Untangling the complexity of nature: Machine-learning for accelerated life-sciences

by

Adam U. Yaari

BSc, Bar-Ilan University (2016)
MSc, Weizmann Institute (2018)
S.M., Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

Author .....
Department of Electrical Engineering and Computer Science
December 31, 2022

Certified by .....
Boris Katz
Principal Research Scientist, Computer Science and Artificial Intelligence Lab
Thesis Supervisor

Certified by .....
Bonnie Berger
Simons Professor of Mathematics
Thesis Supervisor

Certified by .....
Andrei Barbu
Research Scientist, Computer Science and Artificial Intelligence Lab
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Untangling the complexity of nature:

Machine-learning for accelerated life-sciences

by

Adam U. Yaari

Submitted to the Department of Electrical Engineering and Computer Science
on December 31, 2022, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY

Abstract

The fundamental understanding of living processes is one of the main pillars in modern medicine and technology. Biological mechanisms are convoluted and stochastic systems that remain largely misunderstood despite centuries of rigorous scientific work. In recent years, machine-learning (ML) has resurfaced as a powerful framework to identify patterns of interest in complex datasets. Yet, the impact of such methods remains limited in the broad context of life-sciences. This work optimizes the utility of ML to accelerate research of fundamental biological problems. First, we propose a paradigm shift from siloed data curation to multi-purpose cohorts at scale, even in the most restrictive case of human experimentation. The potential of this approach is revealed through the Brain TreeBank, a multi-modal dataset of naturalistic language aligned to intracranial neural recordings. The TreeBank provides the resolution and breadth required to probe the spatio-temporal dynamics of language context dependence and representation in the brain. Second, we argue for the importance of ML interpretability to accelerate the understanding of biology. We develop an explainable general-purpose tool for modeling discrete stochastic processes at multiple resolutions with output certainty estimation. We demonstrate the utility of the method by modeling patterns of somatic mutations across the entire cancer genome and extend it to map mutation rates in 37 types of cancer. The confidence intervals and increased sensitivity of the method identify sets of mutations that likely drive cancer growth in both coding and noncoding regions of the genome. Broadly, this work demonstrates how computational approaches can overcome unique challenges in biological data and how biological problems can drive advances of computational methodologies.

Thesis Supervisor: Boris Katz
Title: Principal Research Scientist, Computer Science and Artificial Intelligence Lab

Thesis Supervisor: Bonnie Berger
Title: Simons Professor of Mathematics

Thesis Supervisor: Andrei Barbu
Title: Research Scientist, Computer Science and Artificial Intelligence Lab

Acknowledgments

A PhD is a race against yourself towards an unknown finish line, but escorted by the right people it becomes a journey of a lifetime. I owe a great deal to those who supported, guided and encouraged me throughout this unforgettable adventure. Chief among these are my incredible supervisors, Boris Katz, Andrei Barbu and Bonnie Berger. I was privileged to have you as mentors and friends in my admittedly capricious exploration of research interests. Thank you for your unwavering belief in me.

I would also like to express my deepest appreciation to Gabriel Kreiman for his substantial contribution to my work through his guidance and non-formal supervision. Another special shout out to my great office mates - Yen-Ling Kuo, Candace Ross, and Ignacio Cases. Thank you for being there through the good and bad.

Thank you to all of my many co-authors who contributed in innumerable ways to the research presented below. These are, in no particular order, Oliver Priebe, Christopher Wang, Aaditya Singh, Vighnesh Subramaniam, Yevgeni Berzak, Helena Aparicio, Felix Dietlein, Jan DeWitt, Sue Felshin, Henry Hu, and Bennett Stankovits.

Thank you to the lab members that have taught me a great deal, but with whom I did not have the pleasure to publish. These include (but are not limited to) Dana Rosenfeld, Ravi Tejwani, Rohit Singh, Alex Wu, David Mayo, and Julian Alverio.

A special thank you to my very dear friend Max Sherman, whom I met on my first day at MIT and has since been an inseparable part of my personal and professional life. Thank you for being a wonderful colleague, teacher and a friend.

Thank you to my supportive parents Nissan and Yael, and my amazing sister Shachar. It is the steps that you laid out that allow me to transcend beyond my wildest dreams.

More than all, I want to thank my incredible family Gal, Ethan and Mowgli for always being there for me. Gal, thank you for your never ending love, friendship, and support. You are my brightest light in the darkest moment. Thank you for being a trailblazing women, fantastic mom, and my best friend.

Finally, I would like to recognize those who made me a better man despite crossing my path for but a moment. Specifically, I want to thank Yair Shiran (R.I.P.). Your fortitude, benevolence, and devotion will forever be an inspiration.

Contents

1 Introduction **21**

1.1 Machine-learning for biological science 21

1.2 Machine-learning biology-specific challenges 22

1.3 Contributions of this research 24

1.4 Roadmap of thesis 26

2 Background **29**

2.1 Neural language processing 29

2.1.1 The hierarchy of language 29

2.1.2 Relevant principles of neural processing 30

2.1.3 Approaches for real-time brain measurement 31

2.1.4 Universal Dependency language formalism 32

2.2 Genetic Determinants of Cancer 32

2.2.1 DNA mutations in human genetics 32

2.2.2 High-throughput sequencing for mutation detection 33

2.2.3 Relevant principles of cancer genetics 34

2.2.4 Relevant principles of epigenetics 34

3 The Aligned Multimodal Movie Treebank: an audio, video, dependency-parse treebank **37**

3.1 Summary 37

3.2 Introduction 38

3.3 Dataset 39

3.3.1 Transcription pipeline 40
3.3.2 Dependency parsing pipeline, annotation and validating annotator performance 42
3.3.3 Performance of existing parsers 43
3.4 Multimodal feature analysis 44
3.5 Tools 45
3.6 Conclusion 47

4 Neural processing of nouns and verbs with large-scale intracranial recordings from naturalistic language 49
4.1 Summary 49
4.2 Introduction 50
4.3 Results 51
4.3.1 Brain TreeBank: a large-scale intracranial naturalistic language dataset 52
4.3.2 How POS neural responses differ? 53
4.3.3 How is the neural POS network distribute? 57
4.3.4 What are the dynamics of POS processing? 60
4.4 Discussion 63
4.5 Materials and Methods 65
4.5.1 Dataset construction 65
4.5.2 Task and stimuli 65
4.5.3 Data acquisition and signal processing 66
4.5.4 Word responsive electrode selection 67
4.5.5 Mean signal peak analysis 67
4.5.6 Generalized linear model 68
4.5.7 Convolutional neural network model 69
4.5.8 Test set construction 71
4.5.9 Decoding significance assessment 72
4.5.10 Held-out trial analysis 74

4.5.11 Functional connectivity analysis 74
4.6 Extended Data Figures 75

5 Multi-resolution modeling of a discrete stochastic process identifies causes of cancer 79

5.1 Summary 79
5.2 Introduction 80
5.2.1 Previous work 81
5.2.2 Our contributions 82
5.3 Materials and Methods 83
5.3.1 Multi-resolution modeling of a non-stationary discrete stochastic process 83
5.3.2 Fitting parameters to predict cancer mutation patterns 86
5.3.3 Identifying genetic drivers of cancer 90
5.4 Results 91
5.4.1 Accuracy of regional rate parameter estimation 91
5.4.2 Accuracy and efficiency of mutation rate prediction 92
5.4.3 Identification of cancer driver mutations 94
5.5 Discussion 95

6 Genome-wide mapping of somatic mutation rates uncovers drivers of cancer 97

6.1 Summary 97
6.2 Introduction 98
6.3 Results 100
6.3.1 Testing mutational excess with probabilistic deep learning . . 100
6.3.2 Small mutation sets increase power to identify drivers 102
6.3.3 Quantifying pan-cancer selection on cryptic splice SNVs 105
6.3.4 Noncoding candidate cancer driver mutations in 5' UTRs . . . 107
6.3.5 The shared landscape of common and rare driver genes 110
6.4 Discussion 113

6.5 Materials and Methods . . . 116
6.5.1 Sequencing data curation . . . 116
6.5.2 Identification of mutational excess with probabilistic deep learning . . . 118
6.5.3 Comparison to existing driver detection methods . . . 121
6.5.4 Power analysis . . . 122
6.5.5 Quantifying selection on cryptic splice SNVs . . . 123
6.5.6 Quantifying mutational excess in promoters and 5' UTRs . . . 125
6.5.7 Driver gene prediction in WES & targeted sequenced samples 126
6.6 Extended Data Figures . . . 127

7 Conclusion 133

A Supplementary Information Related to Chapter 3 137

B Supplementary Information Related to Chapter 4 141
B.1 Supplementary Methods . . . 141
B.1.1 Cortical surface extraction and electrode visualization . . . 141
B.1.2 Audio transcription and alignment . . . 142
B.1.3 Part of speech tagging . . . 143
B.1.4 Confounding features . . . 143
B.2 Supplementary Figures . . . 145
B.3 Supplementary Tables . . . 145

C Supplementary information related to Chapter 5 149
C.1 Supplementary Materials and Methods . . . 150
C.1.1 Data . . . 150
C.1.2 Graphical model derivation . . . 153
C.1.3 Overview of parameter estimation procedure . . . 155
C.1.4 Regional parameters estimation methods . . . 155
C.1.5 Empirical variance estimation . . . 159
C.1.6 Performing a genome-wide search for cancer driver mutations . 160

C.1.7 Environment and compute time . . . 161
C.2 Supplementary Results . . . 162
C.2.1 Negative Binomial Regression does not detect well-known drivers genome-wide . . . 162
C.2.2 Convolutional neural network outperforms other dimensionality reduction alternatives for a Gaussian process . . . 162
C.2.3 Existing whole-genome regression models are time inefficient at multi-resolution search . . . 163

D Supplementary information related to Chapter 6 169
D.1 Supplementary Results . . . 169
D.1.1 Insights into mutation rate prediction accuracy from feature maps 169
D.1.2 Comparison of cancer driver detection methods . . . 170
D.1.3 Additional details on alternative splicing analysis with LeafCutter 171
D.1.4 Investigation of mutational burden in *ELF3* 5' UTR . . . 171
D.1.5 Functional correlates of mutations in rare driver genes . . . 173
D.1.6 Preliminary analysis of enhancer networks . . . 174
D.2 Supplementary Methods . . . 175
D.2.1 Technical details of Dig's deep-learning framework . . . 175
D.2.2 Technical details of Dig's probabilistic graphical model . . . 177
D.2.3 Associating epigenetic structure to mutation density with feature maps . . . 179
D.2.4 Additional details about the comparison of mutation rate models 181
D.2.5 Details about the comparison of driver element detection methods 183
D.2.6 Constructing a genome-browser of genome-wide mutation rate estimates . . . 185
D.2.7 Details about power analysis . . . 185
D.2.8 Additional details about quantifying selection on cryptic splice SNVs . . . 186
D.3 Supplementary Figures . . . 190

D.4 Supplementary Tables 204

List of Figures

3-1 Aligned Multimodal Movie Treebank overview 38

3-2 Efficient Audio Alignment Annotator snapshot 40

3-3 Aligned Multimodal Movie Treebank performance evaluation 44

3-4 Noun-object agreement across COCO DB 46

4-1 Brain TreeBank overview 52

4-2 Noun-verbs neural responses 55

4-3 Generalized linear model analysis of language unique features in the brain 56

4-4 POS functional connectivity map analysis 59

4-5 POS CNN dynamic decoding 62

4-6 Brain TreeBank confounds elimination overview 75

4-7 Nouns-verbs average signal peak attributes comparison across sentence progression 77

4-8 Brain TreeBank all electrode locations 77

5-1 Non-stationary stochastic process modeling predicts mutation patterns and identifies cancer-specific driver mutations. 81

5-2 Data simulation and regional parameters inference accuracy across methods. 87

5-3 SPG accurately models mutation density and detects driver events. . 92

6-1 Modeling the genome-wide neutral somatic mutation rate and identifying cancer driver elements. 103

6-2 Evidence of positive selection on intronic cryptic splice SNVs in tumor suppressor genes. 108
6-3 Enrichment of somatic mutations in the 5' UTRs of *TP53* and *ELF3*. 111
6-4 Enrichment of protein-altering SNVs in "long-tail" genes reveal a shared landscape of common and rare driver genes 114
6-5 Detailed overview of the Dig model. 128
6-6 Epigenetic input features used by Dig to predict mutation density in nine cancer types. 129
6-7 Cryptic splice SNV enrichment in oncogenes and genes not in the CGC. 130
6-8 Examples of distribution of activating mutations in gene-tumor pairs. 131

A-1 Aligned Multimodal Movie Treebank sentence length distribution . . 137
A-2 COCO classes noun-object agreements per movie 138
A-3 EWT POS frequency comparison to AMMT 139

B-1 Neural POS connectivity map example 145

C-1 Distribution of mutation counts in 50kb windows tiled across the genome. 152
C-2 Model robustness to region size. 161
C-3 μ_R and σ_R^2 estimation accuracy. 164
C-4 NBR and RF R^2 accuracy. 165
C-5 Mean (μ_R) vs variance (σ_R^2) for three cancer types. 165
C-6 Observed versus expected mutation counts based on sequence context alone. 166
C-7 Log-log run-time of current whole-genome regression methods. 167
C-8 -log_{10}(P-value) quantile-quantile (qq) plots for expected vs observed number of mutations. 167

D-1 Plate diagram of the probabilistic model that Dig uses to model the number of neutral mutations (M_i) in an element of interest 191
D-2 Comparison of variance explained of SNV counts across methods, annotations, and cohorts. 192

D-3 Precision-recall comparison of gene driver methods in the PCAWG cohort. 193
D-4 Approximate number of false-positive and true positive driver genes identified from 15 whole-exome sequenced cohorts from Dietlein et al. 194
D-5 Precision-recall comparison of noncoding driver detection methods in the PCAWG dataset. 195
D-6 Simulated power to detect driver elements in a pan-cancer cohort by sample size and by size of the elements being tested. 196
D-7 Proportion of excess protein-altering SNVs in TSGs. 197
D-8 SNV enrichment (with 95% CI) and excess analysis excluding samples with >3000 coding mutations. 198
D-9 Estimated SNV enrichment with 95% CI in tumor suppressor genes and oncogenes. 199
D-10 Additional predicted cryptic splice SNV carriers in which LeafCutter identified strong evidence of alternative splicing. 200
D-11 Normalized expression of *TP53* stratified by the type of mutation individuals carry in *TP53*. 201
D-12 Evaluation of neutral mutation model for ten solid cancer megacohorts. 202
D-13 Estimated excess activating SNV rate in oncogenes and TSGs. 203

List of Tables

3.1 Aligned Multimodal Movie Treebank statistics 41

3.2 Aligned Multimodal Movie Treebank inter-annotator agreement . . . 42

4.1 Brain TreeBank confounds description 76

5.1 Comparison of run times by method and region size. 94

6.1 Proportion of variance in observed SNV counts in the PCAWG pan-cancer cohort (n=2,279 samples) explained by different methods. . . . 104

A.1 Aligned Multimodal Movie Treebank POS tags distribution 138

A.2 Aligned Multimodal Movie Treebank movie statistics 139

B.1 Brain-TreeBank subject statistics 146

B.2 Brain-TreeBank movie statistics 147

C.1 All 50bp windows with significant recurrent mutations in the *TP53* gene from genome wide driver search in esophageal adenocarcinoma. . 168

D.1 Information about the 37 PCAWG cancer cohorts used to train Dig's mutation rate models. 205

D.2 Variance explained of SNV counts in 10kb regions in held-out test data per fold. 206

D.3 Variance explained of SNV counts in tiled regions by method. 207

D.4 Variance explained of SNV counts in genes by method. 208

D.5 Variance explained of SNV counts in enhancers and noncoding RNAs. 209

D.6 Accuracy measures of driver gene detection methods at FDR<0.1. . 210
D.7 Areas under the approximated ROC curves of Fig fig. 6-1e and Supplementary Fig fig. D-4. 211
D.8 Accuracy measures of noncoding driver detection methods at FDR<0.1 for Dig. 212
D.9 Accuracy measures of noncoding driver detection methods at FDR<0.1 for DriverPower. 213
D.10 Accuracy measures of noncoding driver detection methods at FDR<0.1 for ActiveDriverWGS. 214
D.11 Accuracy measures of noncoding driver detection methods at FDR<0.1 for Larva. 215
D.12 SNV enrichment across coding and cryptic splice sites in the PCAWG pan-cancer cohort. 216
D.13 Tumor suppressor genes with a FDR<0.1 significant burden of intronic cryptic splice SNVs. 217
D.14 Evidence of cryptic splice events in RNA-seq data. 218
D.15 Genes not in the CGC with a FDR<0.1 significant burden of intronic cryptic splice SNVs. 219
D.16 Recurrent predicted cryptic intronic splice mutations in genes in Supplementary Table table D.15. 220
D.17 Enrichment of mutations in TP53 and ELF3 5'UTRs in the PCAWG pan-cancer dataset. 221
D.18 Enrichment of mutations in the ELF3 5'UTR in the Hartwig Medical Foundation pan-cancer dataset. 222
D.19 Metadata on mega-cohorts of targeted and whole-exom sequenced samples. 223
D.20 Burden of activating mutations in long-tail oncogenes in mega-cohorts. 224
D.21 Burden of pLoF mutations in long-tail tumor suppressor genes. . . . 225
D.22 Genes not in driver gene databases with a FDR<0.1 significant burden of pLoF mutations in exome-sequenced samples. 226

D.23 ABC enhancer elements with a FDR<0.1 significant burden of mutations in the PCAWG pan-cancer cohort. 227
D.24 Metadata about whole-exome sequenced cohorts from Dietlein et al. 2019 Nat. Genet.. 228

Chapter 1

Introduction

1.1 Machine-learning for biological science

The core theme of this work is the exploration and development of computational approaches to accelerate life-science research. Computational tools, such as machine learning (ML), have been rapidly evolving in recent decades with the exponential growth of computing power [230]. However, other research domains, that can benefit from such advancements, have yet to leverage the full potential of these technologies [223, 111, 188]. Specifically, biological sciences, the branch of natural sciences that deals with the processes, structure, organization, and interactions of living organisms [181], can gain immensely from methods to augment the logical reasoning of the practicing scientist.

The fundamental understanding of biological processes is key to progress in our understanding in healthcare, agriculture, environmental studies and more. Through the discovery of Penicillin [90] to the engineering of live cells [63] advances in life-sciences and medical research have nearly doubled the human life span in the past century [158].

However, biological systems' hierarchical complexity has no counterpart outside the realm of biology [266]. Unlike the logical structure of human-made machines, natural systems are simultaneously convoluted, non-linear, stochastic, and multi-dimensional [7, 5, 6, 174, 100, 144]. Thus, making them extremely complex to decipher

and mandating slow progress through experimentation and rigorous reasoning. Unraveling this complexity is therefore of the utmost importance for solving life-sciences' most urgent problems and pushing technology beyond its current limitations.

