focus on the "diagonal" local contact  $h_{ii}$  and model its dependence on the true contacts  $p_i$  as

$$h_{ii} = C_0 p_i g(x_i^0, x_i^1, x_i^2, x_i^3), \quad (\text{A.1})$$

for a function  $g$  to be determined, where  $x^0$  through  $x^2$  encode the numerical, experiment-related covariates (e.g., GC content, cut-site density, and read mappability), and  $x^3$  is a discrete indicator of one of the six chromatin compartments  $\mathcal{C} = \{A0, A1, A2, B1, B2, B3\}$ .

Two observations inform the form of  $g$ :

1. The dependence between Hi-C contact intensities and covariates appears to be sensitive to protocol differences in the experiment. Although *in-situ* Hi-C based covariate structures tend to look very similar to **Supplementary Figure A.1**, those obtained from Fix-C experiments usually look like the plots depicted in **Supplementary Figure A.2**.
2. Even though the details of this effect may differ across protocols, their qualitative shape appears close to linear for  $x^0$  and  $x^1$ , with  $x^2$  exhibiting a cut-off phenomenon (see panels B of **Supplementary Figure A.1** and **Supplementary Figure A.2**) without substantial interaction (see panels C of the same figures). Constructing  $g$  using these qualitative trends leads to satisfactory model fit (see the subsequent paragraph).

We therefore filter genomic bins, retaining only those for which  $x^2 > 0.8$ , and adopt a simple descriptive linear regression model, with different regression line for each type of compartment for  $x^0$  and  $x^1$ . That is,  $g$  is fit as

$$g(x^0, x^1, x^2, x^3) = \sum_{c \in \mathcal{C}} \mathbb{1}_c(x^3) g_c(x^0, x^1), \quad (\text{A.2})$$

where  $g_c$  is linear for all  $c \in \mathcal{C}$ . For purely diploid genomes this model explains 80-90% of the observed variance, satisfying standard model-fit criteria (normality and identical distribution of studentized residuals, independence of residuals and predictions, etc.)

For simplicity we also assume that covariate corrections Eq. (A.1) and Eq. (A.2) derived for karyotypically normal genomes can also be applied to cancer genome (that is, as a first approximation we neglect changes in compartment structure in the cancer genomes). By the superposition principle it follows that Eq. (A.1) and Eq. (A.2) also hold with  $p_i$  replaced by  $\pi_i$ . Such covariate correction was performed for all Hi-C maps used in the main manuscript, based on  $g$  fit to Fix-C from a karyotypically normal reference, corresponding to Sample 1 - I. An illustration of both the need for adjusting raw counts, as well as the protocol sensitivity of any such adjustment is shown in **Supplementary Figure A.11**.

In principle, one could attempt a similar correction procedure on the off-diagonal entries  $h_{ij}, i \neq j$ , with a corrector function  $g' : (\mathcal{C} \times \mathbb{R}^3)^2 \rightarrow \mathbb{R}$  that depends on pairs of covariates. However, as indicated in the main text, the magnitude of off-diagonal signals are difficult to interpret due to uncertainty in the precise break-point within a 50 kb window, and the possibility of substantially altered compartment structure associated with large-scale structural variants (**Supplementary Figure A.4**). Nevertheless, these off-diagonal signals can clearly distinguish  $h_{ij} > 0$  from  $h_{ij} = 0$ . For the precise inference of large-scale structural variants and their copy numbers, this type of on-or-off signal is sufficient as will be illustrated in Appendix (A.4).

If a reference contact map is not available, or the underlying experimental protocol is unknown, then an internal covariate correction is still possible based on read counts reliably identified to correspond to  $\pi_{\text{mode}}$  (see section below). Such correction empirically performs competitively with the procedure described above, and in HiDENSEC is resorted to automatically if no explicit reference protocol and/or data set is specified.

### A.3 Inference of copy numbers & mixture proportions

Given the covariate correction as described in Appendix [A.2](#) the resulting corrected read counts  $H_{ii}$  are effectively modeled as  $H_{ii} = C_0 N \pi_i + \varepsilon_{ii}$  as mentioned in Appendix [A.1](#) (see also **Supplementary Figure A.12A** for an illustration of the underlying generative model).

First we note that, in the general case (arbitrary numbers and forms of cancer genomes), we cannot infer  $\pi$  from  $H$  as a matter of principle, even in the limit of infinite data, without further assumptions:

1. *Relative copy number profiles determine absolute copy number profiles only up to integer scaling.* Since  $\pi = \sum_{g \in \mathcal{G}} f^G p^G$  is only observed up to an overall factor, any two sets of copy number vectors  $p$  and  $p'$  that differ by a multiplicative constant (that is,  $p' = Kp$  for some  $K \in \mathbb{N}$ ) are indistinguishable on the level of  $H$ , since  $C_0$  and  $N$  are generally unknown. E.g., a completely diploid and a completely triploid genome are indistinguishable based on their relative copy numbers. Indeed, without explicit knowledge of  $C$  (which depends, among other factors, on the number of cells involved in the experiment, and so is typically difficult to obtain), any absolute copy number profile consistent with the Hi-C map can be scaled by an integer and remains consistent.
2. *Absolute copy numbers only involve the products of mixture proportions and copy numbers, not either of them individually.* Since  $\pi$  involves products of proportions and copy numbers, an increase in one can often be compensated by a decrease in the other without affecting even absolute values  $\pi$  (i.e., this type of unidentifiability is independent of the scaling factor  $C_0 N$ ). More concretely, for any genome  $G_0 \in \mathcal{G}$ ,

$$\begin{aligned} \pi &= \sum_{G \in \mathcal{G}} f^G p^G = f^{G_0} p^{G_0} + \sum_{G \in \mathcal{G}_{\bar{0}}} f^G p^G = \left( 1 - \sum_{G \in \mathcal{G}_{\bar{0}}} f^G \right) p^{G_0} + \sum_{G \in \mathcal{G}_{\bar{0}}} f^G (p^{G_0} + p^G - p^{G_0}) \\ &= p^{G_0} + \sum_{G \in \mathcal{G}_{\bar{0}}} f^G (p^G - p^{G_0}) = p^{G_0} + \sum_{G \in \mathcal{G}_{\bar{0}}} \left( \frac{f^G}{K^G} \right) \cdot K^G (p^G - p^{G_0}), \end{aligned}$$

where  $\mathcal{G}_{\bar{0}} = \mathcal{G} \setminus \{G_0\}$  and  $K^G \in \mathbb{N}$  are genome-specific constants. That is, even if the copy number profile for a single genome  $G_0$  is completely known (in, e.g., most samples not derived from cell lines, it is reasonable to assume the presence of the purely diploid reference genome), the relative difference of all other genomes to  $G_0$  are, in



the most general setting, not determined by  $\pi$ . For the simplest case of two distinct cell populations, one purely diploid and the other with absolute copy number profile  $p$ , mixed at proportions  $1 - f$  and  $f$ , respectively, the effective copy number profile reads

$$\pi_i = 2(1 - f) + fp_i = 2 + (p_i - 2)f.$$

That is, given the exact absolute copy number, it is only possible to infer the product  $(p_i - 2)f$ , and not either of them individually.

Given these fundamental limitations we must impose additional suitably restrictive, yet biologically plausible, assumptions. HiDENSEC does so by positing that:

- (a) The most common effective copy number is known in advance; i.e., one has an estimate of

$$\pi_{\text{mode}} = \tau \in \bigcup_i \pi_i \# \{i : \pi_i = \tau\}.$$

Where  $\# \{i : \pi_i = \tau\}$  is the number of bins with effective copy number  $\tau$ . This assumption allows rescaling of the  $H_i$  to correct for the factor of  $C_0N$ . While we could use mean or median of  $\pi$  for this purpose the mode of  $\pi_{\text{mode}}$  is particularly appealing in that

1. unless the genomes are extraordinarily complex,  $\pi_{\text{mode}}$  will be 2, and
  2.  $\# \{i : \pi_i = \pi_{\text{mode}}\}$  is often large, so that rescaling is effectively based on averaging many noisy observations, which in general outperforms estimation based on a single observation (as is done when only considering  $\pi_1$ ).
- (b) Copy number profiles are as close to purely diploid as is consistent with the data, in that  $\max_{i,G} p_i^G$  is chosen as small as possible. For instance, consider the hypothetical example of  $\mathcal{G} = \{G_0, G_1\}$ ,  $p^{G_0} \equiv 2$ ,  $\pi_i = 2 \cdot \mathbb{1}_{\{1, \dots, i^*\}}(i) + 2.5 \cdot \mathbb{1}_{\{i^*+1, \dots\}}(i)$ , HiDENSEC will estimate  $f^{G_0} = 0.5$ ,  $f^{G_1} = 0.5$ ,  $p^{G_0} \equiv 2$ ,  $p_i^{G_1} = 2 \cdot \mathbb{1}_{\{1, \dots, i^*\}}(i) + 3 \cdot \mathbb{1}_{\{i^*+1, \dots\}}(i)$  rather than, say,  $f^{G_0} = 0.9$ ,  $f^{G_1} = 0.1$ ,  $p^{G_0} \equiv 2$ ,  $p_i^{G_1} = 2 \cdot \mathbb{1}_{\{1, \dots, i^*\}}(i) + 7 \cdot \mathbb{1}_{\{i^*+1, \dots\}}(i)$ , which is consistent with the principle of parsimony. Prior knowledge favouring a non-parsimonious solution can be explicitly fed into HiDENSEC as an optional argument if desired.

Although these assumptions narrow down the feasible solutions substantially, it can be shown that, in the most general setting, obtaining a solution is computationally intractable:

Identifying the smallest number of genomes  $|\mathcal{G}|$  that explain a given noise-less effective copy number profile  $\pi$  using mixture proportions bounded away from zero (e.g.,  $\min_{G \in \mathcal{G}} f^G \geq o(|\mathcal{G}|^{-1})$ ) and bounded absolute copy numbers (i.e.,  $\max_{G \in \mathcal{G}} \|p^G\|_\infty \leq B$  for some  $B \in \mathbb{N}$ ) is, in general, at least as hard as the subset sum problem, and therefore NP-complete.

Even though its proof does not immediately inform inference (and is therefore deferred to Appendix [A.5](#)), this theorem suggests that any feasible inference procedure must be based on either

1. assumptions that are strong enough to render the subset sum problem efficiently solvable, yet are still biologically plausible, or
2. approximate inference which may work well for few genomes (e.g.,  $|\mathcal{G}|$  is small) but may become inaccurate as  $|\mathcal{G}|$  grows large.

Given that empirically  $|\mathcal{G}| = 2$  appears to typically explain the data well, with  $|\mathcal{G}| > 3$  rarely being required (indeed, none of the cases described in the main text requires more than three genomes, HiDENSEC assumes a modest number of cancer genomes, and proceeds as follows:

1. Normalize the data  $\{H_{ii}\}_i$  by  $\hat{H}_{\text{mode}}/\pi_{\text{mode}}$ , where  $\hat{H}_{\text{mode}}$  is an estimate of the mode of  $\{H_{ii}\}_i$ . Data is typically abundant enough that obtaining  $\hat{H}_{\text{mode}}$  through either a kernel density estimate or a simple histogram is sufficient. This normalized data is referred to as  $\{\Pi_i\}_i$ .
2. For a fixed window size  $w$  (in the main text analysis  $w = 100$ ), choice of  $f$ , and candidate copy number  $p$ , we define  $\rho(f, p) = (1 - f)2 + fp$  and

$$m_x(f, p) = \frac{1}{2w + 1} \sum_{i=x-w}^{x+w} |\Pi_i - \rho(f, p)| \cdot [p \cdot \text{md}(\Pi_{[x-w, x+w]})]^{-1},$$

where  $\text{md}(X)$  is the median deviation of a set of numbers  $X$ , and the normalization involving it is motivated by the heteroskedasticity observed in contact-intensity counts (see **Supplementary Figure A.12B**). For a choice of maximum copy number  $p_{\text{max}}$ , compute a first estimated copy number profile  $\hat{p}_x^1$  and associated mixture proportion  $\hat{f}^1$  as the minimizers of  $m_x(f, p)$  aggregated over the entire genome

$$\hat{f}^1 = f \sum_x \left[ \min_{p \in [p_{\text{max}}]} m_x(f, p) \right] \quad \hat{p}_x^1 = \underset{p \in [p_{\text{max}}]}{\text{argmin}} m_x(\hat{f}^1, p),$$

and estimate the corresponding effective copy number profile  $\hat{\pi}^1$  as  $\hat{\pi}^1 = \hat{f}^1 \hat{p}^1$ . Estimation based on  $m_x(f, p)$  in this manner exploits the strong spatial correlation present in  $\Pi$ , while otherwise remaining fully non-parametric.

3. Refine  $\hat{\pi}_x^1$  by adjusting points of copy number changes, measuring their significance, and fine-tune  $p_{\text{max}}$  (see section below). Call this refined profile  $\{\hat{\pi}_x^1\}_x$  as well.
4. If  $\hat{\pi}^1 \equiv 2$ , then return  $\hat{\pi}^0 \equiv 2$  and  $\hat{f}^0 = 0$ .
5. Otherwise repeat steps 2-5 on the corrected effective copy number profile  $\Pi^1 = \Pi - \hat{\pi}^1$  until  $\hat{\pi}^K \equiv 0$ , and then return  $\hat{f}^{\mathcal{G}} = \bigcup_{k \in [K]} \hat{f}^k$ ,  $\hat{p}^{\mathcal{G}} = \bigcup_{k \in [K]} \hat{p}^k$  and  $\hat{\pi} = \sum_{k \in [K]} \hat{f}^k \hat{p}^k$ .