As datasets grow larger, one approach to unravel this complexity is through the use of ML frameworks [269], suited to identify meaningful information in large-scale data. ML methods have proven to be powerful tools in a variety of different domains [45, 95, 44, 202], with super-human performance in a number of seminal cognitive tasks [125, 238, 145, 49]. Specifically, ML methods are uniquely suited to identify subtle recurring patterns in a plethora of data, often intractable by direct computation or human reasoning [32]. Additionally, ML approaches do not depend on predefined rules that tend to be sub-optimal and limit the hypothesis space. More recently, different ML models have demonstrated impressive results across a variety of biological domains [58, 153, 102, 127, 160, 134]. However, such examples are the exception that proves the rule, as many biological datasets tend to contain smaller sample sizes, complex structures, and more noise with respect to the more classic ML tasks with human-made datasets [223].

This work will focus on approaches to improve ML methodologies to overcome the subset of challenges that are abundant in the natural sciences but do not carry the same weight in the well-defined seminal problems of modern artificial intelligence.

1.2 Machine-learning biology-specific challenges

The innate complexity of natural systems put forth a number of biology-specific limiting factors for current ML solutions. Some biological and medical domains, like radiology or pathology, are easier to reduce to computational reasoning due to their similarity to classical ML problems (e.g. computer vision, natural language processing, etc.). Others, require only marginal computational assistance to perform a human supervised task, like robotic surgery or electronic health record analysis, lowering the performance bar to an attainable threshold. However, the vast majority of life-science research problems are not reducible to these two sub-classes and are

therefore hindered by a number of key drawbacks.

The first and foremost challenge in the space is data acquisition. While cheap and accessible DNA sequencing techniques have revolutionized genomics [233], it is still an outlier in the space. The study of most biological systems requires real-time signal recordings that are either resource-consuming or often technically infeasible. Data restrictions are not only limiting the prospective benefits of ML, but the pace at which new discoveries can arise. For instance, neuroscience research has long been striving for bona fide datasets of human neural measurements, paramount to explore multiple cognitive behaviors and neural diseases [150, 94, 57]. However, the inherent risks involved have pushed the field to a dismal choice between poor resolution and highly specific datasets. The former provides low signal-to-noise ratios, increasing false discovery rates and minimizing the scope of questions to be studied. The latter produce siloed datasets, confined to narrow research questions, limiting the complexity of studies and preventing reproducibility of results. This work will address this issue by proposing the alternative approach of hypothesis-free data curation by exploring the question:

[A] Can multi-purpose datasets at scale accelerate scientific discovery over complex biological systems?

The second fundamental challenge is the black-box nature of the most commonly used ML modality – deep learning. Unlike image processing or language understanding tasks, where *why* and *how* are nice-to-have features on top of systems' accuracy, in biology such questions are arguably the core of the study itself. No new science can be learned if we cannot query the model for *why* a prediction was made [108], and no follow-up decisions can be achieved without an estimate of *how* confident the model is in their estimation [147]. Taken together, these questions boil down to the interpretability of ML models processing and the confidence likelihood of their final outcome. While some progress has been made in recent years, bioinformatics is still largely governed by simpler models that fall short of the rapidly evolving state-of-the-art technologies. This work will demonstrate the importance of incorporating

interpretability mechanisms by suggesting a modular likelihood estimation and input space reasoning framework. Through this development, this thesis will explore the question:

[B] Can improved ML interpretability drive novel discoveries in well-studied domains?

1.3 Contributions of this research

The first part of this work aims to answer question [A] through two major contributions to the study of neurolinguistics. Both rely on the availability of intracranial recordings from intractable epilepsy patients (see chapter 2), which provide unparalleled temporal and spatial resolutions of neural signal recordings [186]. An invaluable resource that has been underutilized thus far due to the limited scale and scope of the data collected from any given patient [220, 82, 205, 113]. This thesis presents the first large-scale naturalistic language dataset with invasive stereoelectroencephalography (SEEG) recordings. Specifically, this work:

Introduces the Brain TreeBank – the largest collection of neural recordings aligned to annotated language.

The Brain TreeBank is a first-of-a-kind dataset, surpassing in scale both its intra and extra-cranial dataset predecessors. It presents:

- **Unprecedented scale:** recordings of 236,400 annotated tokens across 10 subjects, 10 times larger compared to other naturalistic language datasets [137, 28, 105] and 100 times larger compared to the more common controlled studies [80, 79, 260, 82, 23].
- **Augmented flexibility:** multi-modal audio and visual streams aligned to conversational language, parsed in the Universal Dependencies (UD) formalism [193].

- **Supporting analysis tools:** a broad battery of methods to eliminate confounds and provide a structured experiment level of control.

(Definitions are provided in chapter 2 for readers unfamiliar with these concepts).

The second contribution to question [A] is utilizing the Brain TreeBank, curated as a multi-purpose resource for the study of language in the brain, to probe the neural representation of part of speech (POS). Specifically, this work aims to show the applicability of large-scale hypothesis-free datasets to complex systems by presenting new evidence on:

- **The difference in neural activation patterns for words with distinct POS.**
- **The network of brain regions involved in POS processing.**
- **The temporal dynamics of the POS processing network.**

The second part of this work aims to answer question [B] through the development of a novel method for modeling somatic mutation patterns (see chapter 2). Specifically, this work:

Develops a probabilistic deep-learning approach to model the patterns of somatic mutations genome-wide in a tissue-specific manner.

This method builds upon a successful history of prior works that modeled patterns of somatic mutations in specific regions of the genome [151, 169, 71, 187, 161, 236].

This computational contribution is a means to an end. That end is demonstrating the utility of interpretability and confidence estimation methods, even in the extreme case of well-studied fields like somatic mutations in cancer. The probabilistic model uncovers novel understandings in cancer research and highlights the input features used to infer its outcome. Specifically,

- **Somatic mutations outside of protein-coding sequences (noncoding mutations) can serve as high-impact drivers of cancer.**

- **Proteins that frequently drive one type of cancer can act as rare drivers of numerous other types of cancer.**
- **Highly localized chromatin markers govern the likelihood of somatic mutation accumulation.**

1.4 Roadmap of thesis

Given the strong biology focus of this thesis, chapter 2 provides a brief primer on the central concepts used in later chapters. While this background is by no means comprehensive, it is designed to provide sufficient detail for readers with less biological familiarity to follow (more or less) in chapters 3-6.

In chapter 3, we present the Aligned Multimodal Movie Treebank (AMMT), an English language treebank derived from dialog in Hollywood movies which includes transcriptions of the audio-visual streams with word-level alignment and UD parsing. This chapter introduces the groundwork and analysis required to build a large-scale multi-modal treeabank. It presents an overview of the dataset and tools developed to curate it, as well as a collection of statistics for quality evaluation. The work presented in this chapter was originally published in

Adam Yaari, Jan DeWitt, Henry Hu, Bennett Stankovits, Sue Felshin, Yevgeni Berzak, Helena Aparicio, Boris Katz, Ignacio Cases and Andrei Barbu. The Aligned Multimodal Movie Treebank: An audio, video, dependency-parse treebank. ***The Conference on Empirical Methods in Natural Language Processing (EMNLP)*****, 2022.** [4]

In chapter 4, we deploy the AMMT to curate the Brain TreeBank and explore the POS processing in the human brain. The chapter reinterprets the AMMT in the context of its alignment to SEEG recordings and thoroughly explains any additional steps. chapter 4 demonstrates how multi-purpose data at scale can enable the rigor of a quasi-structured experiment and the variety of outcomes it can produce. This chapter presents in-depth description of the methods used to analyze such an unstructured

resource and novel insights into the neural representation of a fundamental language feature.

At the time of writing these lines, this work is yet to be published. It is currently referenced as:

Adam Yaari, Aaditya Singh, Ignacio Cases, Vighnesh Subramaniam, Pranav Misra, Joseph Madsen, Sceillig Stone, Gabriel Kreiman, Boris Katz, and Andrei Barbu. Neural processing of nouns and verbs with large-scale intracranial recordings from naturalistic language.

In chapter 5, we develop a general method to model discrete stochastic processes at multiple resolutions in a computationally efficient manner. We demonstrate the application of this method to model patterns of somatic mutations anywhere in the genome. This work presents a functionality-enhanced ML model for multiple time-series biological tasks, with a unique focus on cancer biology. This chapter was originally published in:

Adam Yaari, Maxwell Sherman, Oliver Priebe, Po-Ru Loh, Boris Katz, Andrei Barbu, and Bonnie Berger. Multi-resolution modeling of a discrete stochastic process identifies causes of cancer. ***International Conference on Learning Representations*****, 2021.** [270].

In chapter 6, we further extend and apply the method from chapter 5 to identify mutations genome-wide that may contribute to the etiology of cancer. chapter 6 is largely an extensive application of the method developed in chapter 5. The work presented in this chapter was originally published in:

Maxwell Sherman, Adam Yaari, Oliver Priebe, Felix Dietlein, Po-Ru Loh, and Bonnie Berger. Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. ***Nature Biotechnology*****, 2022.** [234]

Extensive additional details on results and methods are provided for the motivated reader in appendix A, appendix B, appendix C, and appendix D.

Chapter 2

Background

2.1 Neural language processing

2.1.1 The hierarchy of language

Language has been argued to be the most important evolutionary development for human survival. It is the primary tool we use for expression and communication. We use it constantly and effortlessly. However, it is far from a trivial process. Language is the unique process of transforming a linear input (auditory or visual) into a complex semantic representation of meaning. While some researchers argued recently that artificial models can mimic this behavior, the human brain remains the only system to flexibly produce and comprehensively process language in its broad form [142].

While produced and consumed linearly, the structure of human language is hierarchical by nature. In verbal language processing, an incoming audio stream is broken down into phonemes, such as consonants and vowels. The phonemes are then compiled into minimal logical units, called morphemes, that in turn combine into words, phrases, and whole sentences. There is a plethora of opinions and longstanding debates on the exact logic and set of rules governing this hierarchy. Albeit, the overall consensus is that language is compositional and therefore, so is its one and only processing system – the brain [205].

Language composition theories span two main dimensions. The first is *syntax* vs

semantics, with the former focusing on grammatical rules and the latter on meaning. The second is the specific set of rules used to compile units under the syntactic or semantic assumption. The rules and entities may vary between different languages. However, they all share a small subset of universal components, common across all languages. One example for such universals are *nouns* and *verbs*. Often overlapping the semantic sets of *objects* and *actions*, nouns and verbs enable unbiased hypothesis testing of language processing.

2.1.2 Relevant principles of neural processing

To understand the complexity of how the brain receives sensory input, processes information, produces thought and generates action, one must first understand the functionality of a single nerve cell. Each neuron is a computation unit able to receive input from hundreds to hundreds of thousands (1,000 on average) of its neighbors, aggregate the information and transmit it to multiple of its neighbors (or even itself). While the synaptic interaction between neurons is chemical, the internal transmission of a signal across the neuron is electrical and therefore measurable.

The brain is anatomically divided into hemispheres, lobes, regions, and sub-regions. While some brain functions are loosely associated with certain regions (e.g. sensory-motor, vision, hearing, and even fear), many brain areas are multi-functional with most functionalities regarded to be distributed. Furthermore, complex functions are typically hierarchical across different brain regions, compiling the bigger picture from smaller pieces of the puzzle. For instance, the visual cortex progresses from pixels at the first visual processing core (V1) and lines at V2 to objects at V4. It is now known that the linear hierarchy assumption is an oversimplification of an extremely complicated system, but for the purpose of this work, it will suffice. The study of brain systems and their associated regions has enabled novel therapies and treatments, improved mechanistic understanding of the biological system and even breakthrough technologies like convolutional neural networks.

Unlike sensory systems, language has been associated with an assembly of regions across the brain. While some are robust and reproducible across studies (e.g. Broca's

and Wernicke's areas), most are weakly correlated and were found to correspond to vague components of language processing. For example, the angular gyrus was associated with sensitivity to argument structure [205]. A better understanding of language representation and processing dynamics in the brain can reshape our perception of cognition, help treat communication impairments, resolve centuries-long debates, and power computational language processing as a whole.

2.1.3 Approaches for real-time brain measurement

Deconvolving functionality from the activity of billions of hyper-connected neurons is an onerous task that requires precise measurements. Neural measurement tools can be defined by three distinct features: spatial resolution (number of neurons measured per sensor), temporal resolution (sensor sampling rate), and coverage (percentage of the brain simultaneously measured). Existing techniques fall into one of two categories: intra (inside) or extra (outside) cranial (the scalp) recordings.

Intracranial recordings capture the electrical signaling between neurons at high spatial and temporal resolutions but are limited in coverage. Most typically, an electrode will sample every 0.5 seconds at 1-3 millimeter scale. These resolutions capture local activity (down to a single spike) of hundreds to a few thousands of neurons. This resolution is currently as good as it gets in human recordings. Some methods can achieve even a single neuron resolution (e.g. patch-clamp); however, their restriction to animal studies places them outside the scope of this work. The most commonly used human intracranial recordings are electrocorticogram (ECOG), a multi-electrode grid placed on the surface of the dura mater (the covering surface of the brain beneath the scalp), and streoelectroencephalography (SEEG), multi-electrode wires inserted into the depth of the brain. This study relies on SEEG recordings, which simultaneously measure multiple layers of the brain.

Extracranial recording techniques capture either electrical activity (EEG), magnetic activity (MEG), or blood-oxygen-level-dependent (BOLD) signals (fMRI). All three approaches provide full brain coverage but compromise on resolution. EEG and MEG measure at a milliseconds sampling rate, but with the scalp acting as a

filter, both must average the activity of millions of neurons at each sensor. fMRI on the other hand can achieve an order of magnitude better spatial resolution but has a sampling frequency of seconds due to the nature of the measured BOLD signal.

2.1.4 Universal Dependency language formalism

Universal Dependencies (UD) is an international cooperative project to create treebanks of the world's languages [68]. UD is widely applicable in NLP, primarily in the study of syntax and grammar. The project's main aim is to achieve cross-linguistic consistency of annotation while permitting language-specific extensions when necessary.

UD cohorts are collectively recognized as *treebanks*. Each treebank consists of sentences parsed into trees based on the UD scheme. Each tree begins with a root (typically the main verb), and connects all words of the sentence with dependency edges, labeled by one of 45 syntactic functions (not including the root). Each word in the tree has a POS tag and an incoming dependency edge, defining its grammatical relation and role.

2.2 Genetic Determinants of Cancer

2.2.1 DNA mutations in human genetics

DNA mutations - also known as genetic variants - are classified along several dimensions that will be referenced throughout this thesis. We define these classifications here. For the sake of disambiguation, the definitions are provided in the context of human genetics.

First, a mutation can be germline or somatic. A germline mutation is present in the fertilized zygote from which all cells in the body are derived; thus germline mutations are present in every cell of the body. The vast majority of germline mutations are inherited from parents; however, they can also arise spontaneously in sex cells, leading to a *de novo* germline mutation in a child that is not present in the germline of either

parent. Somatic mutations are those which were not inherited from parental sex cells. Somatic mutations can arise in any cell at any point from conception to death. They can be caused by endogenous factors such as DNA replication or exogenous factors such as UV radiation. Depending on when and where a somatic mutation arises, it can be present in a single cell, present in a small set of cells, or widely dispersed across the body [16, 50, 91]. For example, mutations that arise during early embryonic development will typically be widely dispersed throughout the body; such somatic mutations are often referred to as "mosaic". Mutations that arise in a post-mitotic cell such as a neuron will only exist within that cell.

Second, mutations are classified based on the number of base pairs they affect. Single nucleotide variants (SNVs) change a single base of the DNA to one of the other three possible bases. Small insertions and deletions (indels) are insertions or deletions that alter 1-50 bases. Structural variants (SVs) are rearrangements of the DNA that affect more than 50 bases.

Finally, a note on mutation nomenclature. SNVs and indels will be indicated by their direct DNA change. For example, an SNV that converts a cytosine to a thymine will be indicated as C>T. The base expected to be present (cytosine in the example) is known as the reference allele; the other base (thymine in the example) is the alternate allele.

2.2.2 High-throughput sequencing for mutation detection

Some of the work in this thesis relies on data from short-read high-throughput sequencing to identify somatic mutations. In this type of sequencing, the nucleotide content of a genome is directly assayed massively in parallel. DNA is extracted from many cells, sheared into short segments of typically 51-151 bases in length, and then the sequence of these short segments is directly determined in parallel through a biochemical reaction. Thus the data produced are millions of strings representing the nucleotide sequence of short segments of DNA from a person's genome. The original location of each segment relative to the human reference genome is then inferred in a process known as alignment, and mutations can be read-off as differences between

the observed sequences and the human reference genome.

2.2.3 Relevant principles of cancer genetics

In chapter 5 and chapter 6, we dive into the world of cancer genetics. Apart from the rare cases of an oncoviral spread (where cancer is caused due to a highly specific viral attack), cancer is a genetically driven disease. The initial set of mutations that drive the cancer is called *driver mutations*. It has been estimated that there are 2000-3000 unique locations across the genome that could harbor such mutations [213], either as germline or somatic drivers. While there are known germline-dominated driver locations, such as the infamous breast cancer-associated BRCA1 and BRCA2 driver genes [92], most driver mutations are acquired throughout our lifetime as somatic DNA alterations.

Somatic mutations naturally accumulate throughout one's lifetime. Most of these DNA alterations are corrected by an arsenal of genetic repair mechanisms. However, errors during replication and repair are inevitable and irreversible once the original allele template is lost [255]. Initial mutational burdens and defects in DNA repair mechanisms result in an extremely high somatic mutation rate in tumor cells. Most of these mutations are harmless "passenger" mutations, with only a small fraction being true driver events that provide a proliferative advantage to a cell [166, 261], thus making the task of identifying this limited subset of disease-causing mutations extremely challenging.

2.2.4 Relevant principles of epigenetics

The nucleus of a human cell contains nearly 2 meters of DNA. In order for it all to fit and for genes to be accessible for translation, DNA must be carefully packaged. The set of chemical modifications to DNA and its packaging proteins that enable this intricate compaction is known as epigenetics. Knowledge of epigenetics plays a major role in chapter 5 and chapter 6.

DNA winds around proteins called histones, creating a structure known as chro-

matin. The amino acid residues of histones often carry modifications that are associated with how tightly the DNA is compacted. For example, tri-methylation of the 27th lysine of the H3 histone (H3K27me3) is associated with highly compacted chromatin and limited gene expression. Conversely, tri-methylation of the 4th lysine of the H3 histone (H3K4me3) is associated with open chromatin and active gene expression.

Histone marks (also known as chromatin modifications) can be assayed using a special type of high-throughput sequencing known as Chromatin Immunoprecipitation sequencing (ChIP-seq). ChIP-seq has been applied extensively to characterize chromatin state across human tissues [218].