In this way, HiDENSEC attempts to greedily explain the shape of  $\Pi$  by subtracting the effect of an individual genome one at a time. It can be shown that this greedy procedure accurately recovers ground-truth  $f^{\mathcal{G}}$  and  $p^{\mathcal{G}}$  in the limit of noiseless data if the following conditions are met:

(a)

$$\left| \text{supp } p^{G_k} \setminus \bigcup_{i=k+1}^K \text{supp } p^{G_i} \right| \geq \left| \bigcup_{i=k+1}^K \text{supp } p^{G_i} \right|,$$

for all  $k \in [K]$ , where  $\text{supp } p = \{i : p_i \neq 2\}$ , and the  $G_k$  are ordered such that  $f^{G_1} \geq f^{G_2} \geq \dots \geq f^{G_K}$ .

(b)

$$2 \left\| \rho(f^{G_{k+1}}, \dots, f^{G_K}, p^{G_{k+1}}, \dots, p^{G_K}) \right\|_{\infty} \leq f^{G_k},$$

also for any  $k \in [K]$ , where by slight overloading of notation,  $\rho(f_1, \dots, f_r, p_1, \dots, p_r) = 2(1 - \sum_{k=1}^r f_k) + \sum_{k=1}^r f_k p_k$ .

(We note that if these conditions are not met, then the results of HiDENSEC may not be accurate.)

It is clear that these conditions tend to more easily be met if  $K$  is small, while they become more restrictive as  $K$  grows. Intuitively, they stipulate that genomes of more abundant cell populations should exhibit more substantial copy number changes than those of rare populations, and that mixture proportions be far away from uniformity; which—given the nature of logistic growth, and the fact that more abundant cell types likely had more time to evolve—appear biologically plausible. Indeed, these assumptions are satisfied in all samples analyzed in the main text (the majority of which carries  $K = 1$  cell population in addition to the reference genome, in which case these conditions are trivially satisfied).

## Refining effective copy number profiles

Due to the randomness inherent in chromatin folding and Hi-C experiments, the procedure described in step 2 above will occasionally detect copy number changes that are either imprecisely located or purely the result of stochastic fluctuation rather than biologically meaningful structure, and so it is desirable to correct for such misinference. HiDENSEC does so in various ways.

1. *Refining change points:* A site  $x$  at which copy numbers change is characterized by the fact that  $\mathbb{E}\Pi_{x-\delta} = \pi_{x-\delta} \neq \pi_{x+\delta} = \mathbb{E}\Pi_{x+\delta}$ , for any sufficiently small  $\delta$ . If this gap between  $\pi$  left and right of  $x$  is sufficiently large compared to the variance of the data, then it is reasonable to assume that

$$x =_{j \in \mathcal{N}(x)} \mathbb{E} \text{Var} \left[ \Pi_{\sigma} \mid \mathbb{1}_{[-\infty, j-1]}(\sigma) \right], \quad (\text{A.3})$$

where  $\mathcal{N}(x)$  is a suitably small neighbourhood around  $x$ , and  $\sigma \sim \text{Uniform}(\mathcal{N}(x))$  is a uniform draw from  $\mathcal{N}(x)$ . That is, once likely change point candidates have been identified in step 2 above, their precise location can be refined by choosing suitable neighbourhoods  $\mathcal{N}$  around them, and optimizing Eq. (A.3) accordingly (this essentially corresponds to fitting a depth-1 decision tree regressing sites in  $\mathcal{N}$  against  $\Pi$ ). The resulting refined change points are further adjusted or shifted to ensure that corresponding excursions (see below) do not cross chromosome boundaries.

2. *Interpreting change points:* The optimization procedure described above will return a refined choice of  $x$  even if  $\mathcal{N}$  does not undergo any copy number change, and so it is of interest to quantify the extent to which  $x$  separates copy number levels. To do so, HiDENSEC assesses significance by performing 100 replicates of a permutation test on  $\mathcal{N}$ , and computing the  $p$ -value  $p_{\mathcal{N}}(x)$  of  $\text{Var}[\Pi \mid \sigma_x]$  on the resulting empirical distribution. It should be noted that a priori it is unclear whether  $p$ -values calculated in such manner are well-calibrated even in the limit of large  $\mathcal{N}$  (indeed, they should instead be formed based on the null distribution of counts around  $\pi = 2$  conditional on  $\hat{\pi} \neq 2$ ; which, however, is not accessible), but they do behave super-uniformly empirically (see **Supplementary Figure A.12C**).
3. *Interpreting excursions:* An excursion  $e$  of an effective copy number profile  $\hat{\pi}$  is defined to be any tuple  $e = (x_1, x_2, \hat{\pi}_{x_1+1})$  for two adjacent change points  $x_1$  and  $x_2$ , for which  $\hat{\pi}_{x_1+1} \neq 2$ .  $e$  is likely to be reflective of actual biological signal if  $\max_{x \in \{x_1, x_2\}} p_{\mathcal{N}}(x)$  is small, if the length  $x_2 - x_1$  of  $e$  is large, and if the aggregate read counts on  $[x_1, x_2]$  are broadly no more variable than expected for level  $\hat{\pi}_{x_1+1}$  (if they are significantly more variable, then the change in effective copy number is prone to being merely a result of fluctuation). HiDENSEC thus assigns a significance to each excursion  $e = (x_1, x_2, \pi)$  by incorporating the two  $p$ -values of both  $x_1$  and  $x_2$ , one  $p$ -value associated with  $x_2 - x_1$  based on a reference diploid genome, as well as one calculated from the median deviation of  $\Pi$  on  $e$  in relation to appropriately re-scaled diploid  $\Pi$  values in the vicinity of  $e$  (since read count fluctuations generally exhibit spatial dependence, with stochasticity increasing in smaller chromosomes, it is preferable to construct local empirical null distributions over global ones). Under  $\mathcal{H}_0$ ,  $p$ -values computed in such a manner on a given set  $\mathcal{E}$  of excursions behave broadly uniformly (see **Supplementary Figure A.12C**).
4. *Model selection:* To assess whether  $\hat{\pi}$  likely captures true copy number variation, or simply overfits to a noisy  $\pi \equiv 2$  profile, various empirically well-performing checks are in place. More concretely, a  $\hat{\pi}$  instance is declared overfitting (and whence adjusted to  $\hat{\pi} \equiv 2$ ) if it clears any three of the following criteria:
  - *Inferred mixture proportion*  $\hat{f} < 0.15$ . The amplitude  $A$  of a length- $\ell$  excursion that is purely due to stochastic fluctuations decays broadly as  $O(e^{-A^2\ell})$ , and so detecting excursions consistent with large  $f$  is unlikely under  $\mathcal{H}_0$ .

- $\text{md } \Pi \geq \hat{f}$ . Small mixture proportions are only reliably attributable to biological signal if the fluctuations in  $\Pi$  are of smaller order.
  - *Number of excursions*  $\geq 60$ . Under  $\mathcal{H}_0$ , small-amplitude excursions are typically frequent.
  - $|\{e = (x_1, x_2, \pi_e) \in \mathcal{E} : x_2 - x_1 \leq 200\}| / |\mathcal{E}| > 1/2$ . Under  $\mathcal{H}_0$ , lengths of excursions decay exponentially, and so most observed excursions ought to be short.
  - *The Benjamini-Hochberg threshold calculated on  $p$ -values computed in step 3 above calls less than 10% of  $\mathcal{E}$  significant at  $\alpha = 0.25$* . Under  $\mathcal{H}_1$ ,  $p$ -values tend to be strongly significant.
  - *Fluctuation strength is not monotonically increasing with copy number*. Under  $\mathcal{H}_1$ , larger copy numbers are associated with larger fluctuations.
  - $\#\{x : \hat{\pi}_x = \pi_{\text{mode}}\}$  makes up less than  $\varphi\%$  of all  $x$ , where  $\varphi$  is by default set to 50, but can be adapted based on prior knowledge. Overfitting will lead to erroneous excursions away from  $\pi_{\text{mode}}$ .
5. *Utilizing off-diagonal information*: The presence or absence of off-diagonal signal at either boundary of an excursion  $e$  provides further evidence for the biological significance of  $e$ . Thus, uncertainty quantification of off-diagonal intensities (discussed in the following section) is incorporated into the overall computation of a  $p$ -value associated with an excursion  $e$ .

Illustrations of effective copy number profiles called in this manner on the data sets analyzed in the main text are given in **Supplementary Figure A.13**.

## Confidence intervals for mixture proportions $\hat{f}^G$

There are primarily two sources of noise that contribute to uncertainty in the proportion estimates  $\hat{f}^G$ :

1. The stochasticity of read counts conditional on  $h$ ; that is  $\varepsilon$ .
2. Shifts in the expected intensities  $h$  themselves, due to, e.g., uncaptured covariates.

The former is a commonly encountered complication in statistical inference and can be addressed by classical non-parametric tools like the bootstrap (or versions thereof; e.g., the block or sieve bootstrap to account for the lack of independence and identical distributions in  $\varepsilon$ ; it is also this lack of regularity in  $\varepsilon$  that prevents exploiting more explicit tools based on central limit arguments or semi-parametric assumptions), while the latter is more delicate: It includes systematic biases in the data that may be unique to  $\mathcal{G}$  (and therefore  $h$ ) itself, and therefore can be difficult to estimate. For instance, for a (ground-truth) effective copy number profile  $\pi$  featuring two excursions  $e_1$  and  $e_2$  at levels  $\pi_{e_1} = 2 + \phi - \delta$ ,  $\pi_{e_2} = 2 + \phi + \delta$  (for some  $\phi \in [0, 1]$  and small  $\delta > 0$ ), both of identical length, it is reasonable to either infer

$|\hat{\mathcal{G}}| = 3, \hat{f}^1 = \phi, \hat{f}^2 = \delta$ , or  $|\hat{\mathcal{G}}| = 2, \hat{f} = \phi$  and attribute the shifts by  $\delta$  to systematic biases that have not been captured by the covariate correction described above. HiDENSEC will decide between these two situations based on the fluctuations around these effective copy number values both inside and outside of these excursions, but if the latter is returned, then the difference of  $\delta$  ought to be reflected in any uncertainty quantification of  $\hat{f}^G$ . To do so, HiDENSEC performs the following bootstrap procedure, estimating confidence intervals for each  $f^k$  in turn, and decoupling each excursion  $e \in \mathcal{E}$ :

1. For each  $e \in \mathcal{E}$  with  $e = (x_e, y_e, \hat{\pi}_e = 2\hat{f}^0 + \hat{f}^1\hat{p}^1)$  (that is, every excursions in which only the first non-reference genome in  $\mathcal{G}$  is not diploid) of length  $n_e = y_e - x_e$ , resample  $B_e$  Bootstrap replicates  $\Pi_e^{b*}, b = 1, \dots, B_e$  (where  $B_e$  is determined below) of  $\Pi$  on  $e$ , and compute local proportion estimates  $(\hat{f}_e^1)^{b*}$  as the medians of  $|\Pi_e^{b*} - 2|/f$ .
  - The resample sizes  $B_e$  are chosen as the closest integer to  $Bn_e/(\sum_{e'} n_{e'})$ , where  $B$  is chosen as large as computationally feasible (by default  $B = 10^3$ ).
2. Remove outliers from  $\bigcup_{e,b} (\hat{f}_e^1)^{b*}$  by truncating past 3.5 median deviations, and return a 95% confidence interval around the resulting distribution's median.
3. Repeat steps 1 and 2 on each higher-order  $\hat{\pi}^k$  that is not uniformly diploid, incorporating for each  $e$  whose effective copy number estimate is contributed to by  $f^1, \dots, f^k$  the previously estimated uncertainties of  $f^1, \dots, f^{k-1}$ .

The weighting scheme in step 1 is designed so as to attribute more importance to longer excursions, as these contribute more heavily towards estimating  $\hat{f}^G$ , with the trimming of step 2 encouraging erroneously called excursions to be excluded. Resorting to previously estimated uncertainties in  $f^1, \dots, f^{k-1}$  when estimating confidence intervals for  $f^k$  in step 3 is necessary as fluctuations of  $\sum_{i=1}^k \hat{f}_k \hat{p}_k$  inform fluctuations in  $\hat{f}^k$  only when the fluctuations of  $f^1, \dots, f^{k-1}$  are known. Confidence intervals computed in this manner are likely to be conservative (though proving so requires further assumptions on the uncaptured covariates), since the bootstrapping design above effectively simulates inference of  $\hat{f}^G$  on each excursion individually, while HiDENSEC estimates  $\hat{f}^G$  using all excursions jointly. Nevertheless, **Supplementary Figure A.3** illustrates that the resulting confidence intervals are reasonably small whenever appropriate.

## A.4 Inference of large-scale structural variants

Large-scale structural variants typically result in off-diagonal intensities arranged in either of the six patterns given in **Supplementary Figure A.14**, which in the following will be referred to as  $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 = \{\square, \square, \square, \square\} \cup \{\boxplus, \boxplus\}$ . While events in  $\mathcal{P}_1$  are most often associated with changes in the copy number profile, rearrangements falling into  $\mathcal{P}_2$  typically are not, and so HiDENSEC treats their analysis separately. In particular, while HiDENSEC

largely relies on its previously inferred copy number profiles  $\hat{\pi}$  for detecting the former, the latter are called primarily based on their characteristic diagonal shape.

## Detecting patterns in $\mathcal{P}_1$

Hi-C sub-matrices structured like the patterns in  $\mathcal{P}_1$  can be found abundantly throughout the entire genome, and often correspond to intrinsic DNA geometry, compartment structure, or simply stochastic fluctuations inherent in the underlying biological and experimental processes. Moreover, in particularly noisy data or comparatively complex rearrangements, the area of largest intensity in any given  $p \in \mathcal{P}_1$  may not be straightforward to identify, in which case all  $p, q \in \mathcal{P}_1$  are approximately related to each other by a translation, and assigning one of them to a given empirical Hi-C sub-matrix may be under-determined. HiDENSEC addresses these two sources of uncertainty in two ways:

1. By default, HiDENSEC only reports off-diagonal events associated with excursions and corresponding copy number change points that have been evaluated as significant under the hypothesis testing scheme described in Sec. [A.3](#). Switching to non-default behavior and scanning arbitrary points along the genome is possible, but care should be taken in interpretation, as off-diagonal squares of enriched read counts may be confounded by above-mentioned biological and experimental hidden covariates. Reliably distinguishing signal due to noise from signal due to genomic rearrangement is often difficult even under manual detection by experts.
2. Since biological or experimental noise rarely result in individual  $\mathcal{P}_1$  patterns in isolation, but rather display effects that tend to propagate horizontally, vertically, and locally along the Hi-C matrix (see, e.g.,  $\chi_6^p$  of Sample 1-II in **Figure 4**), each candidate Hi-C sub-matrix  $H[\mathcal{J}, \mathcal{K}]$  that may potentially contain signal reflecting large-scale structural variant is evaluated in comparison to all sub-matrices obtained by translating  $H[\mathcal{J}, \mathcal{K}]$  vertically (i.e.,  $\{H[\mathcal{J}, \mathcal{K} + y]\}_y$ ), horizontally (i.e.,  $\{H[\mathcal{J} + x, \mathcal{K}]\}_x$ ), and locally (i.e.,  $\{H[\mathcal{J} + x, \mathcal{K} + y]\}_{x,y=-w}^w$  for some window size  $w$ ). Only if a suitable summary statistic of  $H[\mathcal{J}, \mathcal{K}]$  (to be discussed below) appears sufficiently significant in comparison to the entire class of shifted sub-matrices, is  $H[\mathcal{J}, \mathcal{K}]$  declared as containing evidence of genomic rearrangement events.

More concretely, HiDENSEC proceeds as follows.

1. For a given set of excursions  $\mathcal{E}$  and associated  $p$ -values as determined in Sec. [A.3](#), select their most significant subset through Benjamini-Hochberg on a given significance threshold  $\alpha$  (by default  $\alpha = 0.05$ ). Call  $C$  the set of boundary points of the so selected candidate excursions.
2. For a choice of weight  $w = (w_1, w_2)$  and off-diagonal point  $x = (x_1, x_2)$ , define the four quadrants

$$Q_w^{jk}(x) = \{(x_1 + m, x_2 + n) : m \in [0, jw_1], n \in [0, kw_2]\},$$

for  $j, k \in \{\pm 1\}$ , and denote by  $H_w(x)$  the associated empirical distribution

$$H_w(x) = \sum_{j', k' \in \cup_{j, k} Q_w^{jk}(x)} H_{j'k'} \delta_{j'k'}$$

of read count locations. If  $X \sim H_w(x), Y \sim \text{Uniform}\left(\bigcup_{j, k \in \{\pm 1\}} Q_w^{jk}(x)\right)$  are random variables distributed according to  $H_w(x)$  and the uniform measure on  $[x_1 - w_1, x_1 + w_1] \otimes [x_2 - w_2, x_2 + w_2]$ , respectively, then HiDENSEC considers as test statistics

$$S_w^1(x) = \eta(\mathbb{E}[X \mid \mathcal{Q}_w(x)]) \quad S_w^{2, \tau}(x) = \mathbb{E} \text{Var}[\mathbb{E}(H(Y) \mid Y_\tau) \mid \mathcal{Q}_w^\tau(x)],$$

for  $\tau \in \{1, 2\}$ , where  $\eta(Z)$  is the entropy of the random variable  $Z$ ,  $\mathcal{Q}_w = \{Q_w^{jk}(x)\}_{j, k \in \{\pm 1\}}$ ,  $\mathcal{Q}_w^1(x) = \left\{ \bigcup_{k \in \{\pm 1\}} Q_w^{jk} \right\}_{j \in \{\pm 1\}}$ , and  $\mathcal{Q}_w^2(x) = \left\{ \bigcup_{j \in \{\pm 1\}} Q_w^{jk} \right\}_{k \in \{\pm 1\}}$ . That is, while  $S_w^1(x)$  essentially captures the extent to which read counts tend to accumulate in only one of the quadrants,  $S_w^{2, \tau}$  measures whether read counts, projected onto the  $X_\tau$  coordinate, exhibit evidence of copy numbers changing at  $x$  (cf. Eq. (A.3) in Sec. A.3).

3. For each pair of boundary points  $\{x_1, x_2\} \in \binom{C}{2}$  that fall into distinct chromosomes, refine its location by maximizing  $S_w^1(x_1, x_2)$  locally through, e.g., coordinate ascent, and denote the resulting  $\binom{|C|}{2}$  off-diagonal indices by  $C$  as well.
4. For each  $\{x_1, x_2\} \in C$ , compute  $p$ -values  $p_{x_1}^1(x_2)$  and  $p_{x_2}^1(x_1)$  from comparing  $S_w^1(x, y)$  against the empirical distributions  $\hat{S}_w^1(x_1) = \{S_w^1(x_1, y)\}_y$  and  $\hat{S}_w^1(x_2) = \{S_w^1(y, x_2)\}_y$ , where the index  $y$  ranges over all genomic locations not part of the chromosome containing  $x_1$  and  $x_2$ , respectively.
5. Compute  $p$ -values  $p^{2, \tau}(x_1, x_2)$  by comparing  $S_w^{2, \tau}(x_1, x_2)$  against the permuted random variable  $\tilde{X} \sim (\sigma_1 \circ X_1, \sigma_2 \circ X_2)$ , where  $\sigma_k$  is drawn uniformly from the symmetric group on  $[-w_k, w_k]$ .
6. Under the null hypothesis of  $S_w^1(x)$  following either  $\hat{S}_w^1(x_1)$  or  $\hat{S}_w^1(x_2)$ , and  $X_1 \perp X_2$ ,  $H(X_\tau) \perp \mathcal{Q}_w^\tau(x)$ ,  $m_w(x_1, x_2) = \min\{p_{x_1}^1(x_2), p_{x_2}^1(x_1)\}$  is super-uniform, while  $p^{2, \tau}(x_1, x_2)$  are uniformly distributed, with all three quantities independent of each other (note that the optimization in step (3) may affect these properties slightly, though as long as the refinement is kept sufficiently local, its impact appears empirically negligible; see **Supplementary Figure A.13**). HiDENSEC thus ranks candidates in  $C$  based on  $p$ -values associated with  $m_w(x_1, x_2) + \sum_\tau p^{2, \tau}(x_1, x_2)$ , and declares a set  $C_+ \subset C$  of significant off-diagonal contacts by means of the Benjamini-Hochberg procedure.
7. For each site  $x$  identified in step (1), denote by  $C_x \subset \bigcup_{c \in C_+} c$  the set of all its refinements partaking in a significant pair, and extract a single refinement by computing a combined on- and off-diagonal statistic akin to Eq. (A.3) on each element in  $[\min C_x, \max C_x]$ .



## Detecting patterns in $\mathcal{P}_2$

Patterns in  $\mathcal{P}_2$  are typically not tied to changes in copy number profiles, and thus require a more global search than what was necessary in the case of  $\mathcal{P}_1$ . However, their characteristic block-diagonal shape is rather more rigid; e.g., translational and within-block rotational symmetries do not apply in the same manner they did in  $\mathcal{P}_1$ , which HiDENSEC exploits for their detection. More explicitly, HiDENSEC proceeds as follows.

1. For each pair of chromosomes  $\{\chi_a, \chi_b\}$ , denote by  $H[\chi_a, \chi_b]$  the Hi-C sub-matrix recording all contacts between  $\chi_a$  and  $\chi_b$ .
2. For a fixed choice of  $r$  (by default,  $r = 50$ ), convolve  $H[\chi_a, \chi_b]$  by  $r^{-2}\mathbb{1}_{[r]} \otimes \mathbb{1}_{[r]}$ , where  $\mathbb{1}_{[r]} \in \mathbb{R}^r$  is the all-ones vector, and replace each entry  $h_{ij}$  of the resulting smoothed matrix  $\tilde{H}[\chi_a, \chi_b]$  by  $\mathbb{1}_{h_{ij} > m}$ , where  $m$  is the median of non-zero values in  $\tilde{H}[\chi_a, \chi_b]$ . Interpret the so-constructed matrix  $\overline{H}[\chi_a, \chi_b]$  as an encoding for a graph  $G[\chi_a, \chi_b]$ , whose vertices  $v$  are labeled  $\{1, \dots, |\chi_a|\} \times \{1, \dots, |\chi_b|\}$  and whose every pair of vertices  $v, w$  is connected by an edge if  $\min_{s \in \{v, w\}} \{\overline{H}[\chi_a, \chi_b]_s\} = 1$  and  $\|v - w\|_\infty = 1$ .
3. Fix a number  $C$  of candidates to be considered per chromosome pair  $\{\chi_a, \chi_b\}$ , and identify the  $C$  largest connected components  $K_1[\chi_a, \chi_b], \dots, K_C[\chi_a, \chi_b]$  (ordered in decreasing size) of  $G[\chi_a, \chi_b]$ .
4. For each component  $K_j[\chi_a, \chi_b]$  identified in the step above, extract the vertex  $v_j[\chi_a, \chi_b]$ , such that  $v_j[\chi_a, \chi_b] = \underset{v}{\operatorname{arg\,max}} \overline{H}[\chi_a, \chi_b](v)$ , and let  $\mathcal{V}[\chi_a, \chi_b] = \bigcup_{j \in [C]} v_j[\chi_a, \chi_b]$  be the collection of these vertices.
5. For a choice of window size  $w$ , sites  $x, y$  and  $Q_w^{jk}, \mathcal{Q}_w, \mathcal{Q}_w^\tau, H_w, X, Y$  as introduced in the previous section, define the statistics  $T_w^{1,\sigma}(x, y), T_w^{2,\sigma}, T_w^{3,\sigma}, \sigma \in \{\pm 1\}$  as

$$\sigma T_w^{1,\sigma} = \mathbb{P}[(X, X') \in (Q_w^{11}(x, y), Q_w^{-1-1}(x, y))] - \mathbb{P}[(X, X') \in (Q_w^{-1+1}(x, y), Q_w^{+1-1}(x, y))]$$

$$T_w^{2,\sigma} = \mathbb{E} [H_w(Y) \mid \|Y - (x, y)\|_\infty \leq 3, Y \in Q_w^{1\sigma} \cup Q_w^{-1-\sigma}]$$

$$\sigma T_w^{3,\sigma} = \tau_{12}\tau_{21} - \tau_{11}\tau_{22},$$

where  $X'$  is an *iid* copy of  $X$ , and  $\tau_{kj}$  is given by

$$\tau_{kj} = \tau \left( m, \mathbb{E} [H_w(Y) \mid \|Y - (x, y)\|_\infty = m, Y \in Q_w^{kj}] \right)_{m \in [\rho]},$$

with  $\tau(\cdot)$  being Kendall's  $\tau$ , and  $\rho$  a pre-specified radius.

6. For each  $v \in \bigcup_{a,b} \mathcal{V}[\chi_a, \chi_b]$ , refine its location by locally maximizing first  $H_w(v)$ , and then  $T_w^{1,\sigma}(v)$ . Call these two collections of refined vertices  $\mathcal{V}^\sigma, \sigma \in \{\pm 1\}$ .

7. For each  $v \in \mathcal{V}^\sigma$ , compute  $(T_w^{j,\sigma}(v))_{j \in \{1,2,3\}}$ , and construct  $p$ -values  $p_w^{1,\sigma}(v), p_w^{2,\sigma}(v)$  for every  $v$  with  $T_w^{3,\sigma}(v)$  larger than some threshold  $t$  (by default,  $t = 5$ ) as

$$p_w^{j,\sigma}(v) = \Phi_{\mu^{j,\sigma}, \nu^{j,\sigma}}(T_w^{j,\sigma}(v)),$$

where  $\Phi_{\mu,\nu}$  is the CDF of a Gaussian distribution with expectation  $\mu$  and variance  $\nu$ , and

$$\mu^{j,\sigma} = |\mathcal{V}^\sigma|^{-1} \sum_{v \in \mathcal{V}^\sigma} T_w^{j,\sigma}(v) \quad \nu^{j,\sigma} = |\mathcal{V}^\sigma|^{-1} \sum_{v \in \mathcal{V}^\sigma} (T_w^{j,\sigma}(v) - \mu^{j,\sigma})^2.$$

8. Under the null hypothesis of  $(X | H_w(X))$  being uniformly distributed on  $\mathcal{Q}_w$ ,  $T_w^{1,\sigma}$  and  $T_w^{2,\sigma}$  become Gaussian as  $w$  and  $\rho$  increase, and so  $p_w^{j,\sigma}$  is approximately calibrated (see **Supplementary Figure A.15C**). HiDENSEC then uses  $\min\{p_w^{1,\sigma}, p_w^{2,\sigma}\}$  to select off-diagonal locations  $v$  likely to exhibit patterns in  $\mathcal{P}_2$ .