Chapter 3

The Aligned Multimodal Movie Treebank: an audio, video, dependency-parse treebank

3.1 Summary

Language is a complex process, involving multiple sensory modalities. However, existing datasets tend to focus on the direct consumption of language, ignoring additional inputs that we as humans use to process text and utterances. Specifically, Treebanks have become a frequent format to study linguistic questions and effects. Treebanks have traditionally included only text and were derived from written sources such as newspapers or the web. We introduce the Aligned Multimodal Movie Treebank (AMMT)†, an English language treebank derived from dialog in Hollywood movies which includes transcriptions of the audio-visual streams with word-level alignment, as well as part of speech tags and dependency parses in the Universal Dependencies (UD) formalism. AMMT consists of 31,264 sentences and 218,090 words, which will amount to the 3rd largest UD English treebank and the only multimodal treebank in UD. We find that parsers on this dataset often have difficulty with conversational speech and that they often rely on punctuation which is often not available from

speech recognizers. To help with the web-based annotation effort, we also introduce the Efficient Audio Alignment Annotator (EAAA)‡, a companion tool that enables annotators to significantly speed up their annotation processes.

3.2 Introduction

Treebanks are fundamental resources in Natural Language Processing [192]. Despite their central role, most existing treebanks are derived from single-modality texts such as newspapers, blogs, and other online communities. The vocabulary, syntax, and statistics of spoken and written language can be quite different from one another [47]. To complement these datasets and aid the advent of multimodal conversational agents, we have created a new dataset, the Aligned Multimodal Movie Treebank, AMMT, the content of which is derived from the language spoken in Hollywood movies. AMMT is released publicly under an open-source license and will be contributed to the Universal Dependencies (UD) [193] treebanks.

Speech-based treebanks have proven to be a resource of enormous importance to the NLP research community [8, 194]. We find Treebank-3 of the Penn Treebank [165], which includes the Penn Treebank Switchboard corpus [104], to be the closest existing dataset to AMMT. This corpus contains nearly one million transcribed words from Switchboard annotated with part of speech tags, dysfluencies, and parse trees,

†https://github.com/abarbu/ammt
‡https://github.com/abarbu/audio-annotation

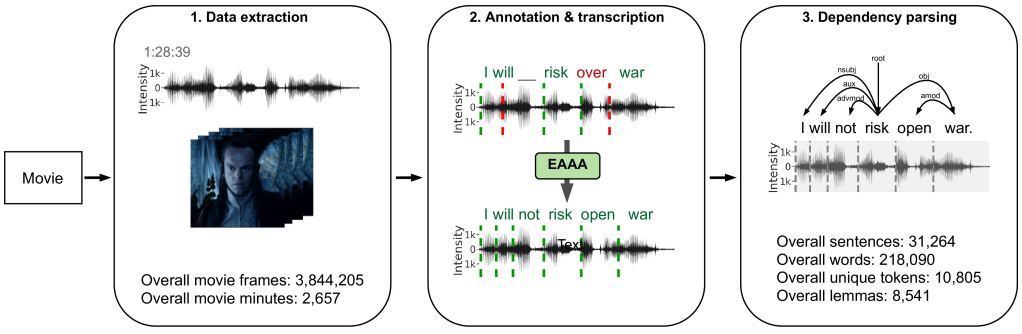

Figure 3-1: An overview of AMMT, our novel multimodal dataset, consisting of transcriptions and parses for 21 movies aligned at the millisecond level. EAAA is a new transcription and alignment tool introduced below.

and it also includes alignment between words and audio. However, there are several key differences between this dataset and our own. AMMT provides alignment to visual as well as audio data; it is annotated with UD rather than Penn Treebank dependencies; and conversations are much shorter (Switchboard was designed to have long 10-minute conversations between strangers on the phone discussing one of a preselected list of topics). While conversations in AMMT can still be considered as prepared speech, topics are way less constrained. AMMT also includes many more speakers and its audio quality allowed us to recover almost all spoken words. For practical experiments, AMMT is significantly more entertaining for subjects, a key feature for researchers aiming to study the neuroscience of language via neural imaging. Finally, with this contribution, AMMT is being made open to the whole research community.

3.3 Dataset

The AMMT dataset is an English language treebank based on 21 Hollywood movies that provide transcriptions with word-level alignment to the audio-visual stream, as well as part of speech tags and dependency parses in the UD formalism. Annotations for speaker identification will be included at the time of release. Due to copyrighted source material, AMMT provides multiple 1-second-long audio-visual sample clips from every movie, and a toolchain allowing users to obtain their own copies and verify alignment with the dataset.

AMMT consists of 31,264 sentences, 218,090 words, 8,541 lemmas, and 10,805 unique tokens. The counts of POS tags and dependencies are shown in appendix A. The 21 movies from which the dataset is derived are listed in table A.2 along with their unique identifiers and relevant statistics.

Movies were chosen to be appropriate for many ages, with the highest rating being PG-13. They belong to a variety of movie genres (including action, adventure, animation, comedy, drama, fantasy, family, and sci-fi, according to IMDb's categorization), and their release dates range from 1995 to the present. They were selected

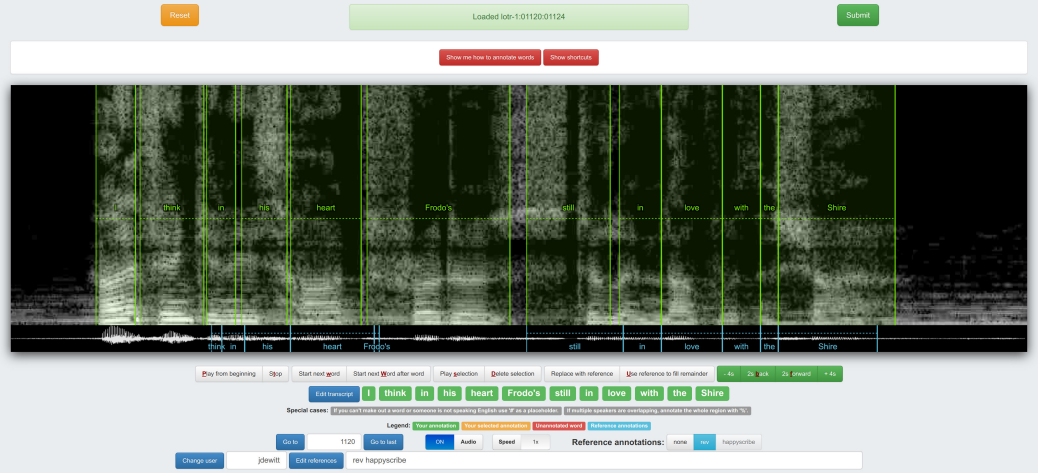

Figure 3-2: A screenshot of EAAA, the Efficient Audio Alignment Annotator. EAAA allows annotators to browse videos, play audio segments, play portions of the audio segments, edit the transcript, review multiple reference annotations, and annotate and change word boundaries. EAAA also includes an in-application walkthrough as well as extensive keyboard shortcuts. The main annotation area shows a spectrogram with annotated words. Words can be dragged with a mouse and similarly, word boundaries can be adjusted with the mouse. The audio for individual words can be played by clicking them, while any audio segment can be played by clicking and dragging the portion that should be played. At the bottom, in blue, one or more reference annotations are shown which can be toggled on the fly. Annotators can start with a blank slate or initialize annotations from any reference annotation. Audio speed can be controlled as necessary.

to have verbose scripts, in the top 50% of randomly sampled movies. Movies that included extensive singing such as musicals were omitted. Copies of the movies were obtained and extracted in full including opening and closing credits. Special features and after-credits scenes were omitted.

3.3.1 Transcription pipeline

The audio track was originally transcribed using the Google Cloud Speech-to-Text API [107]. It was then corrected by annotators, hired from *rev.com* and *happyscribe.com* depending on the movie, and then further extensively corrected by 7 expert annotators. Transcription followed a set of guidelines to deal with problematic audio segments and to enforce coherence. Manual transcription was performed simultaneously with word-boundary annotation using a new tool developed for this purpose, EAAA (see section 3.5), which was also subsequently used by annotators to perform

sentence segmentation and fixing capitalization.

The transcription was verbatim without any corrections for dysfluencies or mistakes. Instructions were provided to the annotators to standardize the transcripts and eliminate problematic audio segments. Foreshortened words (*'round* vs *around*) were transcribed as they were said including the foreshortening. Abbreviations were always expanded (*dr.* vs *doctor*). Cardinal and ordinal numbers were spelled out, while long numbers were written as spoken including conjunctions such as *and* (e.g., *five hundred and five*).

Aligned Multimodal Movie Treebank	
sentences	31,264
tokens	218,090
lemmas	8,541
types	10,805
num. movies	21

Table 3.1: Basic statistics of the AMMT

Manual transcription was carried out simultaneously with word boundary annotation using a purpose-built tool, EAAA (see section 3.5). EAAA presented annotators with a spectrogram for 4-second segments of a movie, along with the ability to search, replay and slow down any sub-segment throughout the movie. As the audio was played, a line marked the location of the audio sample in the spectrogram in real-time. In some cases, annotators could hear specific words but could not clearly identify in the spectrogram where those words occurred (e.g. short words like *to*). Annotators were instructed to annotate what they heard regardless of the spectrogram, sometimes leading to such short words having zero-length intervals. Foreign sentences (e.g., Elvish in the movie *The Lord Of The Rings*) were marked but not included in the corpus, although one-off foreign words in English sentences were transcribed. All cases of singing, unintelligible speech, and multiple speakers overlapping were noted and eliminated from the dataset. Transcripts are as spoken, without correction, even when the speaker erred by omitting a word or using a word inappropriately.

After transcription and word boundary alignment, the text was segmented into

Metric	Precision	Recall	F1 Score	AligndAcc
Words	100.00	100.00	100.00	N/A
UPOS	99.53	99.53	99.53	99.53
UAS	98.95	98.95	98.95	98.95
LAS	98.31	98.31	98.31	98.31
CLAS	97.75	97.71	97.73	97.71
MLAS	96.74	96.70	96.72	96.70

Table 3.2: Inter-annotator agreement bound of AMMT syntactic annotations.

sentences. Annotators marked the end of each sentence manually and fixed capitalization (of both proper nouns and sentences as needed). Throughout this process, some critical punctuation was introduced as annotators saw fit.

3.3.2 Dependency parsing pipeline, annotation and validating annotator performance

We parsed all transcriptions with Stanza [206] using the standard English model.

The AMMT dataset was entirely annotated by an in-house expert annotator over the course of a year. Edge cases were discussed with other three team members with a strong background in linguistics and Universal Dependencies in particular. In this period of time, the expert annotator performed a total of three sequential passes *over the full dataset* with the idea of promoting internal consistency.

Separately, after this annotation process concluded, a subset of AMMT consisting of 300 sentences of length 5 through 20 uniformly sampled across movies were reannotated by an expert annotator. This expert annotator has a strong background in linguistics and did not contribute to the dataset otherwise. The length of these sentences was selected to avoid the effect of very short or very long sentences (see table 3.2).

The inter-annotator agreement of the annotations was 99.53% on correct POS tagging, 98.95% on correctly placing dependencies (UAS), and 98.31% on correctly identifying the type of a dependency relation. Morphology-aware labeled attachment (MLAS) score ties together POS and LAS into a single number, 96.72%, which mea-

sures the inter-annotator agreement of the annotations [243].

Note that the inter-annotator score presented in table 3.2 is thus a measure, for this particular subset of the dataset, of the disagreement between the original expert annotator and the external expert annotator. As such it should only be considered as a bound on the actual disagreement between the two annotators.

We found word-boundary inter-annotator agreement to be remarkably high, with less than 15ms on average for all words in a single movie, *Lord Of The Rings*, annotated by 5 annotators.

3.3.3 Performance of existing parsers

We compared our annotations against those produced by Stanza [206] in fig. 3-3. Stanza was the original parser used to initialize the treebank before extensive human correction. This likely biases the results toward Stanza in subtle ways [27] which we do not investigate here beyond section 3.3.2.

Note that performance on short sentences, fewer than 3 words, and long sentences, with more than 20 words, is far worse than average-case performance (see fig. A-1 for the distribution of sentences in AMMT). This trend is not observed in other corpora such as the English Web Treebank (EWT) [237], where performance increases for short sentences (although these are very infrequent) while the performance drop for long sentences is half or less than that seen in AMMT. While the distributions of POS in both corpora are slightly different (cf. appendix A), the performance drop for short sentences appears to be driven by POS tag errors, see the relative drop in POS accuracy between fig. 3-3(a,b,c) — perhaps such sentences require more context to be correctly interpreted. The performance drop for long sentences appears to be driven by incorrectly identified relationships, see the relative drop in UAS between fig. 3-3(a,b,c).

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.51	99.75	99.63	N/A
UPOS	97.64	97.88	97.76	98.13
UAS	88.02	88.24	88.13	88.46
LAS	85.68	85.89	85.78	86.10
CLAS	83.40	83.01	83.20	83.29
MLAS	81.38	80.99	81.18	81.27

(a) All sentences

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.45	99.53	99.49	N/A
UPOS	91.49	91.56	91.53	92.00
UAS	91.31	91.38	91.35	91.82
LAS	88.76	88.83	88.80	89.25
CLAS	86.49	86.06	86.28	86.71
MLAS	75.87	75.50	75.68	76.06

(b) Short sentences, fewer than 3 words

Metric	Precision	Recall	F1 Score	AligndAcc
Words	99.52	99.78	99.65	N/A
UPOS	98.44	98.70	98.57	98.92
UAS	80.47	80.68	80.57	80.86
LAS	78.78	79.00	78.89	79.17
CLAS	76.32	76.06	76.19	76.28
MLAS	74.02	73.77	73.90	73.98

(c) Long sentences, more than 20 words

Figure 3-3: (a) The overall accuracy of Stanza on AMMT. Performance drops significantly for (b) short sentences which are common in speech as well as for (c) long sentences.

3.4 Multimodal feature analysis

Exploring the utility of the corpus as a multimodal resource for grounded language and vision tasks, we quantified the co-occurrence of nouns and their corresponding objects (i.e. objects that are verbally mentioned as they appear on screen). As an approximation, we considered the 80 object classes of the Microsoft COCO dataset [157]. We extracted all nouns corresponding to a COCO class (580 nouns across all movies) and manually reviewed the middle frame of a word utterance. We find an average of 36.5% noun-object agreement rate (212 co-occurring objects) across all movies ($\mu = 23.7\%$, $\sigma \approx 17.5\%$ per movie); see fig. 3-4.

Considering noun-object agreements across both object classes and movie types

reveals variable distributions. Some nouns are highly likely to appear on screen as their corresponding noun is uttered, like Person (94.4%), types of vehicles (Car: 59.7%, Bicycle: 68.3%), and animals (Giraffe: 100%, Cow: 100%), while others have not co-occurred once despite being uttered multiple times. Moreover, unambiguous nouns (e.g. Laptop: 50%, TV: 42.8%, Toilet: 33.3%) tend to have significantly higher agreement rates than words with multiple POS (e.g. Bear: 2.5%, Orange: 0%, Remote: 0%). Some movie categories are also more likely to have a high noun-object agreement, such as movies aimed at a younger audience (educational and animation genres), perhaps to enable language learning through multimodality. For example *Cars-2* and *Sesame Street* present 79.2% and 74.3% agreement rates respectively, while *The Lord Of The Rings 1* and *2*, and *Avengers Infinity War* score only 17.6%, 14.2%, and 5.9% respectively; see fig. A-2.

3.5 Tools

To efficiently annotate the alignment between word onsets and offsets and the audio stream, we created a new tool, the Efficient Audio Alignment Annotator (EAAA). EAAA enables annotators to start with a rough transcript and approximate alignment between words and the audio track. Annotators can simultaneously correct the transcript while annotating new words. An overview of the EAAA interface is shown in fig. 3-2. Tools such as Praat [35] also allow for annotating audio corpora with word boundaries. Unlike Praat, EAAA is web-based making it easier for annotators to use. Data such as spectrograms and wave files seen by annotators are pre-processed on the server side, making browsing and accessing movies with EAAA near real-time. Since EAAA is a single-purpose tool meant for transcription and fine-grained alignment, it provides custom features that significantly speed up the annotation process like keyboard shortcuts, the ability to handle audio files of any length, and a streamlined interface. EAAA also handles multiple concurrent annotators, sharing and comparing multiple annotations directly.

EAAA pre-processes movie files into 4-second segments that overlap by 2 seconds

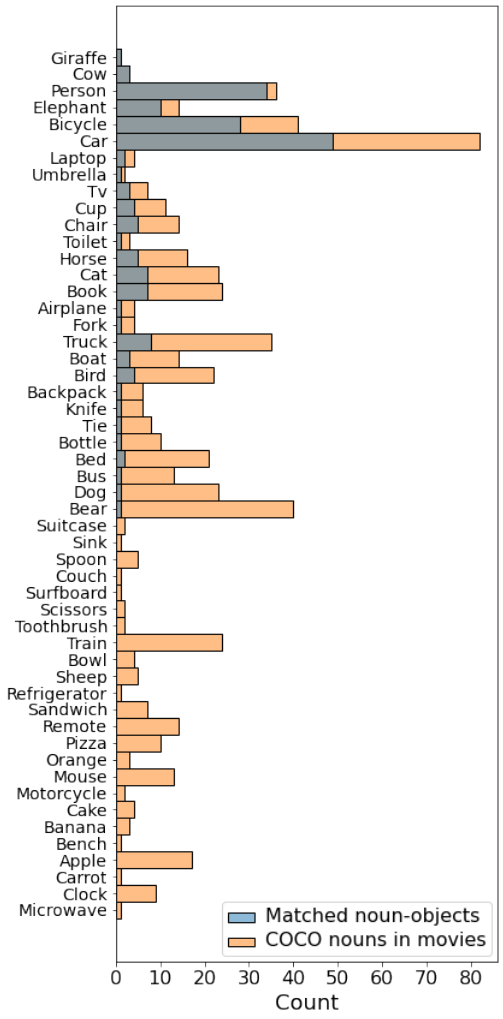

Figure 3-4: COCO classes noun-object agreement across the corpus (sorted by agreement rate). All nouns corresponding to one of the 80 COCO classes (orange) vs their corresponding objects in the video during the noun utterance (blue). Objects were manually detected in the middle frame of a word utterance.

and computes spectrograms for each segment with Librosa [171]. Storage is provided by a local Redis database which is not exposed to the web. In addition, EAAA includes a telemetry server that collects comprehensive information during the annotation process including every transcript change, keyboard shortcut used, and mouse press.

3.6 Conclusion

AMMT and EAAA are open source and AMMT will be contributed to the UD treebanks. In addition to verbatim transcriptions and a treebank, AMMT provides a toolchain to enable access and alignment to the source video and audio. Most datasets for evaluating and training parsers are focused on written rather than spoken language. With the rise of conversational agents, AMMT can serve as a more predictive benchmark in this domain.

At present, no end-to-end systems – from video-and-audio to parses – exist, even if humans often use visual information to disambiguate and contextualize auditory information. In the next chapter, we will show how AMMT will support further work on the neuroscience of language.