**Supplementary Figure A.15A,B** demonstrates the power and accuracy of the selection scheme described above, showcasing its strong calibration, high sensitivity, and precise localization: Manual expert inspection of all Hi-C matrices analysed in the main text yielded three distinct type- $\mathcal{P}_2$  fusion events, two of which are associated with mixture proportions of  $\approx 10\%$ . HiNT does not identify these two events as such (likely precisely due to their small associated proportions), but does declare the remaining third event as significant (alongside a similar number of false positives as discussed in the section above); however, returning a location estimate that differs from the actual signal by about  $\|\hat{x}_{\text{HiNT}} - x\|_1 \approx 67\text{MB}$ . In contrast, HiDENSEC correctly identifies all three—and only these three; i.e., at zero false-positive rate—events as such, with its location estimates coinciding precisely with those obtained from visual inspection.

## Benchmarking

In order to more thoroughly assess the performance of HiDENSEC relative to HiNT outside the context of cell lines, the same benchmarking procedure as displayed in **Figure 3** was employed on all analyzed samples. As **Supplementary Figure A.16** demonstrates, the relative improvement in top- $k$  recall remains as pronounced as, if not more so, in the setting of cell lines.

## A.5 Proof of Theorem

Identifying the smallest number of genomes  $|\mathcal{G}|$  that explain a given noise-less effective copy number profile  $\pi$  using mixture proportions bounded away from zero (e.g.,  $\min_{G \in \mathcal{G}} f^G \geq o(|\mathcal{G}|^{-1})$ ) and bounded absolute copy numbers (i.e.,  $\max_{G \in \mathcal{G}} \|p^G\|_\infty \leq B$  for some  $B \in \mathbb{N}$ ) is, in general, at least as hard as the subset sum problem, and therefore NP-complete. The

proof proceeds by reducing the subset sum problem to two variants of it, one of which will be directly reducible to identifying  $|\mathcal{G}|$ . It begins by recalling the subset-sum problem in one of its most commonly stated form (here referred to as **SSP**<sub>0</sub>): [**SSP**<sub>0</sub>] Given a set  $S \subset \mathbb{Q}_+$  of  $K$  non-negative rational numbers, and a target  $T \in \mathbb{Q}$ , decide whether there exists a subset  $R \subset S$ , so that  $\sum_{r \in R} r = T$ . **SSP**<sub>0</sub> is well known to be NP-complete, and so any reduction of it to a new task **P** will render **P** NP-hard. The **P** of interest in the case here is the following: [**Min** <sub>$|\mathcal{G}|$ ] Given a profile of effective copy numbers  $\{\pi_i\}_i$ , determine the smallest set  $\mathcal{G}$ , so that  $\pi = \sum_{g \in \mathcal{G}} f^G p^G$  for some mixture proportions  $f^G$  and absolute copy number profiles  $p^G$ , with  $\min_{G \in \mathcal{G}} f^G \geq g(|\mathcal{G}|) \in o(|\mathcal{G}|^{-1})$  and  $\max_{G \in \mathcal{G}} \|p^G\|_\infty \leq B$  for some  $B \in \mathbb{N}$ . It is clear that **Min** <sub>$|\mathcal{G}|$   $\in$  **NP**, and so reducing **SSP**<sub>0</sub> to **Min** <sub>$|\mathcal{G}|$  suffices to show that it is NP-complete. To do so, two intermediary reductions are needed:</sub></sub></sub>

$$\mathbf{SSP}_0 \leq \mathbf{SSP}_1 \leq \mathbf{SSP}_2 \leq \mathbf{Min}_{|\mathcal{G}|},$$

where **SSP**<sub>1</sub> and **SSP**<sub>2</sub> are defined to be [**SSP**<sub>1</sub>] Given a set  $S \subset \mathbb{Q} \cap [0, 1]$  whose elements are linearly independent in the  $\mathbb{Z}/B\mathbb{Z}$ -module  $\mathbb{Q}$ , less than  $2g(K)$ , and sum to less than or 1; and a target  $T$ , deciding if there exists a subset  $R \subset S$  for which  $\sum_{r \in R} r = T$  is NP-hard. [**SSP**<sub>2</sub>] Given a set  $S$  as in **SSP**<sub>1</sub>, and a target  $T$ , deciding if there exists a subset  $R \subset S$ , and multiplicities  $m \subset \mathbb{N}^S$  for which  $\sum_{r \in R} m_r r = T$  is NP-hard. Indeed, if **SSP**<sub>2</sub> is known to be NP-hard, then hardness of **Min** <sub>$|\mathcal{G}|$  follows:</sub>

$$\mathbf{SSP}_2 \leq \mathbf{Min}_{|\mathcal{G}|}.$$

[Proof of lemma] Given an instance of **SSP**<sub>2</sub>, enumerate the elements of  $S$  as  $\{s_k\}_{k \in [K]}$ , and construct an effective copy number profile consisting of  $\pi_k = s_k$  as well as  $\pi_{K+1} = T$ . Due to the linear independence and boundedness assumptions on  $S$ , any  $\mathcal{G}$  explaining such  $\pi$  must be of size at least  $K$  (with  $f^{G_k} = s_k$  for  $k \in [K]$ ), and will be of size  $K + 1$  if and only if  $T = \sum_{k=1}^K f^{G_k} p^{G_k} = \sum_{k=1}^K s_k p^{G_k}$  for some  $p_{K+1}^{\mathcal{G}} \in \mathbb{N}^K$ . That is, if a set of genomes  $\mathcal{G}$  of size  $|\mathcal{G}| = K$  explains  $\pi$ , then setting  $m_k = p_{K+1}^{G_k}$  solves **SSP**<sub>2</sub>; while otherwise no solution to **SSP**<sub>2</sub> exists. Thus it remains to show that **SSP**<sub>0</sub>  $\leq$  **SSP**<sub>2</sub>.

$$\mathbf{SSP}_0 \leq \mathbf{SSP}_1.$$

[Proof of lemma] Given an instance of **SSP**<sub>0</sub>,

1. find an invertible linear transformation  $\tau(x) = ax + b$  such that  $\tau(S)$  satisfies the boundedness assumptions of **SSP**<sub>1</sub>, and  $b = b_0 + 10^{-e_0}$  for some  $e_0$  much larger than any of the  $e_k$  discussed below,
2. replace each  $s_k$  by two new elements

$$s'_k = s_k^0 + 10^{-e_k} \qquad s''_k = s_k^1 - 10^{-e_k},$$

where  $s_k^i$  are positive,  $s_k^0 + s_k^1 = \tau(s_k)$ , and  $S' = \cup_{k \in [K]} \{s'_k, s''_k\}$  is linearly independent in the  $\mathbb{Z}/B\mathbb{Z}$ -module  $\mathbb{Q}$ , and  $e_k \in \mathbb{N}$  are exponents larger than the maximum of

$\mathcal{F}_{10}(aT + Kb)$  and  $\max_{k \in [K], i \in \{0,1\}} \mathcal{F}_{10}(s_k^i)$  (where  $\mathcal{F}_{10}(x)$  is the largest index—counting from the left—at which the base-10 expansion of  $x$  is non-zero), distinct from each other; i.e.,  $e_k \neq e_\ell$  if  $k \neq \ell$ , and chosen so as to not violate any boundedness assumptions.

Then  $(S', aT + kb)_{k \in [K]}$  are all valid instances for **SSP**<sub>1</sub>, and any solution must either select both or neither of  $s'_k$  and  $s''_k$ . If one of these instances, say the  $k_*^{\text{th}}$ , accepts on a subset of indices  $R$ , then  $|R| = k_*$  due to the choice of  $e_0$ , and since

$$\sum_{r \in R} a s_r + b = k_* b + a \sum_{r \in R} s_r = aT + k_* b,$$

it must be true that  $\sum_{r \in R} s_r = T$ , and so  $R$  too provides a positive answer to  $(S, T)$ . Conversely, a solution  $R$  to  $(S, T)$  will provide a solution to  $(S', aT + |R|b)$ , and so the lemma is proved.

$$\mathbf{SSP}_1 \leq \mathbf{SSP}_2.$$

[Proof of lemma] A similar proof idea as in the lemma just proved works here as well: Each element  $s_k \in S$  is replaced by two elements that indicate whether  $s_k$  is used once or not at all in the following manner.

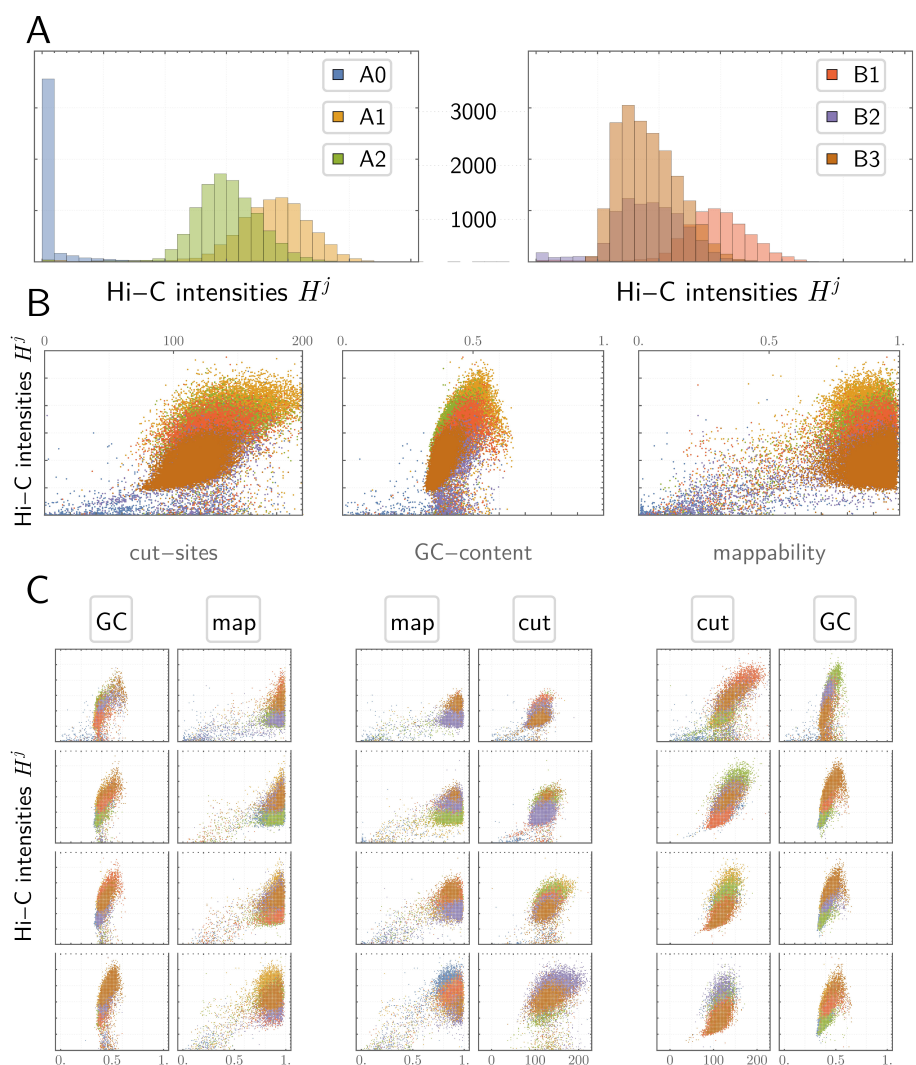
1. Choose  $(e_k)_{k \in [K]}$ , so that  $e_k \geq B' + \max_{x \in S \cup \{T\}} \mathcal{F}_{10}(x)$ , and such that  $|e_k - e_\ell| \geq B'$  for some  $B' > B$ .
2. Replace each element  $s_k$  with two elements

$$s'_k = 10^{-e_k} \qquad s''_k = s_k + 10^{-e_k}.$$

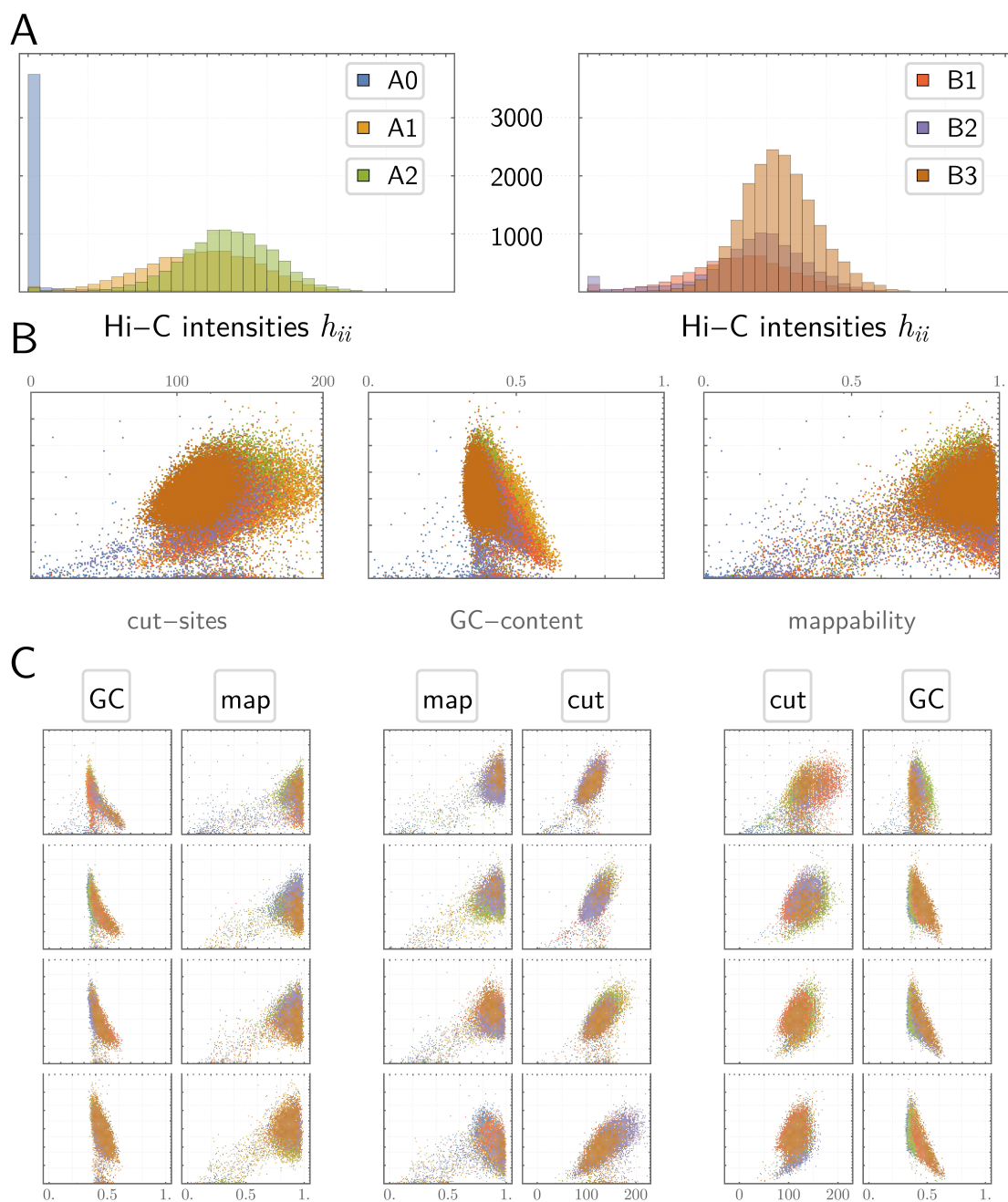
3. Define  $S' = \cup_{s \in S} \{s'_k, s''_k\}$ ,  $T' = T + \sum_{k \in [K]} 10^{-e_k}$ , and query **SSP**<sub>2</sub> on the instance  $(S', T')$ .

If **SSP**<sub>2</sub> returns a solution  $R'$  to this instance, then  $R = \{k : s''_k \in R'\}$  provides a solution of indices to **SSP**<sub>1</sub> on  $(S, T)$ . The converse direction is clear. Chaining together these individual lemmas yields the theorem as desired. Although this proof may appear contrived on first glance, it in fact describes the very difficulty HiDENSEC must deal with: Given various levels  $\pi_1, \dots, \pi_K$  of  $\pi$ , can a new level  $\pi_{K+1}$  be explained by the same genomes that explain  $\pi_1$  through  $\pi_K$  or is the introduction of a new one necessary? The proof shows that even when a set of genomes explaining  $\pi_1, \dots, \pi_K$  is known, answering this question in general is intractable—therefore, in practice, where the genomes explaining  $\pi_1$  through  $\pi_K$  are not known and must be estimated themselves, this must be true too.

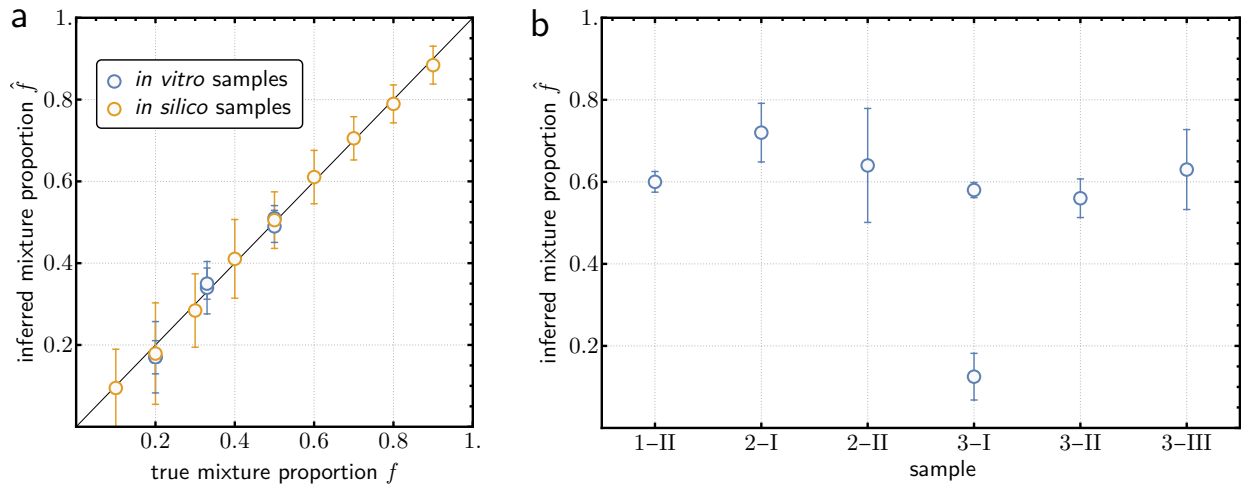
## A.6 Supplementary Figures



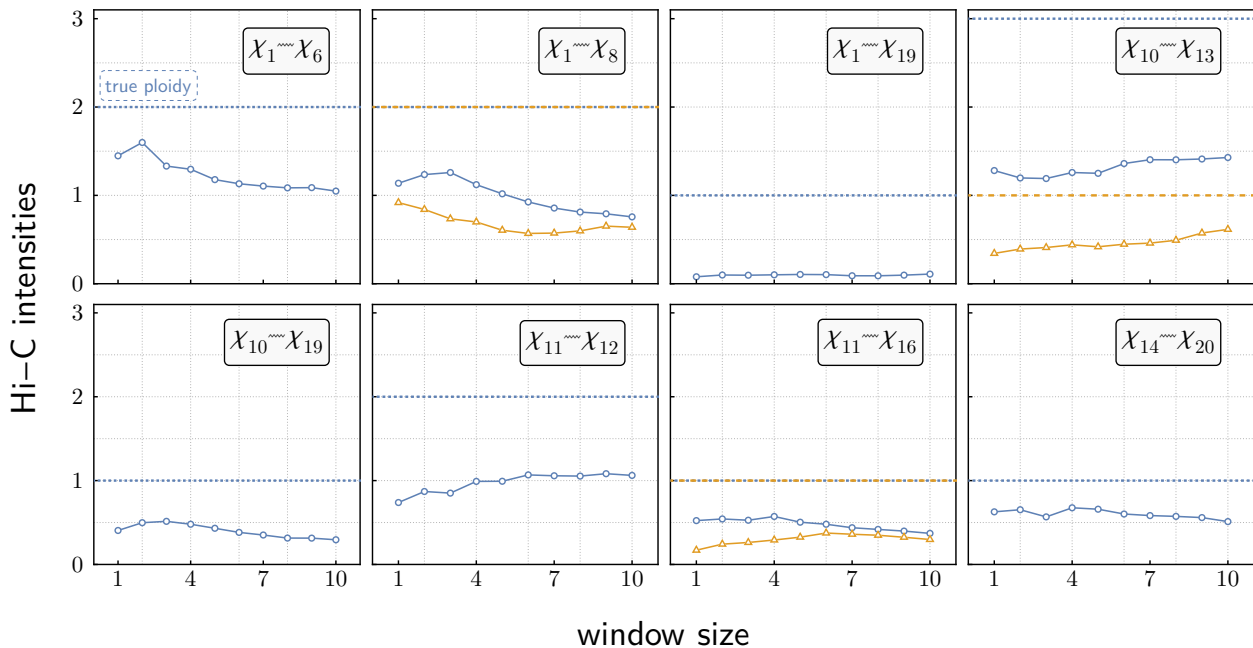
Supplementary Figure A.1: **Covariate dependence of  $H^j = \sum_k H_{jk}$  in GM12878 *in situ* Hi-C data.** The impact of the four covariates compartment structure, GC-content, number of cut-sites and mappability on row sums of Hi-C intensity matrices is displayed. **A:**  $H^j$  conditioned on compartment structure, **B:**  $H^j$  as a function of remaining three covariates; points are coloured by compartment, **C:**  $H^j$  conditioned on quartiles of the corresponding column statistic in B, as a function of the two remaining covariates.



Supplementary Figure A.2: **Covariate dependence of  $H^j = \sum_k H_{jk}$  in Sample 1-I *in vivo* Fix-C data.** Plots are as described in **Supplementary Figure A.1.**

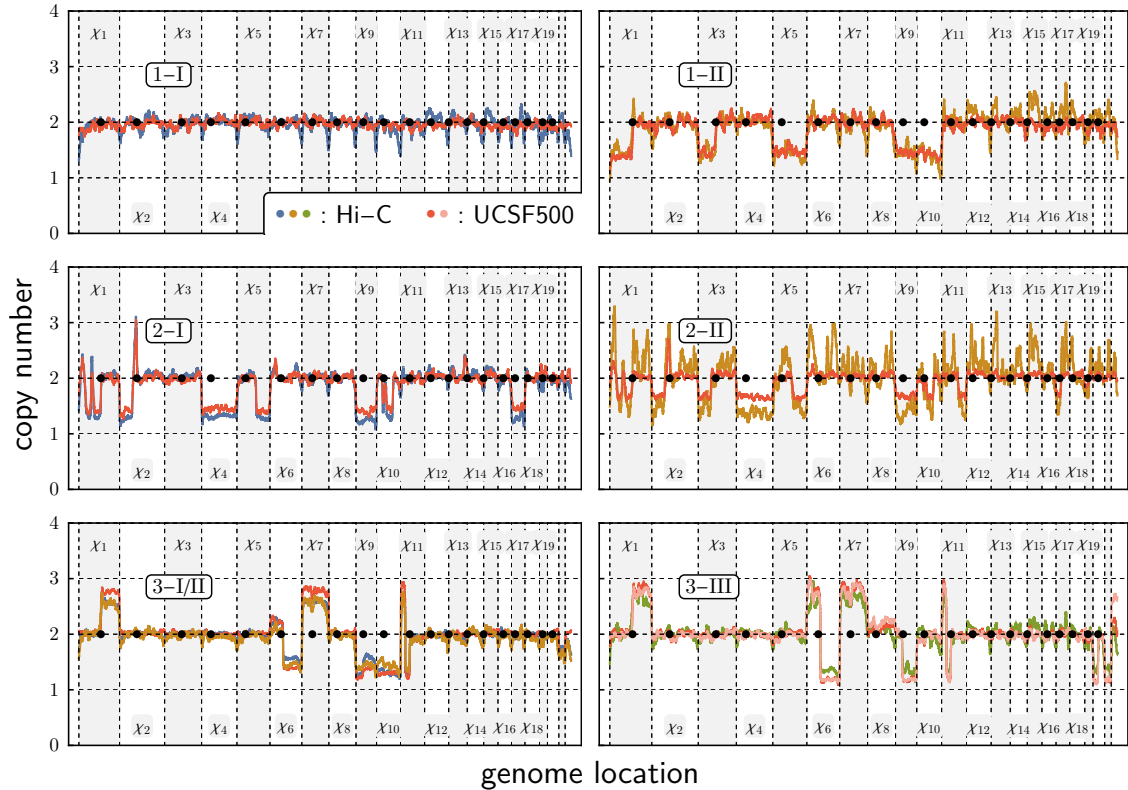


Supplementary Figure A.3: **Predictions of tumor purity correlate well with known tumor purity.** (a) The x-axis represents true tumor purity for *in vitro* and *in silico* samples; while the y-axis represents the HiDENSEC inferred tumor purities. There is high concordance between the two and the error bars represent 95% confidence intervals for the HiDENSEC inferred tumor purity. (b) The x-axis represents different samples used in the analysis (1-II represents Sample 1-II, for instance). The y-axis represents the inferred tumor purity (the fraction of cells that are cancerous). The error bars represent 95% confidence intervals for the inferred tumor purity.

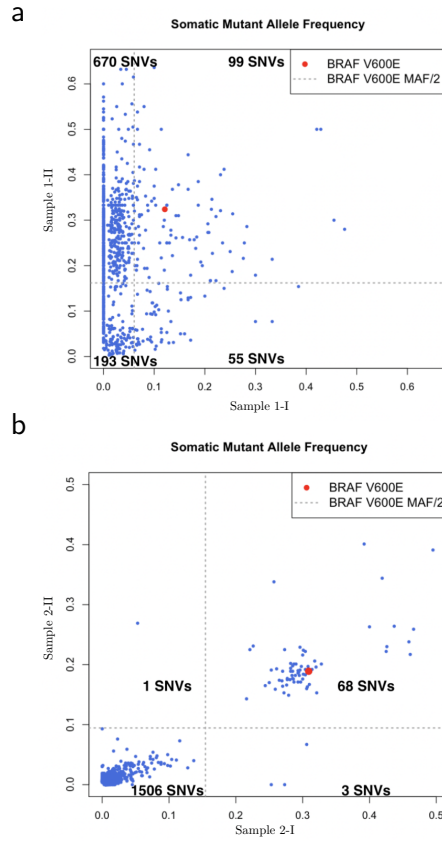


Supplementary Figure A.4: **Off-diagonal intensities of LSSVs aren't reliable estimators of absolute copy number.** Each panel represents the Hi-C intensities corresponding to a particular off-diagonal event in the HCC1187C cell line (for instance the top-left panel represents a translocation between chromosome 1 and chromosome 6). The horizontal axis represents a measure of how large a window around the translocation (which manifests itself as an off-diagonal event on the Hi-C map) was considered to compute the Hi-C intensity. The Y-axis represents the resulting Hi-C intensity. The true ploidies of inter-chromosomal translocations are denoted by the horizontal dotted lines while the colored curves represent the measured Hi-C intensities. For balanced translocations there are two colored lines corresponding to the two fusion events. The fact that the ratio of the true ploidy represented by the horizontal dotted lines and the colored curves is not consistent across the various LSSVs within the same sample, suggests that the off-diagonal intensities are confounded by covariates in addition to absolute copy number and hence HiDENSEC does not use off-diagonal Hi-C intensities to infer absolute copy numbers.