Chapter 4

Neural processing of nouns and verbs with large-scale intracranial recordings from naturalistic language

4.1 Summary

The understanding of language structure and its representation in the brain remains a major challenge with substantial implications for neuroscience, linguistics, and artificial intelligence. The considerable impediments of coarse signal recording resolution, limited data, and confounding features have thus far hindered this area of study. By resolving these constraints, we enable probing the neural spatiotemporal dynamics of nouns and verbs as a proxy for part-of-speech (POS) processing at an unprecedented level of detail. We identify a tightly-connected network of brain areas that respond selectively to nouns and verbs. The network is organized as two semi-overlapping components and clustered around a main processing core. We find that this core anticipates the POS of an upcoming word prior to word onset, takes on most of the computational burden in determining the lexical category during utterance, and transmits the information to auxiliary regions. Finally, we demonstrate the critical nature of context, which differentially changes the neural activity and latency evoked

by nouns and verbs.

4.2 Introduction

The neural representation and dynamics involved in even the most fundamental language processing tasks are still largely unknown despite the mounting evidence linking brain domains to language-related phenomena, such as compositionality [22, 33, 225, 264], semantic categories [180, 126, 277], and surprisal [41, 40, 105, 43, 29]. A particularly interesting case is understanding how the brain ascertains the part of speech (POS) of words. Grammatical classes are of particular importance for their fundamental role in linguistics and natural language processing (NLP). Indeed, the two word classes, nouns and verbs, are widely recognized to be among the few linguistic universals [60, 205]. To date, our understanding of the spatio-temporal course of part of speech processing in the brain is limited by 1) experiments which have coarse spatial resolution such as Magnetoencephalography (MEG) [53] and Electroencephalography (EEG) [112], or coarse temporal resolution, as in functional magnetic resonance imaging (fMRI) [221, 183, 177, 25, 81, 77]; 2) the insufficient amount of naturalistic full-sentence language data used [220, 82], required to engage the full capacity of the language system and free of laboratory-constructed task artifacts [113, 205, 38, 34]; and 3) a plethora of correlated confounding factors [12] which are difficult to disentangle without large-scale data. Previous attempts to understand neural activity evoked by part of speech [78, 251, 232, 53, 1, 178, 207, 129] have yet to overcome all three limitations synchronously, greatly limiting our understanding of the language system's internal structure and dynamics.

To overcome these three hurdles, we leverage the AMMT to create the first large-scale naturalistic language dataset with invasive stereoelectroencephalography (SEEG) neural recordings of 236,400 annotated tokens across 10 subjects – the Brain TreeBank. Powered by our rigorous annotation process, described in chapter 3, this large collection of high-resolution neural recordings aligned with linguistic annotations scales up data per subject by over a factor of 10 compared to other naturalistic

language datasets [137, 28, 105] and by over a 100 for the more common controlled studies [80, 79, 260, 82, 23, 39, 207]. We then introduce a broad battery of methods to eliminate confounds and provide a structured experiment level of control.

The availability of large-scale, high-resolution neural recordings annotated with linguistic information and an exhaustive list of quantified confounds allow us, for the first time, to probe POS at an unprecedented level of detail. To this end, we address the following set of questions (i) how do activation patterns of words with distinct POS differ? (ii) How does the POS network distribute across the brain? And (iii) what are the temporal dynamics of the POS processing network?

Our analysis reveals new information on the neural representation and spatiotemporal localization of POS processing. First, we show a network of language areas that respond selectively to nouns and verbs, enabling direct decoding of POS from neural signals even when their surface forms are identical. Responses are structured hierarchically, such that areas associated with early language processes show increased POS sensitivity relative to higher cognitive function areas. Neural activities are also markedly different: verbs evoke stronger activity and take longer to process than nouns, with a characteristic dependency on the sentence context in which a word was uttered. However, a fine-grained analysis reveals a noun-specific cluster in the inferior frontal lobe (IFL), a previously perceived verb-sensitive region [254]. A high-precision spatiotemporal SEEG analysis shows that the superior temporal lobe (STL) has a central role as a processing core recruiting auxiliary support from nearby areas. The temporal analysis exposes two POS prediction intervals: a primary window occurring between $150ms$ and $500ms$ post word onset, and an anticipatory window where POS is predictable $350ms - 250ms$ before the utterance of that word.

4.3 Results

Subjects participating in the experiment (10 subjects, 5 male, 5 female, aged 4-19, $\mu = 11.9$, $\sigma \approx 4.6$) were patients under treatment for epilepsy at Boston Children's Hospital (BCH), where they had been implanted with intracranial electrodes (total

1,688; $\mu = 169$, $\sigma \approx 40$) to localize seizure foci for potential surgical resection. We collected SEEG recordings while subjects were watching a total of 28 full-length Hollywood films ($\mu = 2.8$, $\sigma \approx 1.8$ per subject) out of a selection of 21 choices.

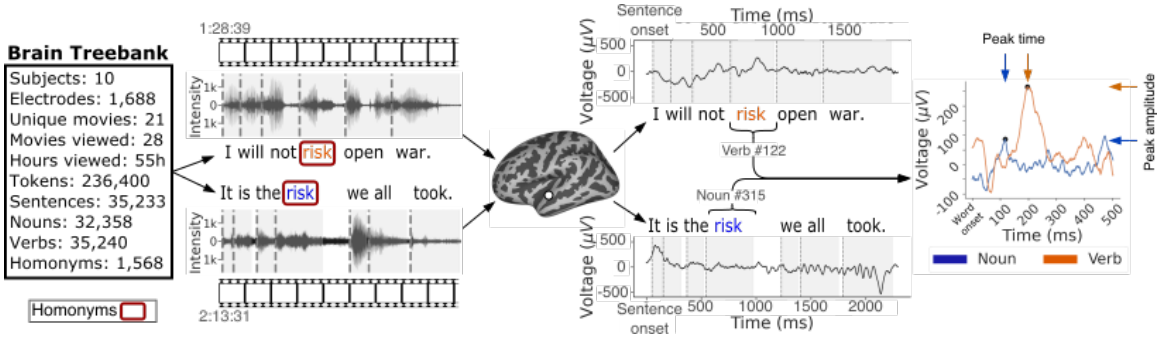

Figure 4-1: Task schematic, data alignment, and definition of word homonyms. noun-verb homonym pair in the Brain TreekBank, a novel large-scale dataset of brain activity, recorded using SEEG as subjects watched Hollywood movies. We show the neural data extracted from one example electrode corresponding to the presentation of a homonym that appears as both a noun and a verb. On average, verbs evoke more activity and take longer to process than nouns.

4.3.1 Brain TreeBank: a large-scale intracranial naturalistic language dataset

The collected SEEG recordings amount to 55 hours ($\mu = 5.6h$, $\sigma \approx 4.2h$ per subject), and form the basis for the Brain TreeBank dataset. The dataset contains 35,223 sentences ($\mu = 3,522$, $\sigma \approx 2,384$ per patient), 46,659 types ($\mu = 4,666$ $\sigma \approx 3,209$ per patient), and 236,400 tokens ($\mu = 23,640$, $\sigma \approx 16,093$ per patient) (see fig. 4-1). The onset and offset of each word for every presented movie were automatically annotated and then manually corrected via an aligned spectrogram and reduced speed audio track. The dataset was also automatically parsed for POS and manually corrected using the Universal Dependencies framework.

Brain TreeBank is to date the only large-scale treebank accompanied by neural data; it is the largest multi-modal treebank (i.e. containing both video and audio tracks), and the third largest treebank altogether. The breadth of the dataset facilitates the construction of pseudo-controlled experiments such as noun and verb pairs

that sound and are written exactly the same (noun and verb homonyms, e.g. risk, love, etc.). We found 4,090 such homonym pairs ($\mu = 584$, $\sigma \approx 432$ per patient) that naturally occurred in the corpus (manually validated as homophones). Like AMMT, Brain TreeBank will be open-sourced (extensive details are provided in Methods section 4.5.1).

4.3.2 How POS neural responses differ?

We first separate all nouns and verbs in the dataset into three distinct categories, by their sentence context, due to potential neural spillover effects from neighboring word attributes (e.g. surprisal, parse tree complexity, etc.). Each word was assigned to one of three subsets – sentence onset (beginning), midset (middle), or offset (end) – manifesting distinguishably different neural representations; see fig. 4-2.a. Additionally, we define electrodes with more than $40\mu V$ min-max range in the average activity evoked by sentence onsets to be language responsive (115 electrodes overall; 10.66% of all electrodes) (see Methods Methods section 4.5.4). This is a simple screening step to denoise language signals independently of POS patterns.

Tracking the evolving activation patterns across sentence progression in language-responsive electrodes revealed clear neural correlate distinctions between nouns and verbs, with growing significance, for both average peak time and amplitude (peak time paired two-tailed t-test: onsets $p = 1.76 \times 10^{-3}$, midsets $p = 1.75 \times 10^{-5}$, offsets $p = 1.95 \times 10^{-8}$; peak amplitude paired t-test: onsets $p = 9.05 \times 10^{-11}$, midsets $p = 8.24 \times 10^{-13}$, offsets $p = 1.71 \times 10^{-27}$) (see fig. 4-7). Moreover, the location of a word in its context sentence appears to modulate underlying key components of the POS-induced activations. As previously reported, verbs typically induce stronger responses [254], but when every subset is considered separately, we find that nouns at sentence onset produce significantly higher average amplitudes (nouns: $\mu = 53.12\mu V, \sigma \approx 21.07\mu V$; verbs: $\mu = 39.44\mu V, \sigma \approx 21.71\mu V$). The trend reverses towards verbs evoking stronger amplitudes at sentence midset (nouns: $\mu = 17\mu V, \sigma \approx 11.21\mu V$; verbs: $\mu = 20.48\mu V, \sigma \approx 11.95\mu V$). This difference further expands at sentence offset (nouns: $\mu = 14.53\mu V, \sigma \approx 9.14\mu V$; verbs:

$\mu = 27.18\mu V, \sigma \approx 12.56\mu V$). Similarly, verbs induce enhanced latency in average peak times, with an increasing difference after sentence onset (onset – nouns: $\mu = 373.18ms, \sigma \approx 77.93ms$, verbs: $\mu = 395.9ms, \sigma \approx 93.14ms$; midset – nouns: $\mu = 235.89ms, \sigma \approx 114.43ms$, verbs: $\mu = 298.73ms, \sigma \approx 120.76ms$; offset – nouns: $\mu = 153.84ms, \sigma \approx 150.39ms$, verbs: $\mu = 255.33ms, \sigma \approx 164.19ms$). Crucially, the observed increased latency comes as a contrast to longer noun utterances as found in the dataset (nouns: $\mu = 411.4ms, \sigma \approx 180.4ms$; verbs: $\mu = 287.9ms, \sigma \approx 158.2ms$).

To further validate the trend of increasing peak property differences (verbs - nouns) and overcoming potential measurement artifacts, we simulated the peak times and amplitudes based on the observed mean and variance in every language-responsive electrode. For every electrode, we computed per time-point (1,024 samples across $0ms - 500ms$ post word onset) distribution properties ($\mu \pm \sigma$), separately for nouns and verbs. We sampled 10,000 independent signal vectors from the multi-variate normal distribution and simulated peak differences via a non-parametric kernel dissimilarity density estimation (KDE) (2.5%-97.5% inter-quantile region) for all three sentence subset classes (see Methods section 4.5.5). As expected, the difference peak amplitude distributions shift upwards across sentence progression (onset mean: -7.89μV, midset mean: 3.63μV, offset mean: 6.86μV), from a heavy left tail (nouns > verbs) onset distribution to a right tilting offset distribution (nouns < verbs). All distributions were found significantly different (T-test onset $\neq$ midset: $p = 2.97510^{-35}$, midset $\neq$ offset: $p = 5.08510^{-8}$, onset $\neq$ offset: $p = 5.12110^{-18}$). Correspondingly, the peak time distributions recapitulate later peak latency for verbs over nouns, that significantly increase post sentence onset (onset mean: $15.819ms$, midset mean: $74.618ms$, offset mean: $75.748ms$; T-test onset $\neq$ midset: $p = 3.17510^{-16}$, onset $\neq$ offset: $p = 3.7710^{-15}$); see fig. 4-2b. top and side panels. When comparing the peak properties differences, we find positive correlations for the midset and offset subsets (midset – Pearson R=0.386, $p = 2.08 \times 10^{-5}$; offsets – Pearson R=0.42, $p = 3.03 \times 10^{-6}$), but a negative correlation for the onset subset (Pearson R=-0.534, $p = 7.84 \times 10^{-10}$), an indication that sentence-leading evoked activity is guided by a nonidentical underlying process (see fig. 4-2b.).

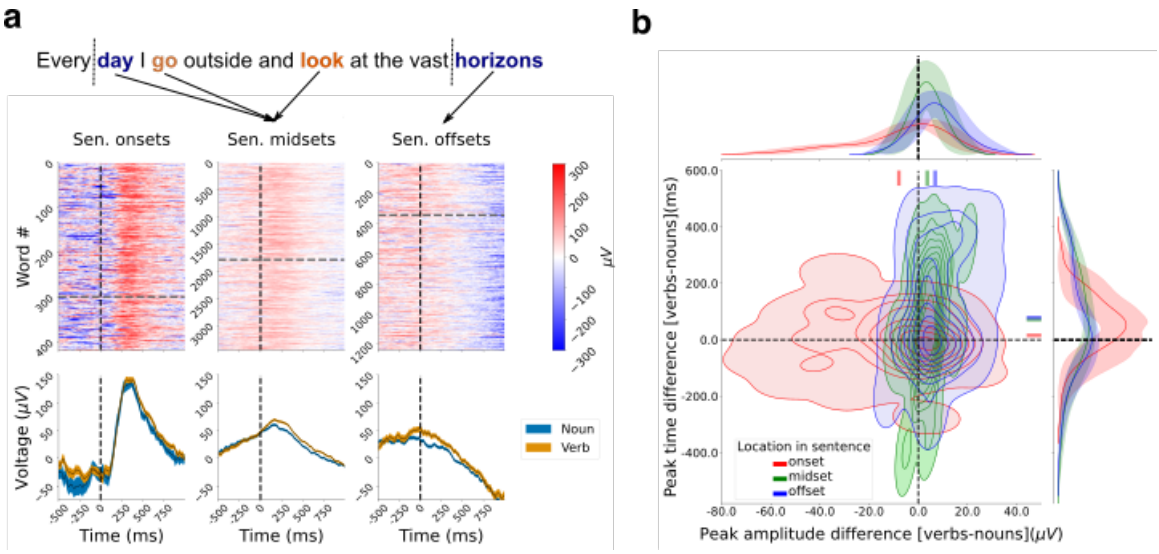

Figure 4-2: a. Neural responses to individual words depend on the position within a sentence and not on their part of speech. Data from a single electrode in one subject from one movie. A representative of the activity seen in language-sensitive electrodes across subjects. (top) Raster plots to every word in a movie in a window of $500ms$ before and 1 second after the onset of the word. (bottom) Average IFP to all nouns (blue) and all verbs (orange). **b.** A density plot of the differences in peak time and amplitude for every language-sensitive electrode across all subjects. Top and right show the marginals for each axis independently. Verb peaks are delayed and have higher amplitudes on average, with a strong dependence on sentence context.

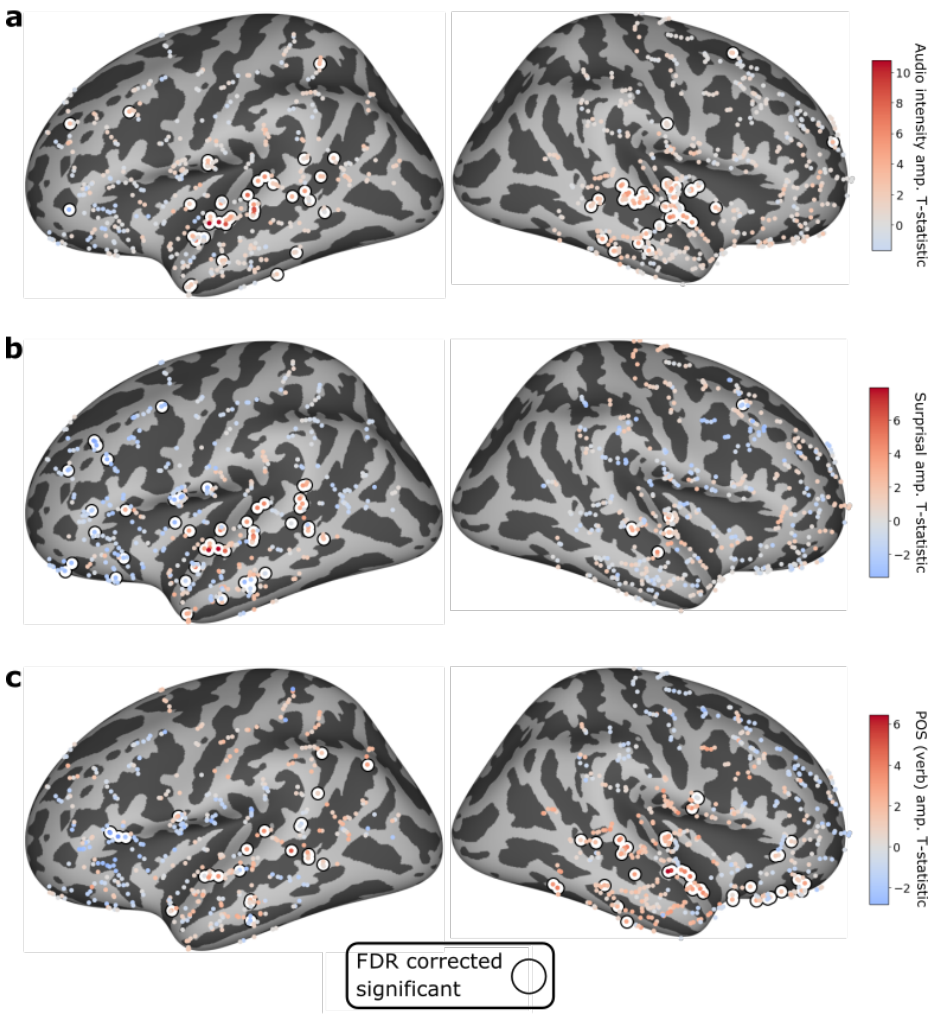

Figure 4-3: Effects of GLM features (captured by T-statistic) on the neural signal peak amplitude within $0ms - 500ms$ post word onset. Significant electrodes have $p \leq 0.05$ combined over peak time and amplitude p-values, FDR corrected via 2-stage Benjamini-Krieger-Yekutieli. **a.** Effects of audio intensity over signal amplitude computed as audio stream magnitude. **b.** Effects of word surprisal over signal amplitude as computed from the GPT-2 model. **c.** Effects of POS over signal amplitude, verb sensitivity marked in red, noun sensitivity marked in blue.