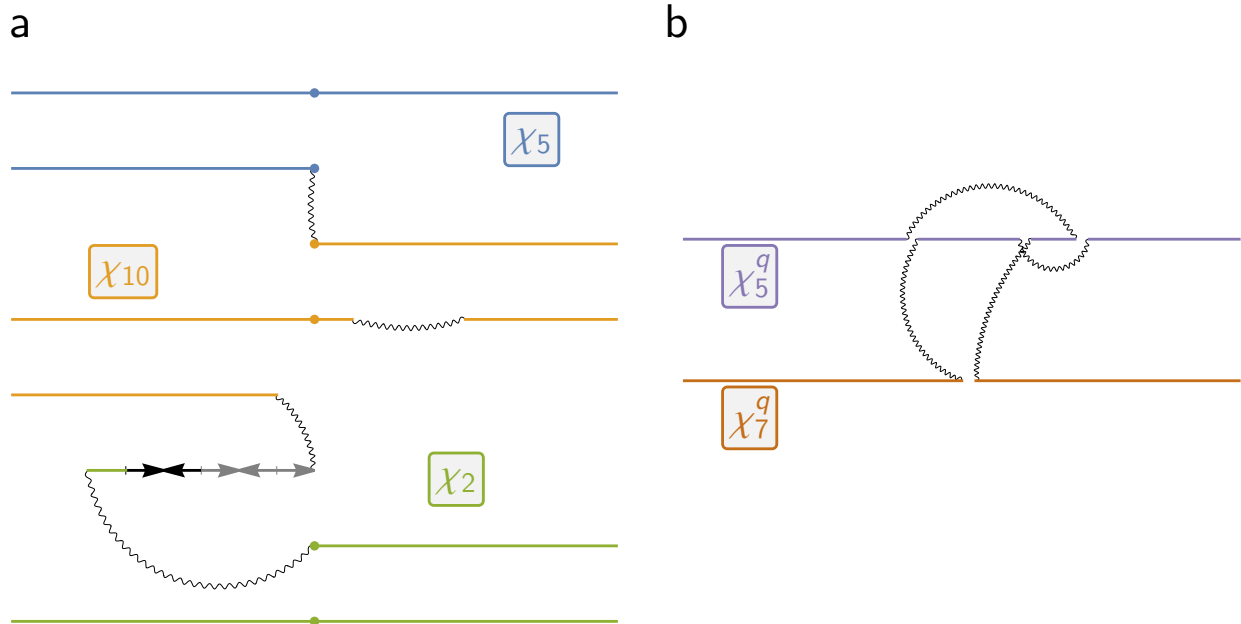




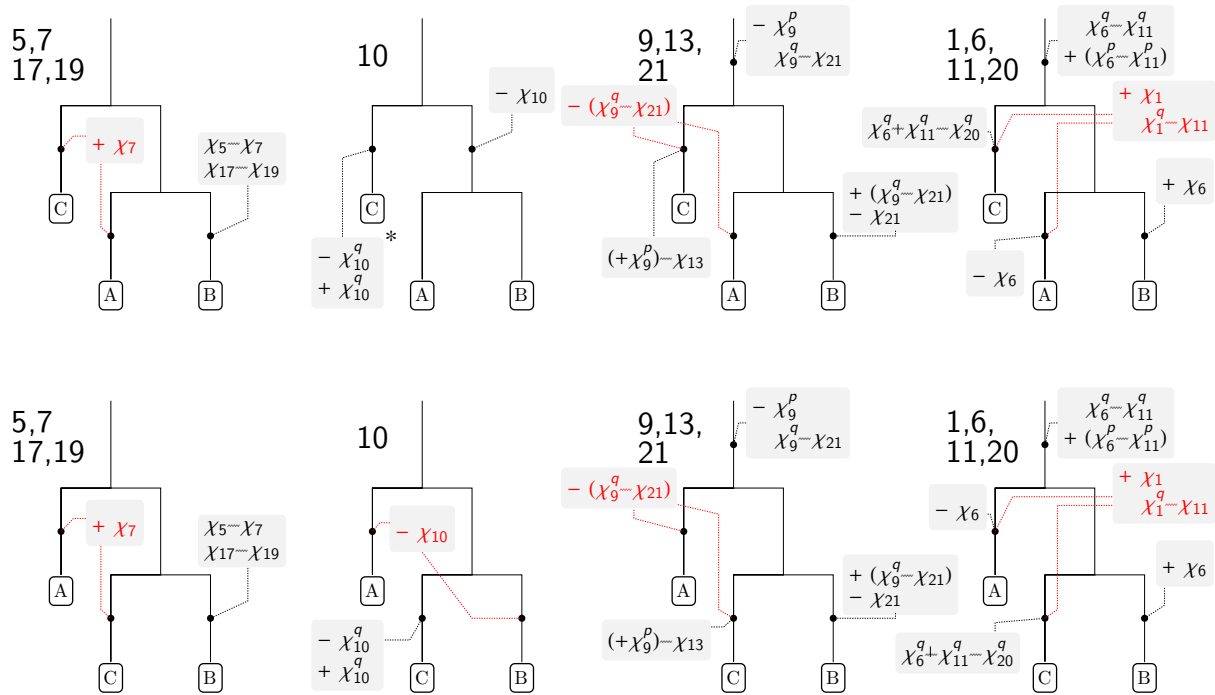
Supplementary Figure A.5: **HiDENSEC absolute copy number predictions correlate well with genome-wide copy numbers inferred from next-generation sequencing.** Each panel represents HiDENSEC absolute copy number predictions for a particular sample (whose identity is indicated inside each panel; e.g., 1-II in the top-right represents Sample 1 - II), compared to UCSF500 or Exome sequencing based (relative) copy number calls. Copy number profiles inferred by HiDENSEC from Hi-C data use color codes consistent with the main text (that is, blue curves correspond to Samples 1 - I, 2 - I, 3 - I, beige curves to Samples 1 - II, 2 - II, 3 - II, and the green curve depicts Sample 3 - III), while red and pink profiles represent relative copy numbers (transformed by  $x \rightarrow 2 \times 2^x$ ) from CNVkit using UCSF500 or Exome sequencing. For Sample 2 - I, the discordance between the levels of the HiDENSEC absolute copy numbers and the CNVkit scaled relative copy numbers inferred using UCSF500 data is likely due to differences in samples for the Hi-C data and for the UCSF500 data, since the UCSF500 based tumor purity lies outside the 95% confidence intervals of the HiDENSEC inferred tumor purity (**Supplementary Figure A.3, Supplementary Table 4**). For Sample 3 - III, two UCSF500 based curves are displayed, as UCSF500 data from two different metastases corresponding to this sample exists; both of these are concordant with the HiDENSEC absolute copy number inferred using Hi-C data from the metastasis sample, Sample 3 - III.



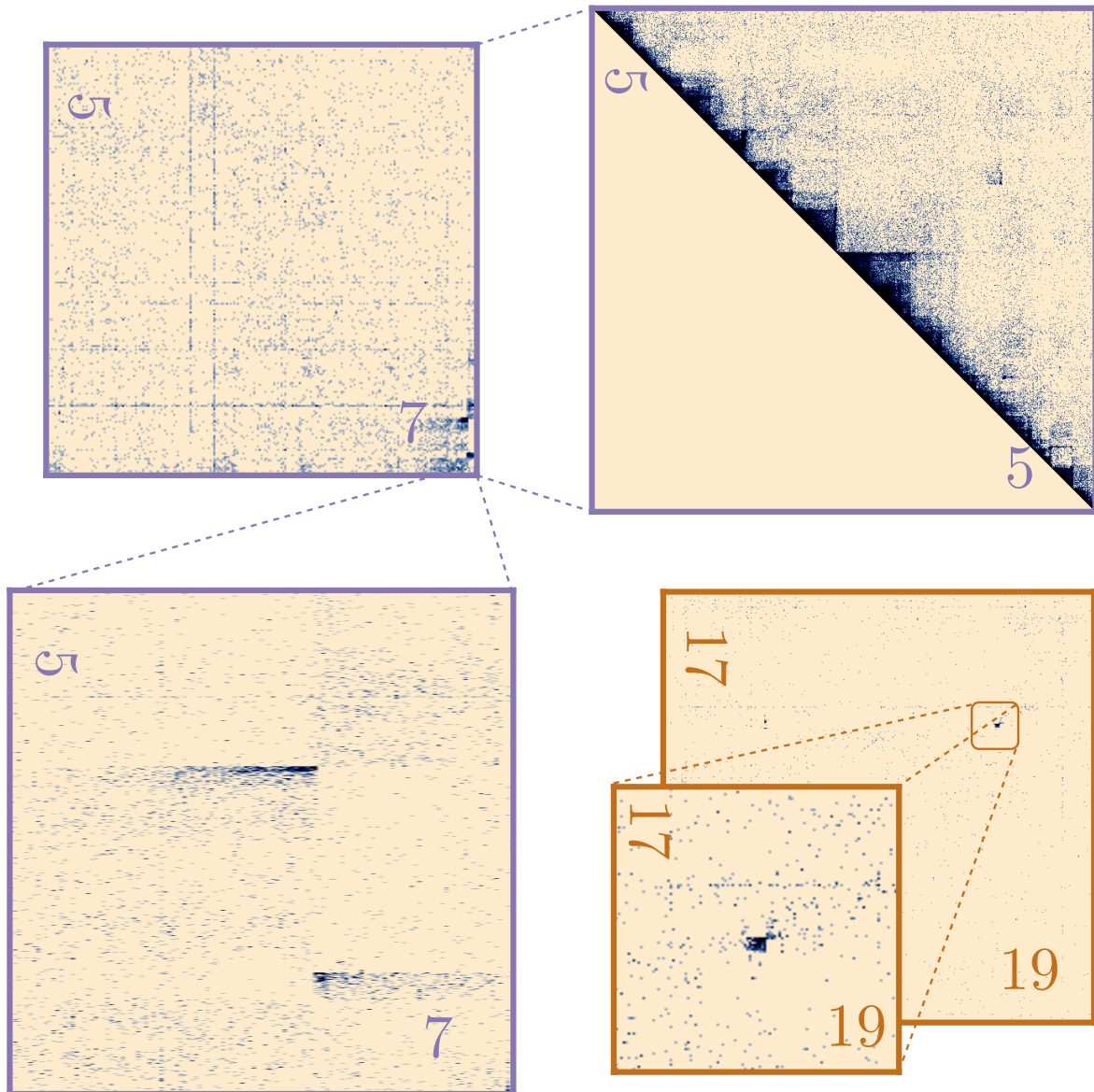
Supplementary Figure A.6: **Somatic mutant allele frequencies for Patient 1 and Patient 2, based on somatic variant calls derived from Exome sequencing data and UCSF500 data, respectively.** (a) Somatic variant calls were obtained using Mutect2 from exome sequencing data in samples derived from Patient 1 and (b) UCSF500 data in samples derived from Patient 2. These somatic variant calls were then filtered out for false positives and the intersection of mutations observed in both samples within a patient was considered. The x-axis and the y-axis in each of the two panels represents somatic mutant allele frequencies. Each individual data point is a particular somatic variant call, from among the intersection of filtered somatic variant calls within the two samples of a patient. The red point denotes the well-known BRAF V600E somatic variant while the dashed lines are drawn at exactly half of its mutant allele frequencies. The resulting quadrants are intended to denote somatic variants common to both samples (corresponding to the top-right quadrant), somatic variants present in only one of the samples (corresponding to the top-left and the bottom-right quadrants), and somatic variants which are likely false positives (bottom-left quadrant).