4.3.3 How is the neural POS network distribute?

When examining POS effects over single electrodes we must control for a wide variety of confounding factors biasing the results by affecting the neural signal patterns. For instance, multiple measures of word surprisal have been found to be predictive of nouns and verbs [12], word length distributions of POS are typically distinguishable and word index-in-sentence effects peak attribute differences as shown in section 4.3.2. Additionally, serendipitous confounds could be introduced through the auditory or visual scenes presented to the subject in our dataset. We curated a list of 33 confounding features from the video, audio and language of every presented trial (see full list on Extended Figures Extended Figures table 4.1 and additional details on Supplementary Methods appendix B.1.4). The features were provided as additional regressors to a generalized linear model (GLM), along with POS label, inferring properties of the neural activity (peak time and amplitude) (see Methods Methods section 4.5.6). The GLM analysis enables estimating the POS neural modulations, independent of the effects of all additional regressors. Considering the GLM's T-statistics per-electrode reveals the directionality, power and spatial distribution of the different predictor effects across the brain. This analysis made no prior assumptions about language responsiveness and evaluated all recording electrodes.

We first establish that our methodology recapitulates known mechanisms through audio intensity (average magnitude $0ms - 500ms$ post word onset) effects on the neural signal. Indeed, out of the 82 intensity significant electrodes ($p \leq 0.05$ FDR corrected with 2-stage Benjamini-Krieger-Yekutieli over Fisher's combined peak time and amplitude two-sided significance), we find substantial bilateral enrichment in the intensity-encoding sub-region of the auditory cortex [30] (56 electrodes, 68.3%) and insula (20 electrodes, 24.4%), previously associated with auditory stimuli tuning and attention [19]. Moreover, as expected we find positive correlations between peak amplitude and audio intensity levels in 95.3% of significant electrodes (see fig. 4-3a.).

Beyond supporting previously described effects of auditory surface features, the analysis provides insights into higher cognitive functions, such as word surprisal, an

established predictor of behavioral measures in language comprehension [242]. We computed preceding-context-aware word-surprisal estimates via a pretrained GPT-2 model [265]. Surprisal increases neural signal amplitudes in 32 out of 55 surprisal significant electrodes, amassed across the STL and insula (93% of significant positively correlated electrodes combined) (see fig. 4-3b.). In addition, results show previously unreported surprisal-significant but negatively correlated regions (23 electrodes). These IFL, medial frontal lobe (MFL) and medial temporal lobe (MTL) (accounting for 73.9% of significant negatively correlated electrodes) demonstrate lower peak amplitudes for words with higher surprisal values.

A GLM-based analysis of the POS processing network finds 72 nouns and verbs sensitive electrodes across the brain (see fig. 4-3c.). The found electrodes are clustered in major language network areas: STL (38.9%, 28 electrodes), IFL (22.2%, 16 electrodes), insula and central operculum (19.4%, 14 electrodes), MTL (8.3%, 6 electrodes), inferior temporal lobe (ITL) (6.9%, 5 electrodes), and the supramarginal gyrus (4.2%, 3 electrodes). While higher amplitudes and increased peak latency are primarily correlated with verbs in the temporal lobe and insula (as shown in section 4.3.2), a dense cluster of IFL electrodes measure the reverse response, with stronger peak effects of the noun class. This observation provides a complementary POS processing view to claims of noun-specific sub-regions during POS production within the IFL, commonly assumed to be verb-specific [119]. Notably, these results are lateralization-oblivious since electrode locations were guided by a preceding language lateralization screening [21], biasing against the canonical left hemisphere language dominance.

Moreover, a language-specific functional connectivity analysis recovers a dense POS network (see fig. B-1). We compared the correlation between electrodes during utterances of nouns and verbs, versus equal scale segments with no recorded speech (see Methods Methods section 4.5.11). A pair of electrodes is determined connected if the distribution of correlations during utterances is strictly higher than the no-speech counterpart distribution (one-sided Mann-Whitney-U, $p \leq 0.05$ Bonferroni corrected for the number of un-directed connections). We find the subset of POS-sensitive

electrodes vastly enriched with inter-connected electrodes ($p = 3.28 \times 10^{-117}$, one-sided Mann-Whitney-U test) (see fig. 4-4a.). A clustering analysis of the connectivity map across all electrodes reinforces our previous POS region results and finds a strongly connected network spanning across the STL, medial temporal lobe (MTL), IFL and insula (see fig. 4-4b.). Interestingly, we find two information flow loops. A bottom loop that contains the STL, insula, and MTL (number of connections between the STL and insula: 340, insula and MTL: 221, MTL and STL: 350), and a top loop that contains the STL, insula and IFL (number of connection between the insula and IFL: 211, IFL and STL: 160). In contrast, the MTL and IFL share only 81 significant connections. In addition, we find the STL and IFL tightly inter-connected (280 and 202 connections respectively), while the insula has only 138 and the MTL has no intrinsic connections whatsoever.

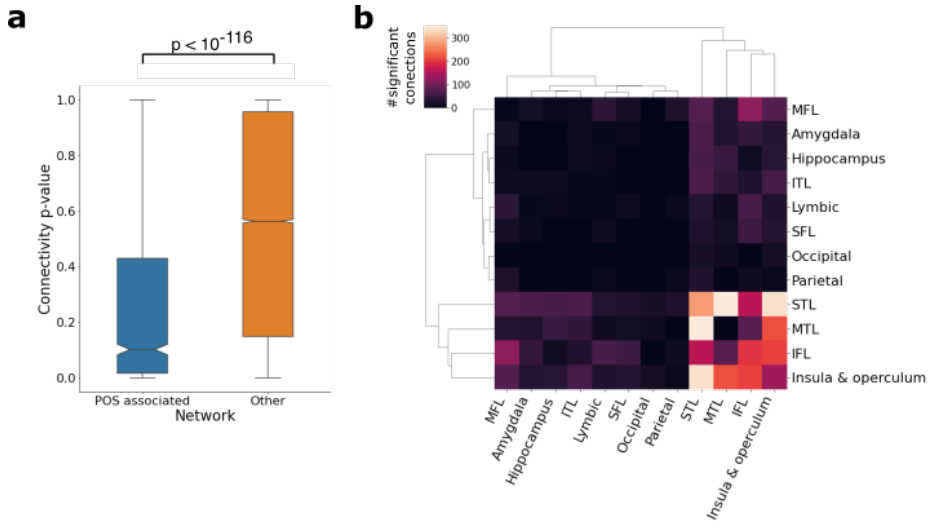

Figure 4-4: POS functional connectivity map analysis. **a.** A box plot comparing the p-values across the POS-associated network and all other electrodes. p-values represent the difference significance between the distributions of electrode correlation during POS utterance and no-speech segments for every pair of electrodes. POS-associated network connectivity is statistically larger ($p = 5.06 \times 10^{-46}$, one-sided Mann-Whitney-U test). **b.** Cluster analysis of brain areas POS connections. Each heatmap cell counts the number of un-directed connections between the two labeled areas. Regions were hierarchically clustered by their Euclidean distance.

4.3.4 What are the dynamics of POS processing?

We further explore POS processing activity as it propagates across the brain. As explicit POS neural correlates become indefinable pre and post-word utterance, we leverage the flexibility of neural networks to decode lexical categories from the raw signal ($0ms$ to $500ms$ post-word onset). Specifically, we use a convolutional neural network (CNN) to perform a binary classification task between nouns and verbs (see Methods section 4.5.7). To further disentangle convoluted features we: 1) introduce a pseudo-controlled experimental design of homonym noun and verb pairs (see fig. 4-1 and Methods section 4.5.8); 2) compute two per-word dense vector representations of the auditory and visual scenes to negate complex effects (e.g. auditory envelope, visual background cues, etc.) (see full list on Extended Figures table 4.1 and additional details on Supplementary Methods appendix B.1.4); 3) leverage the set of 33 previously described properties utilized in section 4.3.3; and 4) perform a held-out trial analysis for multi-trial patients (see Methods section 4.5.10).

For each alternative feature (33 scalar features and 2 vector representations) we first sample a test set with balanced feature distributions between the two classes and then sample the train and validation sets at random, enabling a true estimate of the model's feature-specific biases. The datasets are re-sampled 5 times and models are re-initialized and trained 5 times per data split (a total of 25 independent models per feature) to overcome sampling biases (see Methods section 4.5.7). An electrode is determined confounds-significant only if it performed above chance ($p \leq 0.05$ one-sided Binomial test) across all 35 confound-balanced experiments (see Methods section 4.5.9). The random sampling of the train and validation sets, independent of the selection of the test set balancing feature, makes it unlikely to rely on alternating confounds for different tests. Therefore, the false discovery rate for confounds-significant electrodes is incredibly slim. A homonym set analysis was performed in a similar manner, where a held-out set of all homonym pairs was extracted once per subject and significance was determined via a permutation test (see Methods section 4.5.9).

The homonym analysis finds 89 significant electrodes ($\mu = 12.7, \sigma \approx 11.7$ elec-

trodes per subject) with respect to the null distribution. Separately, 18 electrodes are confounds-significant across all 35 tested features with an intersection of 17 electrodes (see fig. 4-5a.). A region analysis (pixel-wise corrected for multiple comparisons [62]; see Methods section 4.5.9) recapitulates results found in section 4.3.3. The CNN finds 24 region-significant electrodes (in 4 subjects) across the STL (75%, 18 electrodes), insula (12.5%, 3 electrodes), supramarginal gyrus (8.3%, 2 electrodes), and IFL (4.2%, 1 electrode).

Additionally, for electrodes that were found significant under all conditions, we use the homonym set to perform the first-ever held-out-trial decoding experiment of a high-level language feature (see Methods section 4.5.10). Crucially, the held-out-trial analysis shows the ability to perform inference over data recorded on different days and generalize across a variety of linguistic distributions (e.g. train over fantasy genre and predict over a kids animation trial). Interestingly, 16 of the 17 tested electrodes maintained the same decoding performance, notably reinforcing the robustness of the signal.

Lastly, we performed a sliding window fine-grained search ($100ms$ adjacent windows, $50ms$ overlap) between $500ms$ pre and $600ms$ post-word onset. This analysis goes beyond the static view of active regions during word utterance. POS processing appears to contain two distinct intervals. The first ranges between $150ms - 500ms$ post word onset and corresponds to the current state of the field [205]. The second, a more subtle indicator – perhaps driven by linguistic clues like the previous word POS – is an anticipatory interval ranging between $350ms - 250ms$ pre-word onset (see fig. 4-5). Specifically, as suggested in section 4.3.3, the lateral aspect of the STG acts as the main POS processing core by predicting next-word POS, performing analysis between $150ms - 300ms$, and broadcasting the information to auxiliary POS cores at the STL, insula and more, after $350ms$.

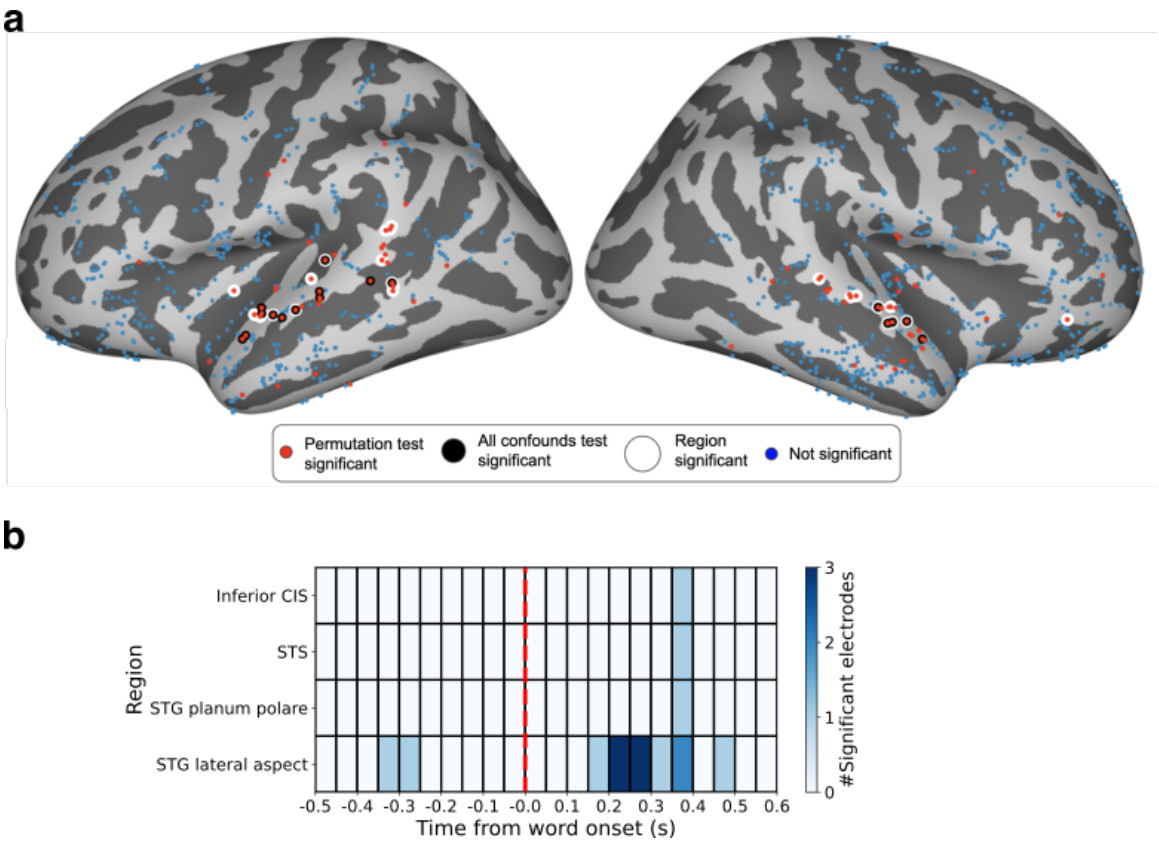

Figure 4-5: POS CNN dynamic decoding. **a.** Overlaid static decoding results for all electrodes across all subjects, projected to an inflated average brain. Electrodes are colored by significance in the per-electrode permutation test (small red), balanced confounds test (medium black), and region max-permutations test (big white). All other non-significant electrodes are colored in blue. **b.** POS processing temporal progression at $100ms$ windows ($50ms$ overlap) across all multi-trial subjects. Significant time-region tuples are colored according to the number of significant electrode hits. Colored cells preceding word onset (dashed red line) are signs of anticipatory signals. The STG acts as the POS processing core, broadcasting to additional auxiliary regions.

4.4 Discussion

Small and highly controlled datasets of neural recordings produce numerous negative results in the study of language in the brain. Conversely, NLP has advanced tremendously by scaling datasets and adapting methods to those larger corpora. We import this methodology – of high spatiotemporal precision at scale through naturalistic language – to studying a critical question in the neural processing of language: how the brain determines the part of speech of words in context.

We uncover differential context-dependent responses to nouns and verbs, where typically verbs evoke longer and stronger neural activity than nouns. This increased verb-associated activity is supported by their role in a sentence, often integrating multiple syntactic arguments, which perhaps require additional neural resources. Context modulates this response significantly, typically overpowering other effects, as shown by the radically different attributes of words that appear at the beginning of sentences. This provides an argument for steering away from investigating individual words or short phrases and towards naturalistic stimuli. Similarly, we find a focal noun-sensitive sub-region in the IFL, commonly assumed to be verb-specific based on imaging and impairment analyses [254], highlighting the shortcomings of low-resolution methodologies. Combined with additional intracranial studies that find comparable patterns during POS production [119], there is growing evidence for the necessity of intracranial methods in neurolinguistics due to the complexity of the language system.

The variety of methods and strict control of confounding features crystallize the POS processing network, spanning mainly across five brain regions – the STL, MTL, IFL, insula, and to some extent the supramarginal gyrus. These regions are rediscovered by both the encoding of POS-evoked modulations to the signal properties and the decoding of POS categories directly from the neural signal. In like manner, on one hand, a POS functional connectivity analysis finds the uncovered network to be markedly more inter-connected than its surroundings. On the other, clustering analysis of the connectivity map recapitulates a robust synchronization, restricted to the

same group of regions. Therefore, providing evidence that the above network is both exhaustive and exclusive. The two loops structure of the network (MTL-bottom and IFL-top), overlapping the STL and insula, along with the high STL and IFL internal connectivity suggest dependent yet separate POS understanding processes. A main short-term process that relies on local cues, such as auditory attributes and previous word lexical category, and an auxiliary long-term process supported by top-down modulation from the IFL.

Within the POS two-loop network the STG seems to act as the main processing core. The uniqueness of the core supports the shared POS processing network theory [254] and its proximity to low-level language areas (e.g. acoustic and phonological features [176, 240, 115, 114]) suggests POS assignment is a primal task in language processing. Specifically, the lateral aspect of the STG anticipates POS before word onset, performs the processing heavy-lifting, and broadcasts the information to auxiliary regions. POS predictability is presumably supported by POS of previous words, however, further validation is required to determine the source of this percussive signal. The temporal localization analysis only finds noticeable levels of POS activity in the STL and insula, probably due to the reduced window size, required for the temporal resolution of this experiment, which considerably diminishes the decoding power of the model.

In this work, we demonstrate a proof of principle for large-scale uncontrolled language datasets. These datasets can be flexibly reduced to highly controlled subsets, such as the noun-verb homonym set, to be reused for the investigation of additional linguistic tasks over high-quality intracranial data. We exhibit the applicability of the dataset by further fortifying previous claims of next-word prediction in language-specialized areas [231] and providing newly found insights on reversed effects in executive control regions. Specifically, we find multiple electrodes with strong positive surprisal correlations restricted to the insula and superior temporal lobe. Moreover, a few significant negatively correlated electrodes in the IFL and MFL allude to a potential "reset" operation of long-term prediction processes in cases of uncertainty across high-level language and multiple-purpose mechanisms.

By reusing uncontrolled data, researchers will be absolved from the need to collect their own. Thereby accelerating research and improving the reproducibility of results that will be extracted from the same large cohort.

4.5 Materials and Methods

4.5.1 Dataset construction

Stereoelectroencephalography (SEEG) neural recordings were collected from 10 subjects (5 male, 5 female), aged 4-19 (mean 11.9 σ ≈4.6), under treatment for epilepsy at Boston Children's Hospital (BCH); see Supplementary Figures table B.1 for per-subject statistics. All subjects were implanted with intracranial electrodes to localize seizure foci for potential surgical resection. All experiments were approved by BCH/Harvard IRB and were carried out with the subjects' informed consent. Electrode types, numbers, and positions were driven solely by clinical considerations.

4.5.2 Task and stimuli

Stimuli consisted of 21 recent animated/action Hollywood movies; see Supplementary Figures table B.2 for per-movie statistics. On average, movies were 2.07 hours long (σ ≈0.68) and contained 1322 sentences (σ ≈303), 8927 tokens (σ ≈2104), 1769 types (σ ≈324), 1358 unique lemmas (σ ≈259), 1219 nouns (σ ≈282), 615 noun types (σ ≈133), 1334 verbs (σ ≈299), and 504 verb types (σ ≈100). Movies were extracted from DVDs and are unchanged other than being re-encoded to a fixed frame rate (23.976 fps). Transcripts and all annotations described in this work will be made publicly available. Due to copyrights prohibiting the release of the raw stimuli (movies) source material, multiple audio-visual sample clips and tools allowing users to verify the alignment of their own movie copies will be publicly provided.

Each subject was given a choice of which movies to watch, viewing an average of 2.8 movies (σ ≈1.8) corresponding to 5.6 hours (σ ≈4.2). Movies were shown in full to each subject. Movies were displayed via a custom video player created in

Matlab 2018b. The player ensured that the presentation was at a fixed frame rate to keep the audio and video synchronized. The presentation of movies was accompanied by regular electrical triggers sent to the neural recording system to enable accurate temporal alignment between the movie and the neural data. A 15.4 inch (resolution 2880×1800) Apple MacBook Pro Retina was placed 60-100cm in front of the subject. Subjects adjusted the volume and paused/resumed the movie as needed. The movie was paused by the experimenter any time someone entered the room or when subjects were distracted and were resumed when subjects could direct their full attention back to the movie. Subjects could freely change position but were instructed by the experimenter, who watched the movies with the subjects, to remain focused on the stimulus or pause the movie. Subjects did not speak during the presentation of the movie nor did they overhear any other speech other than that found in the movie.

4.5.3 Data acquisition and signal processing

Clinicians implanted subjects with intracranial stereo-electroencephalographic (SEEG) depth probes containing 6-16 0.8 mm diameter 2 mm long contact electrodes (Ad-Tech, Racine, WI, USA) recording Intracranial Field Potentials (IFPs) with 1.5 mm separation. Each subject had multiple (12 to 18) such probes implanted in locations determined by clinical concerns entirely unrelated to the experiment. Electrodes placement was informed by a functional analysis [21]. The number of electrodes per subject ranged between 106 and 246 (μ =167, σ ≈40) for a total of 1688 total electrodes; see Supplementary Figures table B.1 for a per-subject breakdown. Data collected during periods of seizures or immediately following a seizure was discarded. Data was recorded using XLTEK (Oakville, ON, Canada) and BioLogic (Knoxville, TN, USA) hardware with a sampling rate of 2048 Hz. For each electrode, a notch filter was applied at 60 Hz and harmonics. No other processing (downsampling, filtering specific frequency bands, referencing, etc.) was performed on the neural recordings.

During the movie presentation, triggers were sent to a separate channel on the neural recording device via a USB connection to a dedicated trigger box (Measurement Computing USB-1208FS) using the Psychtoolbox 3 Matlab package. Each pulse

was logged with both its wall-clock timestamp and its movie timestamp. Individual triggers were sent every $100ms$. Specific events (movie start, pause, resume, and end) were marked by bursts of triggers (10, 8, 9, and 11 respectively) separated by 15ms. All triggers consisted of a 15ms electrical burst at a magnitude of 80mV. An automated tool found triggers and aligned the movie and neural data.

4.5.4 Word responsive electrode selection

As electrodes are placed according to medical needs, some electrodes may not record language processing-related activity. To be able to single out the subset of electrodes that are more likely to register linguistic activity, relevant to some aspects of this study, we took a straightforward threshold approach, independent from POS processing. We set a 40 μV min-max range criterion (selecting 115 electrodes overall; 10.66% of all electrodes) over the average $500ms$ electrical signal window of the sentence onset subset. Corrupted signal electrodes with extensive durations of static signal recordings were manually removed from consideration prior to any downstream analysis. Subjects with no word-responsive electrodes were removed from further analysis (subjects 4, 6, and 7 in Supplementary Figures table B.1).

4.5.5 Mean signal peak analysis

Considering the neural signal recorded by each electrode as the word processing local representation, we extracted the electrical current registered during the first $500ms$ post-word onset for every word in the corpus (see Methods section 4.5.3). We define the maximum voltage (in micro-volts) measured in this window as a word's peak amplitude and the corresponding time difference from the word's onset (in milliseconds) as its peak time. To investigate the dissimilarities and their progression through sentence utterance, we subdivided the data by their POS and location of the word in its context sentence. Specifically, the nouns and verbs in the dataset were grouped into words appearing at the beginning (onset), middle (midset), or end (offset) of a sentence (onsets: 153 ($\sigma \approx 44$) verbs, 72 ($\sigma \approx 26$) nouns; midsets: 1031 ($\sigma \approx 237$)

verbs, 699 ($\sigma \approx 191$) nouns; offsets: 161 ($\sigma \approx 45$) verbs, 471 ($\sigma \approx 106$) nouns per movie).

To obtain reliable statistical estimates and overcome artifacts from electrodes that did not register language-relevant information, for this analysis we only considered electrodes that were found to be word responsive, as defined in section 4.5.4. We directly compare the average peak amplitude and time distributions of nouns and verbs for every subset via a paired t-test (Python SciPy (1.3.0) Stats package `ttest_rel` function [256]). Then, to account for stochasticity and measurement artifacts we balance the number of nouns and verbs for each of the sentence subsets (by selecting a random subset out of the larger POS set) and compute the distribution ($\mu \pm \Sigma$) of all signals recorded per electrode during the utterance of either all nouns or all verbs per subject. We then sample 10,000 independent signal vectors from the multi-variate normal distribution (via the Numpy `multivariate normal` function of the `Random` library [118]) defined by the μ and σ vectors and extract a distribution of the mean peak time and amplitude. Each peak property is subtracted (verb - noun) per subset to estimate the representational difference.

Given the simulated peak differences, we compute the average non-parametric kernel dissimilarity density estimation (KDE) and 95% confidence intervals (2.5%-97.5% inter-quantile region) across the three subsets via the Statsmodels package `KDEUnivariate` function (kernel: gaussian, bw: scott, grid size: 512, fft: true) [227]. The KDEs are compared across sentence subsets for both peak time and amplitude jointly and independently.

4.5.6 Generalized linear model

A GLM was used to study features' contribution to neural signal properties. Two GLM variants were used in this work: 1) a simple linear model for subjects that watched a single movie trial, and 2) a random intercept linear mixed model for subjects that watched multiple movie trials. Both models were estimated via Python's Pymer4 package (0.7.0) [130] wrapping the R software Lme4 linear mixed effect model package [20]. Simple linear models were estimated via the Pymer4 `Lm` function. Ran-

dom intercept fixed mean linear mixed models were estimated via Pymer4 `Lmer` function, with an independent intercept per trial.

Independent models were estimated per electrode for each subject, pulling together all nouns and verbs of all movies a subject viewed. In addition to a word POS label (noun=0, verb=1) all scalar word features, described in Extended Figures table 4.1, were used as model regressors. Regressors' mean was subtracted from samples before analysis. Each set of coefficients served to construct two models, estimating either the signal peak time or amplitude separately. The significance of coefficients contribution was determined against each individual variable separately. Words with a peak time between $0ms - 50ms$ post-word onset were excluded to disregard instances where signal peaks are governed by previous word response or stochasticity. Electrodes' two-sided significance is determined over the FDR 2-stage Benjamini-Krieger-Yekutieli [24] corrected Fisher's combined [184] peak time and amplitude p-value.

4.5.7 Convolutional neural network model

A convolutional neural network (CNN) model was used to classify the nouns and verbs' lexical categories from the neural signal. Each signal sample of length m had a corresponding binary noun or verb label. The input provided to the model included the signal recorded by the desired electrode and its two adjacent neighbors located on the same SEEG probe, amounting to an input matrix of size $3 \times m$. Including signal from adjacent electrodes allows the model to flexibly optimize the re-referencing kernel, similarly to fixed methods (e.g., Laplacian re-referencing [154]) which also rely on adjacent signals. Electrodes at the SEEG probe edges had just one reference electrode, resulting in a $2 \times m$ input matrix.

The CNN model, designed in Python via the Pytorch (1.5.1) package [198], included 11 one-dimensional convolutional layers with skip connections and 2 fully-connected layers. All convolutional layers had a one-dimensional kernel of size 3, were batch-normalized, and had ReLU non-linear activation. The network input layer included 128 channels, and layers 2 and 5 doubled the number of channels to 256 and 512 respectively. In addition, layers 1, 2, 5, 8, and 11 had a stride of length

2 to reduce input length to $\frac{m}{32}$ at the last convolutional layer, with skip connections transferring information from the output of layers 2, 5, and 8 to the input of layers 5, 8, and 11 respectively. The last two fully-connected layers of the model reduced the convolutional layers flattened output matrix (of size $512 \times \lceil \frac{m}{32} \rceil$) to feature vectors of sizes 128 and then 2 for the final binary classification, with no non-linearity function. The predicted class was chosen to be the label corresponding to the cell with the higher value in the 2-dimensional output vector.

The model was trained for 10 epochs with Adam optimizer (learning rate: 10^{-4}, batch size: 32) over a single NVIDIA Titan RTX GPU (Cuda version 11.0). The data train-validation-test non-overlapping split ratio was 64%-16%-20% for all numerical test set balancing features and an 80%-20% split to train-validation sets after removing all possible matches for a textual test set balancing feature (see Methods section 4.5.8). The final model held-out test set accuracy was computed over the model version with the highest validation set score. The validation set was evaluated every 20 batches (640 samples) through training to select the top-scoring model version. All learning rates, batch size, epoch number, and number of batches between validations were selected to maximize validation performance via a grid search.

Nouns vs. verbs decoding

As the two largest POS open-class categories, nouns and verbs were selected to be a proxy for POS processing, providing the best balance between corpus magnitude and linguistic variety, as well as a highly controlled subset of homonym words (see Methods section 4.5.8). Due to their distinct linguistic nature, the noun class excluded proper nouns [72, 189] and the verb class excluded light verbs [46].

To account for the stochasticity of the CNN initialization and optimization process we retrained the model 5 times per data split and selected the test accuracy of the model with the best validation performance. To account for the stochasticity of the data split procedure, for every experiment (i.e. electrode and test set balanced feature), we defined the final experiment accuracy as the average across 5 independent data split reruns. Overall, performance for each experiment was based on 25 model

re-initializations across the 5 data splits.

In this study, we performed two types of decoding analyses, a static input window during a word utterance and a dynamic sliding window seeking predictive and lingering POS information. In the static test case, we provided the network with an extensive time window, capturing the majority of a word utterance (81.4% of nouns and verbs across all movies; $\mu = 357ms$, $\sigma \approx 185ms$) to optimize decoding accuracy. In these conditions, the window size was set to $500ms$ post word onset ($0ms - 500ms$, $m = 1024$ samples in $2048Hz$ sampling rate). In the dynamic test condition, we aimed to explore the rapidly evolving neural signal of word processing and minimize the dependence between adjacent windows. To that effect, the window size was set to $100ms$ ($m = 204$ samples) and 20 independent experiments were made between $500ms$ pre-word onset and $600ms$ post-word onset. Windows had $50ms$ overlap with their adjacent segments. The predicted label for all dynamic window experiments was the POS of the word beginning at time zero.

4.5.8 Test set construction

In this study we tested our trained models over a wide range of test sets, each one eliminates the effect of a potential confound by balancing the distribution of the confounding feature between the noun and verb test set subsets. Notably, in aggregate across all features, this procedure of training a model over uncontrolled train-validation sets while controlling the feature distributions of the test sets, assures that a model will provide an unbiased evaluation of the true POS information in the signal, as it is oblivious of the balanced confound. To construct the confound-balanced test sets, we considered two feature types: 1) numerical features (e.g. word length and surprisal estimates), that account for the vast majority of the features; and 2) identity features (e.g. dependency label and lemma).

Numerical feature distributions were balanced across a test set by randomly sampling 20% of the nouns per movie, and iteratively extracting the verbs with the closest Euclidean distance in the feature space. The feature space is either scalar discrete, scalar continuous, or multi-dimensional continuous, depending on the feature type.

Low-distance noun-verb pairs were prioritized to make the confound distributions of the noun and verb sets indistinguishable. Class sizes were balanced to obtain 50% chance level accuracy.

Due to their discrete and un-hierarchical nature, identity features were balanced based on a Boolean distance metric, where a verb-noun pair could either match or not. Since the language in the movie data is not optimized for any one task, some features (e.g. homonym set) were too strict to enable a test set with 20% of all samples. Therefore, we first extracted all matching pairs and then randomly sampled the test set out of the matched pairs if set sizes exceeded 20% of the data (otherwise, all found pairs were used for the test set). If a noun matched more than a single verb, a verb was randomly sampled out of all matches, and vice versa (homonym test set average size is 3.23% ($\sigma \approx$0.29%)).

If a subject watched more than one movie throughout the experiment, the train, validation, and test sets were first split per movie and then combined by group to form a single training dataset. The train, validation, and test sets were sampled to be mutually exclusive and all extracted test samples were held-out during training and validation, to be used only once at test time.

4.5.9 Decoding significance assessment

Static homonym set

To assess the statistical significance of the homonym set experiments we used a permutation test analysis per electrode, permuting all train labels and re-training in every iteration. We first estimated the number of permutations required for convergence over a randomly sampled subset of electrodes across multiple subjects. We found 120 permutations to be sufficient for the mean performance to remain bounded by $\pm$0.005 for 10 consecutive iterations. As an additional assurance of convergence, we doubled the number of permutations to 240 in our analyses.

To account for multiple comparisons when testing for the significance of different brain regions we used the pixel-based statistics method [62], taking any sepa-

rate electrode as an independent "pixel". We used the same random seed across all 240 permutations for all electrodes to create a null region permutation distribution. Specifically, for each of the 240 permutations we selected the highest accuracy across all of the electrodes in a given region per subject. An electrode region significance was determined as the percentile of its computed accuracy against its region's null distribution.

Dynamic homonym set

Since computing a full permutation distribution for all electrodes across all timing experiments is computationally intractable, in this study we pre-selected the subset of electrodes and corresponding time windows that showed high significance potential via a one sided binomial test (Python SciPy (1.3.0) Stats package *binom_test* function [256]). Comparison of p-values generated by the binomial and permutation tests over the static homonym set analysis found the first to be stricter in 82.8% of electrodes.

For every time window and every subject with n significant electrodes in a time interval (according to the binomial test) where $1 < n < 10$ for the given subject, we randomly sampled $10 - n$ additional electrodes from the subject's non-significant electrodes to construct a reliable time window null distribution. We then ran the permutation analysis as stated above over the pre-selected set of significant and randomly sampled electrode sets. A time window was found significant if it had at least one-time significant electrode.

Other test sets

Due to the computational intractability of computing a permutation distribution for each of our confound-balanced test sets, the significance of all non-homonym features was computed via a one-sided binomial test (Python SciPy (1.3.0) Stats package *binom_test* function [256]).

4.5.10 Held-out trial analysis

Considering only subjects that saw $k > 3$ movies we tested decoding performance over unseen stimuli in electrodes that were found significant across the homonym pairs and all confound sets analyses (see Methods section 4.5.8 and Supplementary Methods appendix B.1.4). In this experiment, a model was trained over the data extracted from 2/3 of the movies (rounded up to full movies, excluding all homonym word pairs) and tested over the homonym set of the remaining held-out 1/3 (1 movie for the 3 movie subjects and 2-3 movies for the 7 movie subject) in a k-fold approach. Final accuracy was computed as the average accuracy across all 3-folds. The significance of electrodes was determined through an additional permutation test, over 120 permutations. The accuracy of each permutation was determined as the average of 3 rounds to match the k-fold methodology.

4.5.11 Functional connectivity analysis

For each electrode, we subtracted the two immediately neighboring electrode IFPs (single neighbour for electrodes in probe edges) to remove any potential synchronization due to the common average reference. Non-overlapping time intervals of 1s (2048 samples) were extracted from nouns and verbs utterances or during segments with no annotated speech. The IFP for each time interval was normalized per electrode by subtracting its mean and dividing by the standard deviation.

The extracted language segments are $0ms - 1000ms$ interval post sentence onsets with no overlap to other included segments. No-language segments are non-overlapping 1s intervals extracted from no-speech durations across the movie, excluding the subtitles. The number of intervals per class was evened out by randomly downsampling the larger set.

The coherence in the responses of every pair of electrodes x and y at frequency f for every time interval was calculates using Welch's method [263]:

$$\frac{C_{xy_f} = |S_{xy}(f)|}{\sqrt{S_x(f)S_y(f)}}$$

Where $S_x(f)$ and $S_y(f)$ are power spectral density estimates of X and Y, and $S_{xy}(f)$ is the cross-spectral density estimate of x and y. Coherence and spectral densities were computed with Hann window and 50% overlap (Python `Coherence` function of the Scipy `Stats` library).

Coherence scores were averaged across frequencies for all time intervals and the distributions of average coherence scores for the language and no-language groups were compared via a one-sided t-test (Python `ttest_ind` function of the Scipy `Stats` library). A connection between two electrodes was found significant if $p \leq 0.05$, Bonferroni corrected for the number of un-directed electrode connections.

4.6 Extended Data Figures

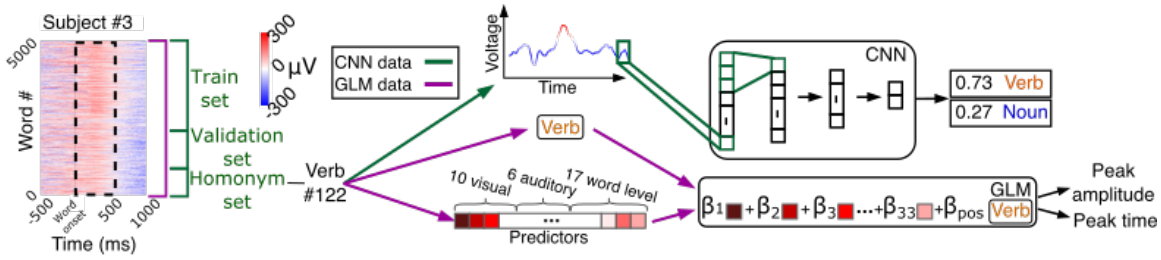

Figure 4-6: CNN (green) and GLM (purple) decoding experiments flow. We hold out all word pairs that appear as both a noun and a verb, train a CNN to decode the part of speech from individual electrodes, and test on the held-out homonym set. Conversely, a GLM predicts either the neural signal peak time or amplitude for all nouns and verbs given the POS as a predictor. Together these serve as detectors for which brain regions are sensitive to noun-verb distinctions.