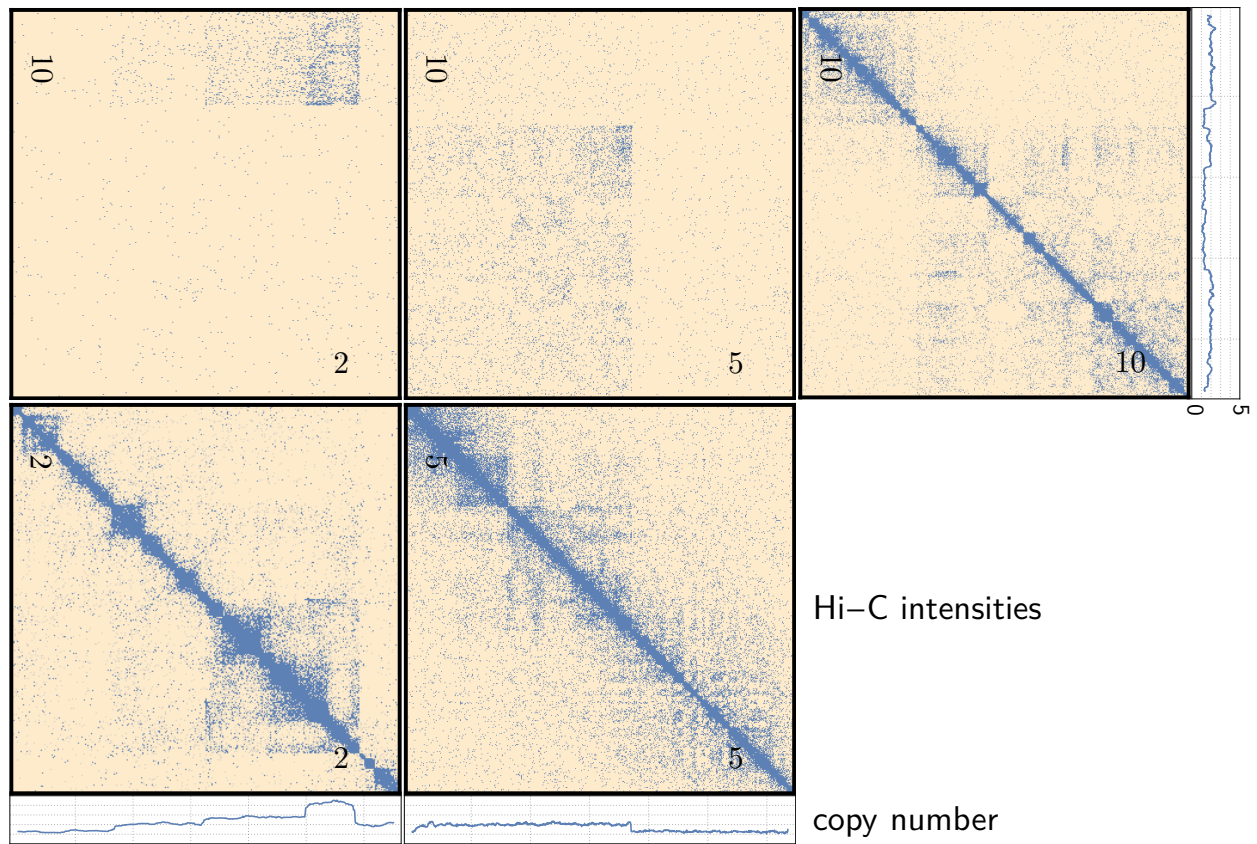
Supplementary Figure A.7: **Schematics demonstrating two particularly complex structural variants.** The translocations involving chromosome 2, 5 and 10 in both samples of Patient 2 (Sample 2 - I and Sample 2 - II) and the inter-chromosomal translocation between chromosome 5 and 7 in Sample - I have been depicted here. In order to infer the contacts constituting these complex structural variants, the Hi-C maps of the relevant chromosomes were carefully analyzed in conjunction with the HiDENSEC inferred copy numbers, as shown in **Supplementary Figure A.9, Supplementary Figure A.10**. Arrows indicate duplications and inversions of genomic segments of chromosome 2.



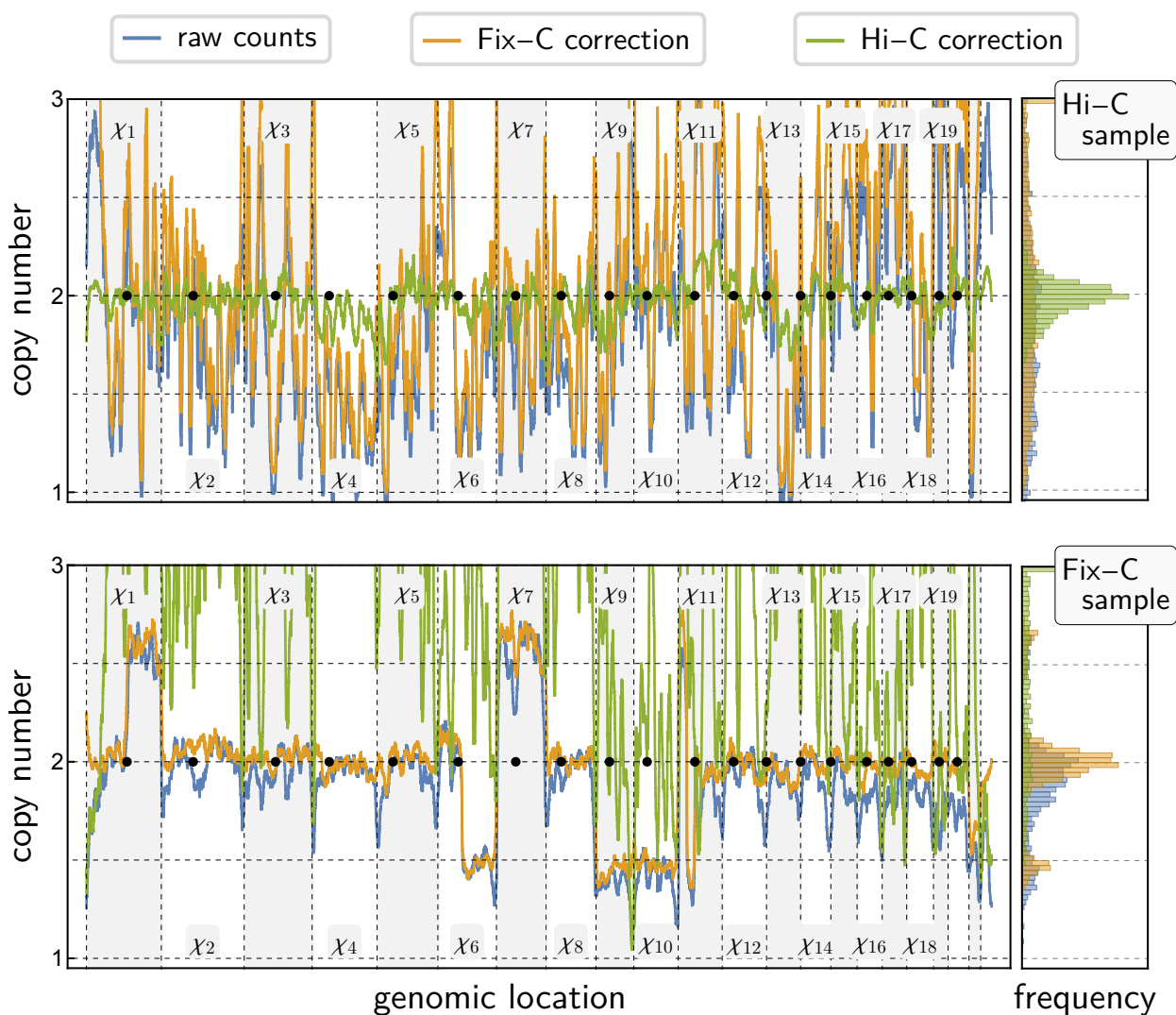
Supplementary Figure A.8: **Alternatives to the phylogenetic relationship between the three cell types in Patient 3** . Two alternative phylogenies that are consistent with the LSSVs in Patient 3. The phylogeny in **Figure 7b** was chosen over these two alternatives following the principle of parsimony since the number of convergent events, denoted in red, are higher in these two phylogenies, than the one described in **Figure 7b**.



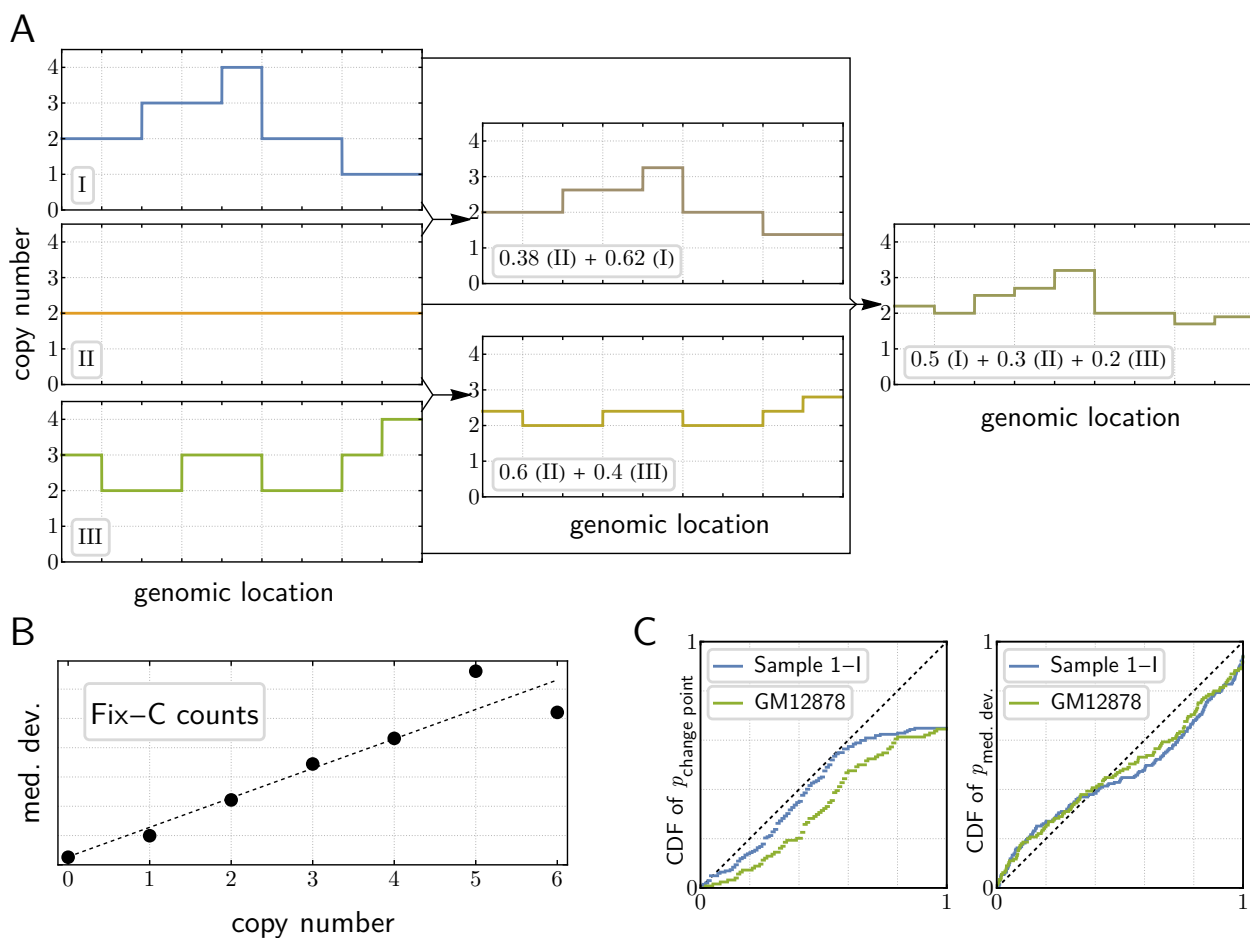
Supplementary Figure A.9: **Zoomed-in Hi-C maps of two translocations observed in Sample 3 - I.** Sample 3 - I contains two structural variants that are relatively smaller in size than chromosome arms. The first of which is a complex structural variant between chromosome 5 and chromosome 7, a schematic of which is depicted in **Supplementary Figure A.7b**. The second structural variant is a short balanced translocation between chromosome 17 and chromosome 19, depicted in the bottom right panel and its inset.



Supplementary Figure A.10: **Hi-C contact maps used to infer the inter-chromosomal translocations involving chromosomes 2, 5 and 10 in Patient 2.** This figure shows the zoomed in Hi-C maps for chromosome pairs involved in the complex structural event depicted in **Supplementary Figure A.7a**. The three line plots adjacent to the Hi-C maps represent HiDENSEC inferred copy numbers for the three chromosomes, which were used to determine the contacts constituting this complex structural variant.

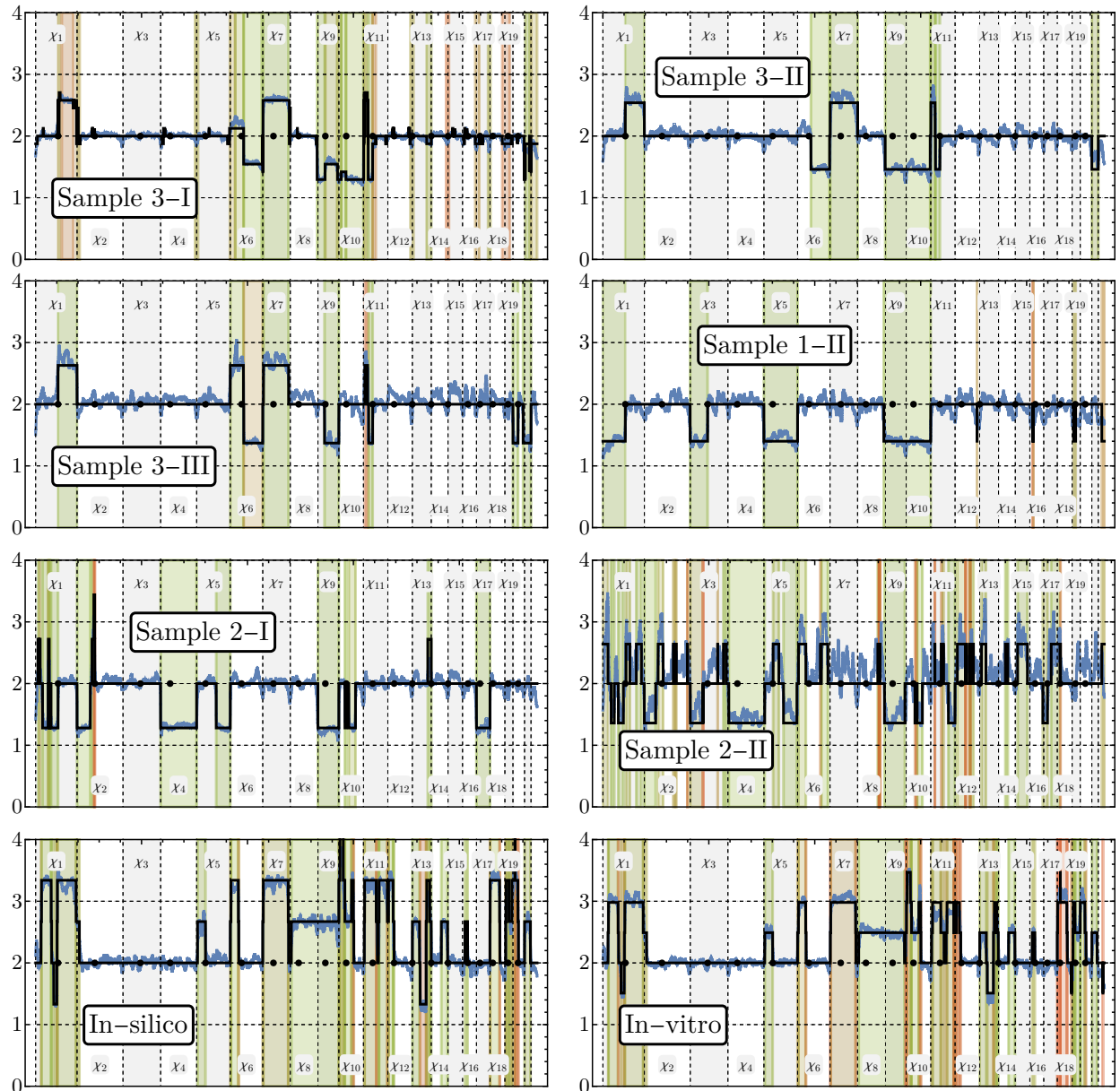


Supplementary Figure A.11: **Covariate correction is protocol-dependent.** A Hi-C sample of the reference genome as well as the Fix-C Sample 3-II illustrate the necessity for both covariate correction in general, as well as its protocol-specific nature.

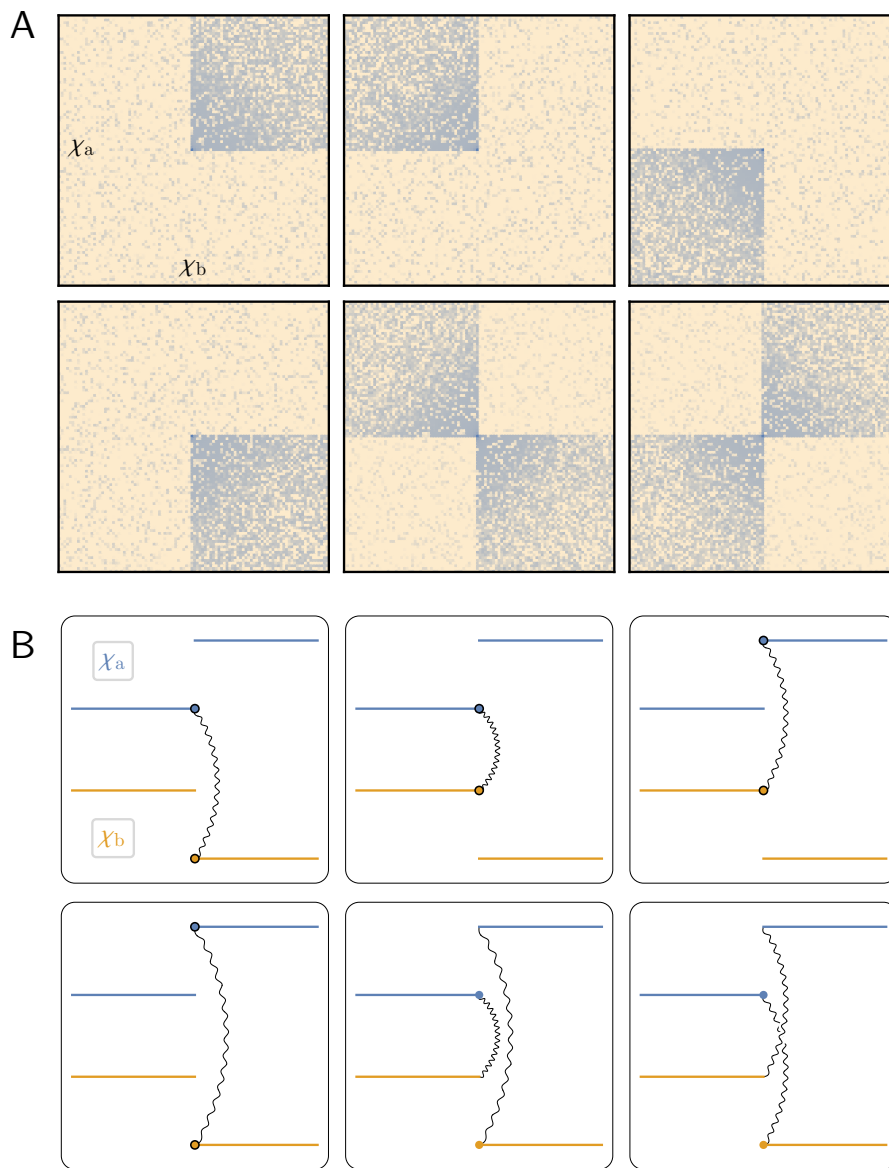


Supplementary Figure A.12: **Generative model of signal & noise.** Observed on-diagonal contact intensities are modeled as an underlying effective copy number profile comprised of a convex combination of individual, cell-population-specific absolute copy number profiles (A), which is perturbed by heteroskedastic noise (B) and scaled by a generally unknown constant  $C_0N$ . Under  $\mathcal{H}_0$  of  $\pi \equiv 2$ ,  $p$ -values associated with HiDENSEC's test statistics behave super-uniformly or close to uniform (C).

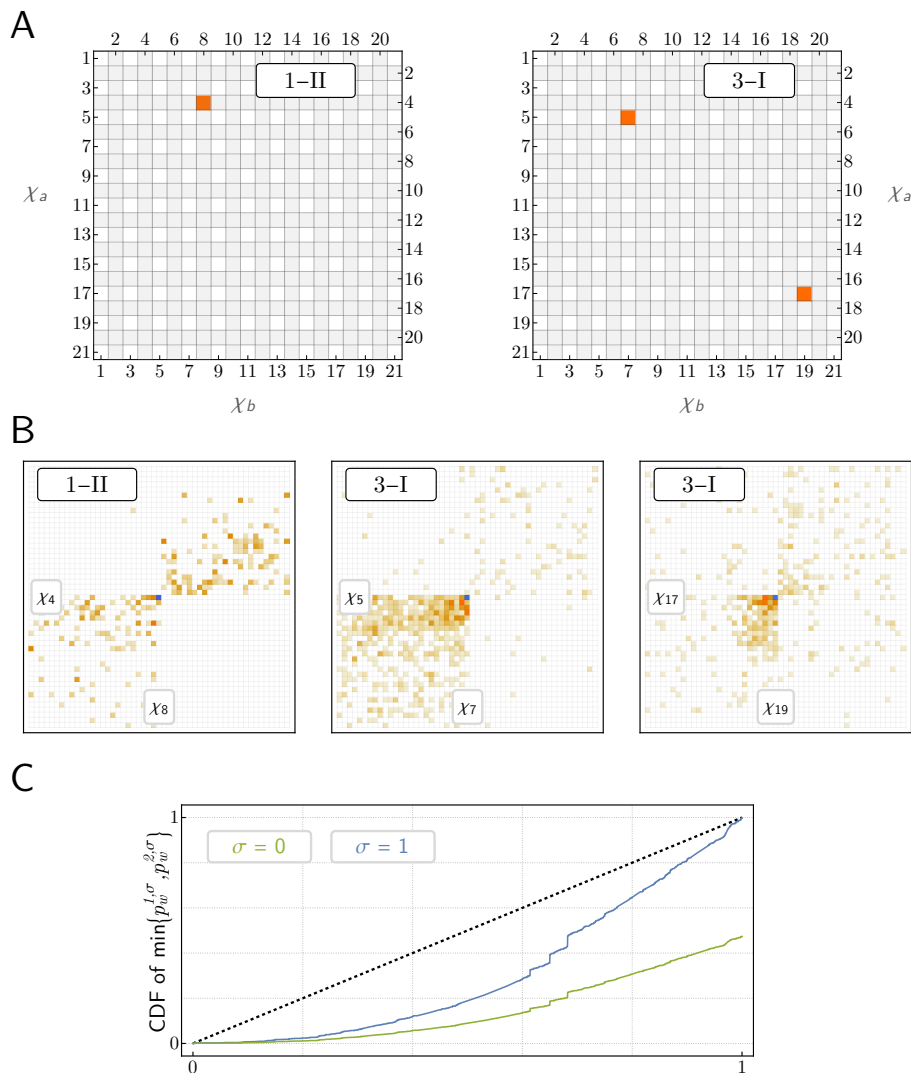




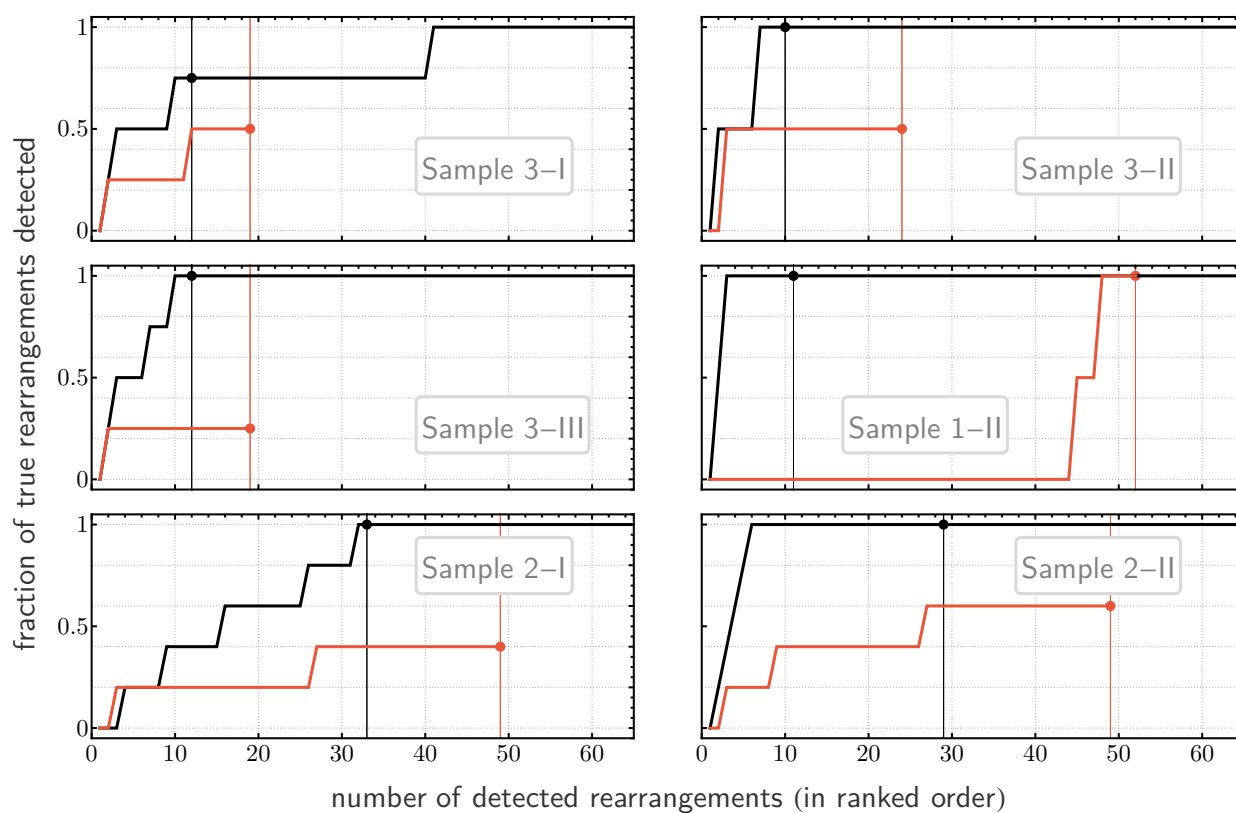
Supplementary Figure A.13: **Inference of effective copy number profiles & interpretation of excursions.** Inferred  $\hat{\pi}$  (solid black line) for each of the non-diploid samples discussed in the main text as well as for in-silico and in-vitro mixtures (with  $f = 0.7$  and  $f = 0.5$ , respectively) are shown against  $\Pi$  (blue line). Each excursion  $e$  is associated with a  $p$ -value reflecting its biological significance, with greener colors mirroring higher significance.



Supplementary Figure A.14: **Hi-C intensity patterns and associated large-scale structural variants.** HiDENSEC detects Hi-C sub-matrices of six distinct patterns (**A**) associated with six types of large-scale structural variants (**B**) (note: non-fusing segments may interact with chromosomes other than  $\chi_a$  and  $\chi_b$  or be deleted without qualitatively affecting the local Hi-C patterns of (A)).



Supplementary Figure A.15: **HiDENSEC reliably detects off-diagonal exchange patterns.** In those samples that do contain patterns in  $\mathcal{P}_2$ , HiDENSEC correctly recovers them at zero false-positive rate (**A**), and identifies their precise locations accurately (**B**, blue highlights indicate fusion sites inferred by HiDENSEC). Calibration of HiDENSEC is primarily a result of computed  $p$ -values behaving super-uniformly (**C**, empirical distributions based on all samples analyzed in the main text). Of the three events, HiNT only detected the  $\chi_4 \sim \chi_8$  fusion, locating it, however,  $\approx 42$  and  $\approx 25$ MB away from the true signal on  $\chi_4$  and  $\chi_8$ , respectively.



Supplementary Figure A.16: Comparison of HiDENSEC's (black) and HiNT's (red) top- $k$  recall on the samples analyzed in the main text. As in the corresponding main figure, filled regions indicate rearrangements deemed significant by either method.

## A.7 Supplementary Tables

Supplementary tables associated with this work can be found here:

<https://tinyurl.com/HiDENSECSupplementalTables>