#	Feature	Category	Type	Description
1	Max frame brightness	Visual	Scalar	The maximal brightness during word utterance computed as all pixels average HSV value
2	Min frame brightness	Visual	Scalar	The minimal brightness during word utterance computed as all pixels average HSV value
3	Max inter-frame difference	Visual	Scalar	A scene cut proxy The maximal inter-frame gray-scaled difference averaged over all pixels
4	Max global flow vector magnitude	Visual	Scalar	A camera motion proxy The maximal average dense optical flow vector magnitude
5	Max global orientation	Visual	Scalar	As above averaged over orientation (degrees) and selected by maximal magnitude
6	Max flow vector magnitude	Visual	Scalar	A large displacement proxy The maximal word utterance optical flow vector magnitude
7	Max flow vector orientation	Visual	Scalar	The orientation (degrees) of the above flow vector
8	Max mean flow vector magnitude	Visual	Scalar	The maximal mean magnitude flow vector per frame during word utterance
9	Max median flow vector magnitude	Visual	Scalar	The maximal median magnitude flow vector per frame during word utterance
10	Max number of faces	Visual	Scalar	The maximal number of faces per frame during word utterance
11	RMS (loudness)	Auditory	Scalar	Average root mean squared watts during the utterance
12	Mean magnitude	Auditory	Scalar	Average audio magnitude (dB) during word utterance
13	Mean pitch	Auditory	Scalar	Average pitch during word utterance
14	RMS difference	Auditory	Scalar	The difference in average RMS of the 500ms windows pre and post word onset
15	Mean magnitude difference	Auditory	Scalar	The difference in average magnitude of the 500ms windows pre and post word onset
16	Mean pitch difference	Auditory	Scalar	The difference in average pitch of the 500ms windows pre and post word onset
17	GPT-2 word surprisal	Surprisal	Scalar	Negative-log transformed GPT-2 word probability (given sentence preceding context)
18	LSTM word surprisal	Surprisal	Scalar	Negative-log transformed LSTM word probability (given sentence preceding context)
19	LSTM context entropy	Surprisal	Scalar	Word preceding context entropy computed over the LSTM probability distribution
20	GPT-2 most likely surprisal	Surprisal	Scalar	A context entropy complement Negative-log transformed GPT-2 most likely word probability
21	LSTM most likely surprisal	Surprisal	Scalar	A context entropy complement Negative-log transformed LSTM most likely word probability
22	N-gram word surprisal	Surprisal	Scalar	Negative-log transformed 5-gram word probability (given sentence preceding context)
23	Word infrequency	Surprisal	Scalar	A model-agnostic surprisal Negative-log normalized word frequency in the BLLIP corpus
24	Word time length	Length	Scalar	Word length (ms)
25	Word time difference	Length	Scalar	Difference between previous word offset and current word onset (ms)
26	Number of phonemes	Length	Scalar	Word number of perceptually distinct units of sound (phonemes)
27	Number of syllables	Length	Scalar	Word number of pronunciation units having one vowel sound (syllables)
28	Number of characters	Length	Scalar	Word number of characters
29	Sentence index in text	Location	Scalar	The index of the sentence containing the word in the movie transcript
30	Word index in text	Location	Scalar	The index of the current word in the movie transcript
31	Word index in sentence	Location	Scalar	The word index in its context sentence
32	Is sentence onset	Location	Scalar	True if the word is the first word in its context sentence and false otherwise
33	Head index	UD	Scalar	The index of the word's dependency tree head
34	Visual vector	Visual	Vector	The mean-normalized word utterance first frame ResNet-50 feature vector (size 2048)
35	Auditory vector	Auditory	Vector	Log transformed Mel-spectrogram flattened to a feature vector of size 6016 (127x47)
36	Form	UD	String	The word as annotated in the transcript
37	Lemma	UD	String	The stem of the word
38	Part of speech tag	UD	String	The word Universal Part-of-Speech (UPOS) tag
39	Dependency tag	UD	String	The head-word relation label in dependency tree

Table 4.1: All extracted word features. All scalar-type features were used as regressors in the GLM analysis and all scalar and vector features were used as test set balancing features in the multi-confounds CNN analysis.

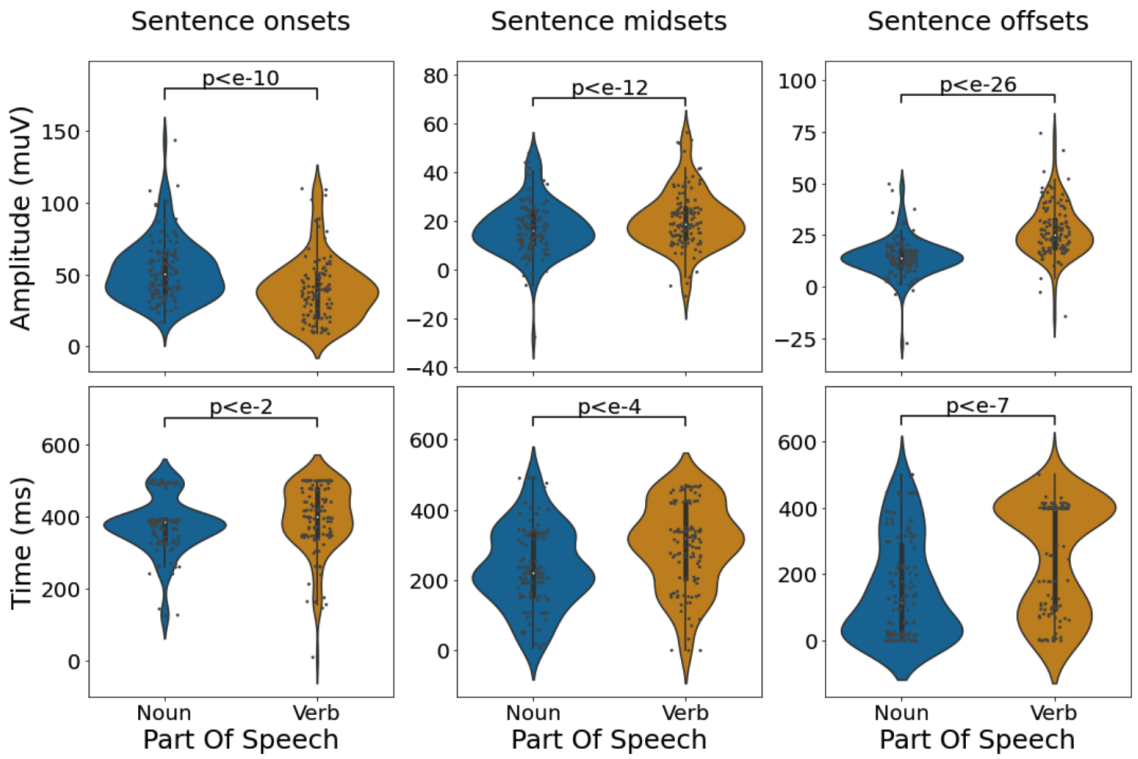

Figure 4-7: Nouns and verbs average signal peak attribute comparison across sentence progression. Peak amplitude (top row) and peak time (bottom row) of the average signal were computed separately per electrode for nouns (blue) and verbs (orange) at the on-set (left), mid-set (middle), and off-set (right) of their context sentences. The distributions of nouns and verbs were computed based on the set of 115 (10.66% of all electrodes) word-responsive electrodes and compared for every subset via a paired t-test.

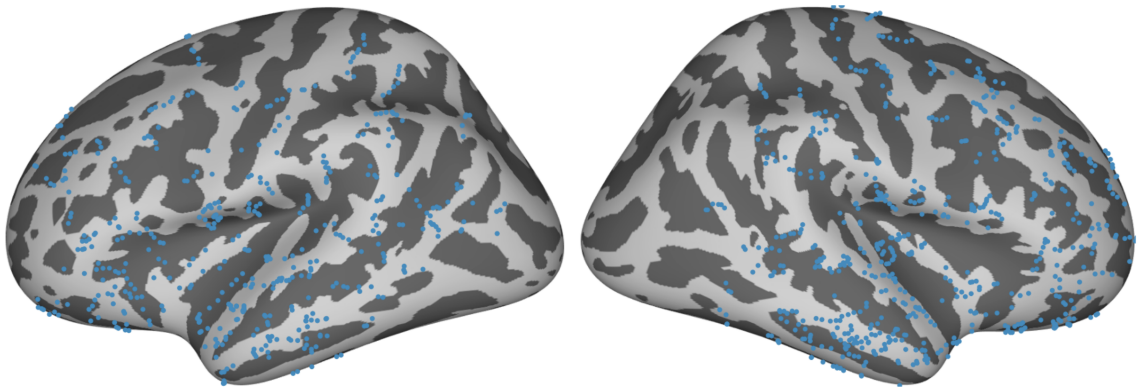

Figure 4-8: All electrode locations from the 7 subjects analyzed in this study projected on the temporal aspects of the average inflated brain. Electrodes located more than 1.5mm away from predefined gray matter regions were removed from the analysis.

Chapter 5

Multi-resolution modeling of a discrete stochastic process identifies causes of cancer

5.1 Summary

Detection of cancer-causing mutations within the vast and mostly unexplored human genome is a major challenge. Doing so requires modeling the background mutation rate, a highly non-stationary stochastic process, across regions of interest varying in size from one to millions of positions. Here, we present the split-Poisson-Gamma (SPG) distribution, an extension of the classical Poisson-Gamma formulation, to model a discrete stochastic process at multiple resolutions. We demonstrate that the probability model has a closed-form posterior, enabling efficient and accurate linear-time prediction over any length scale after the parameters of the model have been inferred a single time. We apply our framework to model mutation rates in tumors and show that model parameters can be accurately inferred from high-dimensional epigenetic data using a convolutional neural network, Gaussian process, and maximum-likelihood estimation. Our method is both more accurate and more efficient than existing models over a large range of length scales. We demonstrate the usefulness of

multi-resolution modeling by detecting genomic elements that drive tumor emergence and are of vastly differing sizes.

5.2 Introduction

Numerous domains involve modeling highly non-stationary discrete-time and integer-valued stochastic processes where event counts vary dramatically over time or space. An important open problem of this nature in biology is understanding the stochastic process by which mutations arise across the genome. This is central to identifying mutations that drive cancer emergence [151].

Tumor drivers provide a cellular growth advantage to cells by altering the function of a genomic element such as a gene or regulatory feature (e.g. promoter). Drivers are identifiable because they reoccur across tumors, but there are two major challenges to detecting such recurrence. First, driver mutations are rare and their signal is hidden by the thousands of passenger mutations that passively and stochastically accumulate in tumors [244, 167]. Second, because functional elements vary dramatically in size (genes: 10^3-10^6 bases; regulatory elements: 10^1-10^3 bases; and single positions), driver mutations accumulate across regions that vary by many orders of magnitude. Accurately predicting the stochastic accumulation of passenger mutations at multiple scales is necessary to reveal the subtle recurrence of driver mutations across the genome.

In this chapter, we introduce the split-Poisson Gamma (SPG) process, an extension of the Poisson-Gamma distribution, to efficiently model a non-stationary discrete stochastic process at numerous length scales. The model first approximates quasi-stationary regional rate parameters within small windows; it then projects these estimates to arbitrary regions in linear time (10-15 minutes for genome-wide inference). This approach is in contrast to existing efforts that model fixed regions and require computationally expensive retraining (e.g. over 5 hours) to predict over multiple scales of interest [191, 169]. We apply our framework to model cancer-specific mutation patterns (fig. 5-1). We perform data-driven training of our model's parameters

and show that it more accurately captures mutation patterns than existing methods on simulated and real data. We demonstrate the power of our multi-resolution approach by identifying drivers across functional elements: genes, regulatory features, and single base mutations. Despite the method having no knowledge of genome structure, it detects nearly all gene drivers present in over 5% of samples while making no false discoveries and detects all previously characterized regulatory drivers. Detected events also include novel candidate drivers, providing promising targets for future investigation.

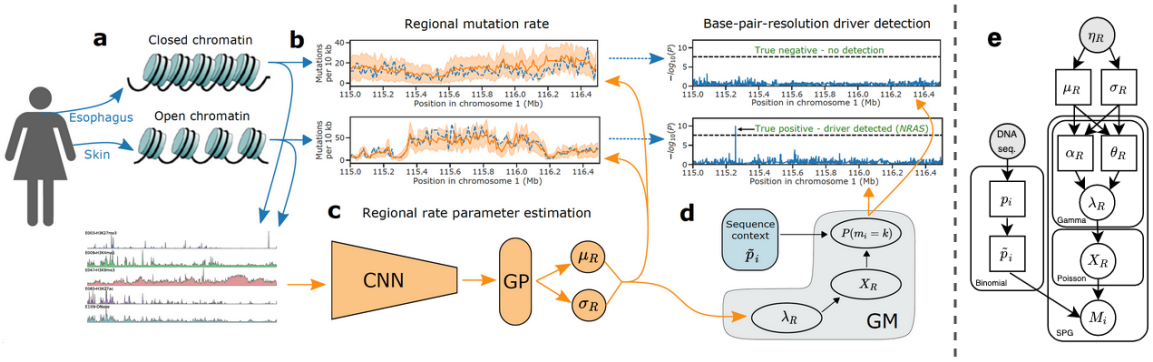

Figure 5-1: Non-stationary stochastic process modeling predicts mutation patterns and identifies cancer-specific driver mutations. Biological processes are shown in blue, data processing is shown in orange. **a.** Areas of the genome have varying epigenetic states (e.g. accessibility for transcription) depending on the tissue type. **b.** These epigenetic states set different mutation rates in different tissues. **c.** Our model takes these epigenetic tracks as input to estimate the regional mutation density across the genome (95% confidence interval in orange). **d.** Regional rate parameters and sequence context are integrated via the split-Poisson-Gamma (SPG) distribution to provide arbitrary resolution mutation count estimates. Deviations between the estimated and observed mutation rates identify mutations that are associated with cancers in different tissues. **e.** The split-Poisson-Gamma (SPG) model plate diagram (squares: inferred parameters; grey: observed input data).

5.2.1 Previous work

Numerous methods exist for modeling stationary stochastic processes [159]. Far fewer exist for non-stationary processes because they are difficult to capture with the covariance functions of parametric models [216]. Non-stationary kernels have been introduced for Gaussian processes [197], but these may not be tractable on large

datasets due to their computational complexity. More recently, there has been work developing Poisson-gamma models for dynamical systems [224, 110], but these methods have focused on learning relationships between count variables, not predicting counts based on continuous covariates.

In the particular case of modeling mutation patterns across the cancer genome, numerous computational methods exist to model mutation rates within well-understood genomic contexts such as genes [151, 169, 257, 187, 135]. These models account for $< 4\%$ of the genome [214]. They are not applicable in non-coding regions, where the majority of mutations occur [103]. A handful of methods to model genome-wide mutation rates have been introduced [200, 191, 26]. However, they operate on a single length-scale or set of regions and require computationally expensive retraining to predict over each new length-scale. Several methods rely on Poisson or binomial regression; however, previous work has extensively documented that mutation counts data are over-dispersed, leading these models to underestimate variance and yield numerous false-positive driver predictions [161, 169, 136]. Negative binomial regression has recently been used to account for over-dispersion [191] and perform genome-wide mutation modeling and driver detection. However, resolution was coarse, and it only found a few, highly recurrent driver mutations.

5.2.2 Our contributions

This work makes three key contributions: 1) we introduce an extension of the Poisson-Gamma distribution to model non-stationary discrete stochastic processes at any arbitrary length scale without retraining; 2) we apply the framework to capture cancer-specific mutation rates with unprecedented accuracy, resolution, and efficiency; and 3) we perform a multi-scale search for cancer driver mutations genome-wide, including the first-ever base-resolution scan of the whole genome. This search yields several new candidate driver events in the largely unexplored non-coding genome, which we are working on validating with experimental collaborators. Crucially, our approach allows fast, efficient, and accurate searches for driver elements and mutations anywhere in the genome without requiring arduous retraining of a model, a feat which is

not possible with existing approaches.

5.3 Materials and Methods

5.3.1 Multi-resolution modeling of a non-stationary discrete stochastic process

We consider a non-stationary discrete stochastic process $\{M_i;\ i = 1, 2, ...\}$ where M_i is the integer-valued event count at position i. Associated with each position i is a real-valued, L-dimensional feature vector η_i that determines the instantaneous event rate λ_i via an unknown function. Thus a region $R = \{i, i+1, ..., i+N\}$ of N contiguous positions is characterized by an $L \times N$ feature matrix η_R and an event count $X_R = \sum_{i \in R} M_i$. As training data, η_R, X_R, and M_i are observed for some set of regions $\{R \in \mathcal{T}\}$. Then given a set of feature matrices from unobserved regions $\{\eta_R;\ R \in \mathcal{H}\}$, the challenge is to predict the distribution of event counts over any arbitrary set I of unseen positions that may or may not be contiguous. Real-world examples include traders in a stock market, packets delivered to routers in a network, and mutations accumulating at positions in the genome.

The split-Poisson-Gamma process

We assume that the process is near-stationary within a small enough region $R = \{i, i+1, ..., i+N\}$ and that the $L \times N$ covariate matrix η_R is observed. Thus the rate of events λ_R within R is approximately constant and associated with η_R, albeit in an unknown way. A number of events (X_R) may occur within R dependent on λ_R and are then stochastically distributed to individual positions within R, implying a hierarchical factorization of the scalar random variables λ_R, X_R, and M_i (fig. 5-1e) as

$$Pr(M_i = k, X_R, \lambda_R; \eta_R) = Pr(M_i = k | X_R; \eta_R) Pr(X_R | \lambda_R; \eta_R) Pr(\lambda_R; \eta_R). \quad (5.1)$$

X_R and λ_R are unknown nuisance variables and are marginalized in general as

$$Pr(M_i = k|\eta_R) = \int_0^\infty Pr(\lambda_R; \eta_R) \sum_{X_R=k}^{\infty} Pr(M_i = k|X_R; \eta_R) Pr(X_R|\lambda_R; \eta_R) \mathrm{d}\lambda_R. \tag{5.2}$$

Since applications often require many posterior predictions over regions of varying sizes, we propose a prior parameterization that builds on the success and flexibility of the classical Poisson-Gamma distribution while ensuring the marginalization has an easy-to-compute posterior distribution:

$$\lambda_R \sim \text{Gamma}(\alpha_R, \theta_R) \tag{5.3}$$

$$X_R \sim \text{Poisson}(\lambda_\text{R}) \tag{5.4}$$

$$M_i \sim \text{Binomial}(X_R, \tilde{p_i}) \tag{5.5}$$

where α_R and θ_R are shape and scale parameters dependent on η_R, p_i is the time-averaged probability of an event at i and $\tilde{p_i} = \frac{p_i}{\sum_{j \in R} p_j}$, the normalized probability within R. A plate diagram of the hierarchical model is presented in fig. 5-1e.

The above formulation provides a simple, closed form solution to eq. (5.2) as a negative binomial (NB) distribution (See Appendix for details):

$$Pr(M_i = k|\alpha_R, \theta_R, \tilde{p_i}; \eta_r) = NB\left(k; \alpha_R, \frac{1}{1 + \theta_R \cdot \tilde{p_i}}\right). \tag{5.6}$$

Eq. 5.5 implicitly assumes that events are distributed independently to units within R. Exploiting this assumption, eq. (5.6) immediately generalizes to consider *any* set of units $I \subseteq R$ as

$$Pr\left(\sum_{i \in I} M_i = k|\alpha_R, \theta_R, \{\tilde{p_i}\}_{i \in I}; \eta_R\right) = NB\left(k; \alpha_R, \frac{1}{1 + \theta_R \cdot \sum_{i \in I} \tilde{p_i}}\right). \tag{5.7}$$

The above formulation is an extension of the classical Poisson-Gamma distribution whereby the Poisson is randomly split by a binomial. We term this a split-Poisson-Gamma (SPG) process. While the derivation of the SPG solution makes simplifying

assumptions, the benefit is that the parameters α_R and θ_R need to be estimated only once for each non-overlapping region R. *Estimates for a region of any other size can then be computed in constant time* from eq. (5.7). If a new region R' is larger than R, we approximate the gamma distribution in a super-region containing R' as a superposition of the previously inferred parameters of each region of size R within the super-region (see section 5.3.1).

Theoretical underpinnings of parameter inference

Inferring regional rate parameters The statistical power of SPG depends on the accurate estimation of the regional gamma rate parameters α_R and θ_R. We propose a variational approach to enable flexible, accurate, and non-linear inference of these parameters from a set of covariates. Let $G(\alpha, \theta)$ be a gamma distribution. By the central limit theorem, $\lim_{\alpha\to\infty} G(\alpha, \theta) = N(\mu, \sigma^2)$ where $\mu = \alpha\theta$ and $\sigma^2 = \alpha\theta^2$. We thus use a Gaussian process (GP) to non-linearly map covariates to regional estimates for μ_R and σ_R^2. The variational estimates for the gamma parameters are then

$$\alpha_R = \mu_R^2/\sigma_R^2, \qquad \theta_R = \mu_R/\sigma_R^2 \tag{5.8}$$

For a super-region $R' = R_i + R_j$, $\mu_{R'} = \mu_{R_i} + \mu_{R_j}$ and $\sigma_{R'}^2 = \sigma_{R_i}^2 + \sigma_{R_j}^2$.

A limitation of this approach is that GPs can only operate on vectors of covariates. Thus a dimensionality reduction method must be applied to the input matrix η_R. In cases where η_R includes spatial relationships, a convolutional neural network can be a powerful approach to dimension-reduction; however, other approaches are feasible (see section 5.3.2 and section 5.4.1).

Inferring time-averaged event probabilities The time-averaged parameters $\{p_i;\ i = 1, 2, ...\}$ must also be inferred. Crucially, as seen in eq. (5.5), these parameters are never used directly; instead, they are always renormalized to sum to one within a region of interest. Thus, estimates do not need to reflect the absolute probability of an event at i but merely the *relative* rate of events between positions.

Indeed, because of the renormalization procedure, the estimates need not even be a true probability distribution. Estimating p_i can thus be accomplished by clustering units with similar relative rates of events. How this clustering should be performed will depend on the application of interest (see section 5.3.2 for a concrete example).

5.3.2 Fitting parameters to predict cancer mutation patterns

We obtained publicly available mutation counts from four cancer cohorts previously characterized by the Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG) [51]: esophageal adenocarcinoma (N = 98 tumors; n $\approx$ 2.7M mutations), skin melanoma (N = 70 tumors; n $\approx$ 7.8M mutations), stomach adenocarcinoma (N = 37 tumors; n $\approx$ 480k mutations), and liver hepatocellular carcinoma (N = 264 tumors; n $\approx$ 3.3M mutations). Crucially, these data contain only the total number of mutations at each position in the genome. *We do not know a priori which mutations are background mutations and which are driver mutations. We also do not know the true mean and variance of the underlying mutation rate in any region.*

We do know that the mutation rate is highly associated with chemical modifications of the DNA that set the way it is processed in a cell, collectively termed the epigenome [226, 200]. We obtained 733 datasets characterizing the patterns of these chemical modifications in 111 human tissues from Roadmap Epigenomics [218]. These data are the largest compendium of uniformly processed human epigenome sequencing currently available. Each track provides the -$\log_{10}$ P-value that a particular modification is present at each location of the genome in a given tissue type. We additionally created two tracks that provide the average nucleotide and GC content in a region based on the human reference genome GRCh37. See Appendix and supplementary data for additional information on the epigenetic tracks. The input matrix for each region η_R thus has 735 rows. We fixed the number of columns to be 100 irrespective of the size of R, where each column is the mean across $R/100$ adjacent positions.

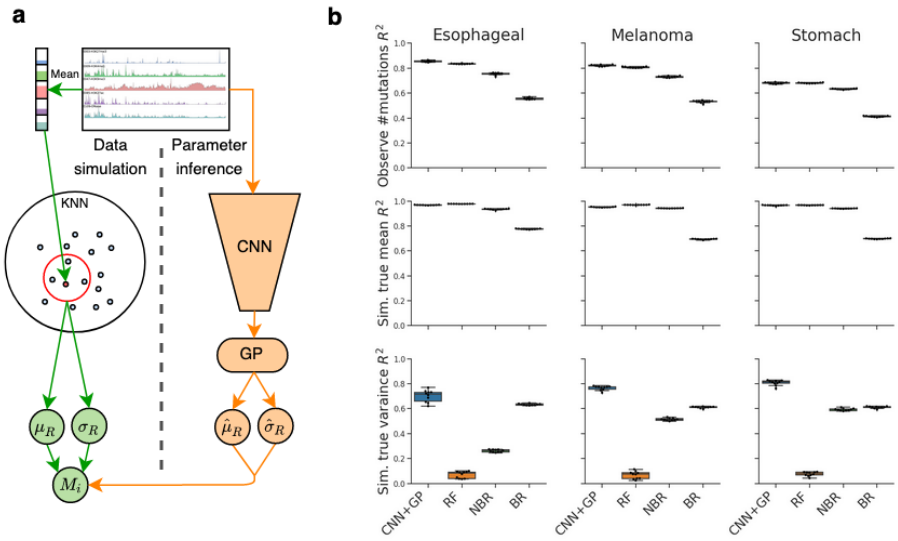

Figure 5-2: Data simulation and regional parameters inference accuracy across methods. **a.** Simulated data experiment. Data simulation is shown in green, parameter inference is shown in orange. Simulated μ_R and σ_R are computed as the mean and variance of a sample's KNN cluster. The model is trained over randomly sampled event counts; the parameter estimates $\hat{\mu}_R$ and $\hat{\sigma}_R$ are then compared to their true values. **b.** Pearson R^2 to the observed mutation count in the true (**unsimulated**) data (top) simulated mean (middle) and simulated variance (bottom) for the CNN+GP parameter estimation strategy (CNN+GP), random forest (RF), negative binomial regression (NBR) and binomial regression (BR). Results for additional estimation techniques are in Appendix.

Artificial dataset

In order to evaluate the ability of SPG and other models to estimate the unknown mean and variance of regional rates, we created simulated datasets with known mean and variance parameters dependent on the observed input matrix (fig. 5-2a). We created input matrices of size 735×100 from the epigenetic tracks (described above) for non-overlapping regions of 50,000 positions. To define the non-stationary mean and variance of mutation rate dependent on each region's input matrix, we reduced η_R to a feature vector of size 735 by taking the mean across columns and used a k-nearest-neighbors (KNN) strategy to identify 500 regions with similar epigenetic feature vectors; we then defined μ_R and σ_R^2 for each region as the mean and variance of the observed event counts across its 500 neighboring regions. The number of observed events for that region was then randomly drawn from a negative binomial distribution defined by those parameters (full technical details in Appendix). Models were trained

on the randomly drawn counts and evaluated on their ability to accurately infer the true mean and variance. We simulated 50kb regions following previous work [214].

Estimating dynamic regional rates with uncertainty

The input matrices $\eta_R \in \mathbb{R}^{735\times100}$ required significant dimension reduction before we could employ our GP-based variational strategy to infer SPG regional rate parameters. Columns encode the high-resolution spatial organization of the epigenome which have recently been shown to be important determinants of local mutation rate [106, 9]. Therefore, we hypothesized that a convolutional neural network (CNN) would provide a powerful approach to produce a low-dimensional embedding that retains information about this local structure; the supervised nature of a CNN further enables the resulting embedding to be optimized for the cancer of interest, which is crucial to performance since the epigenetic determinants of mutation rate vary drastically between cancer types [200]. We constructed a 1D CNN model with 4 residual blocks and 3 fully-connected layers to map mutation-rate-associated local epigenetic patterns to regional mutation rates. The CNN non-linearly reduces $\eta_R \in \mathbb{R}^{735\times100}$ to a 16-dimensional feature vector in its last feature layer. The CNN was trained to minimize the mean squared error between observed and predicted mutation counts. Due to the interchangeable nature of the rows, the 1D kernels allow the network to identify arbitrary inter-track interactions. The final 16-dimension feature vector was then passed as input to a sparse GP [250], fit to maximize the likelihood of the observed mutation counts using 2000 inducing points and a radial basis function kernel (fig. 5-1b). We found that results were robust to the particular choice of kernel and hyperpriors placed over kernel parameters. While end-to-end training is possible [36], we did not find it necessary to achieve high accuracy in this particular application. A CNN is not the only method available to reduce dimensionality prior to GP inference; we investigated numerous other methods but found the CNN+GP to produce the most accurate results (see Appendix).

Estimating time-averaged event probabilities

In the case of cancer mutation patterns, previous work showed that the mutation rate at any position i is heavily influenced by the nucleotide at i and the two nucleotides directly adjacent to i; positions with this same "trinucleotide context" will have similar mutation patterns [11]. Following previous works [187, 257, 169, 262], we used trinucleotide context to estimate p_i. Let ntn' be the trinucleotide context centered at position i. We estimate the probability that i is mutated using the ensemble maximum-likelihood estimate of its cluster

$$p_i = p_{n,t,n'} = \frac{v_{n,t,n'}}{N_{n,t,n'}}. \tag{5.9}$$

where $N_{n,t,n'}$ is the number of ntn' trinucleotides in the genome and $v_{n,t,n'}$ is the number of times t is mutated within ntn'. This approach alone explains little variance in sub-megabase regions (see Appendix) because it does not account for regional mutation rates.

Comparing to benchmark models

We compared SPG to three alternative approaches that have been previously used to learn both the mean and variance of regional mutation patterns genome-wide. The alternative models are random forest (RF) regression [200], binomial regression (BR) [26], and negative binomial regression (NBR) [191, 169]. For the RF, we used the Jackknife method [258] to estimate the variance; this method requires $O(n)$ trees where n is the number of samples in the training set. BR and NBR directly specify the variance as a function of the mean: BR as $\sigma_R^2 = \mu - \mu^2/n$ and NBR as $\sigma_R^2 = \mu_R(1 + \beta\mu_R)$, where β is an overdispersion parameter. Benchmarking comparisons were performed on the skin melanoma, esophageal adenocarcinoma, and stomach adenocarcinoma cohorts.

Model training

For every region of size R, epigenetic features were extracted into matrices of size 735 tracks by 100 binned position columns, where each column was the mean across $R/100$ adjacent base-pairs. Regions with highly repetitive DNA sequences (<70% of 36mer sub-sequences being unique) were excluded from the training set to ensure high data quality as in previous analyses [200]. Before training, high-quality data regions were strictly split into a train (64%), validation (16%) and test (20%) sets. Predictions for excluded regions and held-out test sets were obtained after model training. Genome-wide predictions were generated using 5-fold cross-validation. The CNN received the full 735×100 matrices as input. Vector-based methods (RF, NBR, BR) received the 735-dimension vector of epigenetic values averaged across position columns. Following previous work, we also included the expected number of mutations based on the trinucleotide composition of a region as an offset term in NBR and BR when predicting mutation counts [191]. Additional details on training (e.g. number of epochs) are in Appendix.

5.3.3 Identifying genetic drivers of cancer

Because cancer drivers reoccur across tumors, driver elements (genes, regulatory structures, and individual base-pairs) will contain an excess of mutations relative to the expected background mutations. The SPG model provides a simple, efficient, and accurate method to search for this recurrence. We first estimate the mean and variance of the background mutation rate using the CNN+GP estimation method. We then apply eq. (5.7) to search for statistical evidence that the number of observed mutations, k, exceeds expectation within every gene, known regulatory structure, and 50 bp window in the genome by changing the set of tested positions I. For a gene, k is the number of observed missense or nonsense mutations and I is the set of all possible mutations in the gene. For both a regulatory element and a window of fixed size, k is the number of mutations observed in the element / window and I is the set of all positions within the element / window. If an element overlaps multiple 10kb regions,

we merge the mean and variance estimates for overlapped regions as described in section 5.3.1. To maintain strict train-test separation, both the rate parameters and p_i are estimated excluding the element being tested. We controlled the family-wise error rate at the $\alpha = 0.05$ level using a Bonferroni correction for the total number of tests in genes, regulatory elements, or 50bp windows. Gene information was obtained from [169] and regulatory element information from [214]. Driver detection was performed in all four cancer cohorts.

5.4 Results

5.4.1 Accuracy of regional rate parameter estimation

We first evaluated various methods' abilities to infer regional rate parameters, considering both new (CNN+GP) and existing (RF, NBR, BR) methods. We assessed each method's ability to learn the expected mutation rate by directly assessing the amount of variance (Pearson R^2) it explained over observed mutation counts in 50kb real data windows (fig. 5-2b top), and found the CNN+GP estimation method performed the best, although random forest was a close second (results were similar when estimating the mean in simulated data; fig. 5-2b middle). We then evaluated each method's ability to capture the variance σ_R^2 in the simulated data, quantified as the Pearson R^2 to the true variance. The CNN+GP method again outperformed the others (fig. 5-2b bottom). Notably, RF was unable to infer the variance beyond chance level, and thus we did not consider this method further because its inability to infer variance precludes accurate driver detection.

We also considered other dimensionality reduction techniques including both non-neural and neural approaches as well as supervised and unsupervised approaches, as an alternative to the CNN; no other approach achieved accuracy comparable to the CNN+GP over both mean and variance (see Appendix). Moreover, we validated the necessity of the GP by directly optimizing the CNN to predict both parameters and found it significantly reduced model performance (7% decrease over mutation counts

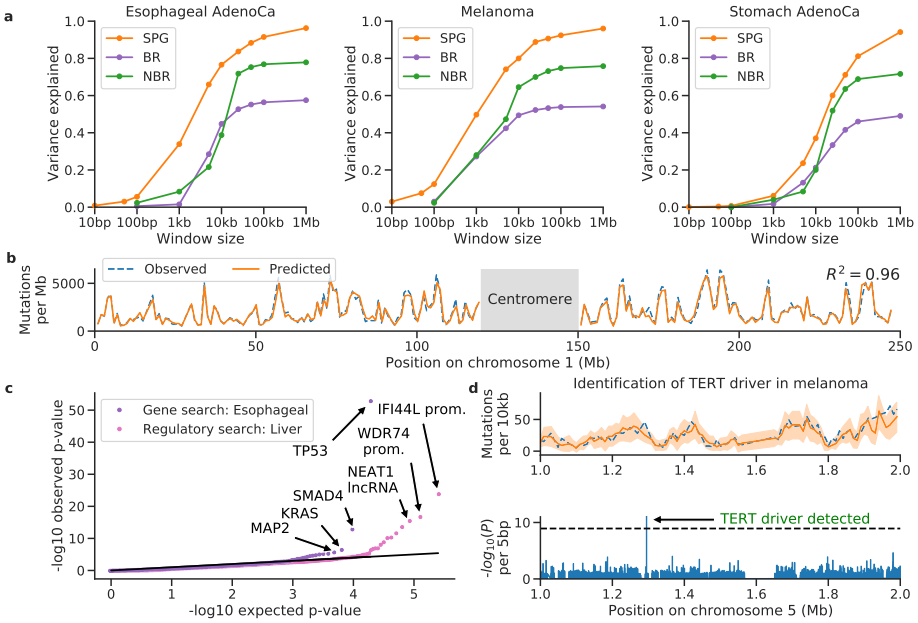

Figure 5-3: SPG accurately models mutation density and detects driver events. **a.** Variance explained (Pearson R^2) of the observed mutation count by our SPG, binomial regression (BR) and negative binomial regression (NBR) across length-scales. **b.** Observed (dashed blue) and predicted (solid orange) mutation density from our GM at 1Mb regions across chromosome 1 in melanoma. (**c**) Quantile-quantile plots of expected and observed P-values for gene driver detection in esophageal adenocarcinoma (purple) and driver regulatory element detection in liver cancer (pink). Esophageal and liver were chosen only for the sake of readability; qq-plots are similar for all cancers. (**d**) Model detection of a well-known non-coding driver in the *TERT* promoter in melanoma at 1kb resolution. Black dashed lines: Bonferroni-corrected genome-wide significance thresholds.

and 13% over σ_R^2 within 10kb windows in melanoma).

5.4.2 Accuracy and efficiency of mutation rate prediction

To further compare the SPG performance to existing methods, we evaluated the accuracy and efficiency of each method over length scales ranging over 5 orders of magnitude (10-10^6 positions). To evaluate SPG, we estimated the background mutation rate parameters, μ_R and σ_R^2, in 10kb regions genome-wide using the CNN+GP estimation strategy; we then applied the SPG distribution to estimate mutation count distributions over all other region sizes. The existing methods with reasonable performance on both mean and variance prediction (BR and NBR) were trained to directly predict the count distribution in each region for each length scale genome-wide.

Across all tested window sizes and cancers, SPG outperformed existing methods, with performance particularly improved in esophageal adenocarcinoma and skin melanoma (fig. 5-3a), crucial for high-accuracy driver detection downstream. Across 1Mb windows, SPG explains $> 95\%$ of the variance in mutation density across all three cancers (fig. 5-3a,b); this is $>15\%$ more variance than both existing methods (Fig. 5-3a), highlighting the ability of SPG to accurately capture regional distribution parameters and project them upwards. The decrease in variance explained in smaller window sizes is expected because observed mutation counts become increasingly stochastic relative to the expected number of mutations predicted by each method. The theoretical foundations of negative binomial regression and SPG are similar, both are built upon the classical Poisson-gamma model. SPG differs from NBR in three key ways that help explain its improved performance: 1) SPG models mutation patterns over arbitrary sets of positions enabling it to dynamically pool information across positions after a single training; in contrast, NBR operates on fixed regions, and must be retrained for every new region size. 2) SPG's variational inference method estimates the gamma parameters for each region independently; NBR estimates only the shape parameter independently for each window and uses a single scale parameter for all windows. 3) SPG's CNN data reduction enables non-linear mapping of spatial covariate information to mutation rate, whereas NBR can perform only linear inference and disregards the spatial organization of the genome.

SPG is also the most efficient method for multi-resolution search (appendix C.2.3). Initial training of parameters using the CNN+GP method for one fold of 10kb regions required 36 minutes using 1 GPU. Projection to each additional scale using 8 CPUs required at most 4 minutes (table 5.1). In contrast, training time for BR and NBR increases considerably as the resolution decreases. Performing a search across resolutions of 50bp, 100bp, 500bp, 1kb, and 10kb would require >5h for negative binomial, >2h for BR, and only 52 minutes for SPG (Appendix). We have also found that parameter estimation on windows as large as 100kb does not significantly reduce accuracy across scales (Appendix), allowing SPG parameter estimation in a considerably shorter time (e.g. only 8 minutes for 50kb).

Method	100bp	1kb	10kb	100kb	1Mb	Multi-Scale
SPG	4m11s	3m33s	36m35s	19s	2s	**42m40s**
NBR	1h30m	7m3s	43s	6s	4s	1h37m56s
BR	44m36s	3m8s	15s	5s	4s	48m8s
RF	>15h	>15h	14h2m	5m24s	28s	>15h

Table 5.1: Run times for SPG, NBR, BR, and RF for five region sizes and multi-scale search. Reported times are for a single train-validation-test split per model over 8 CPUs and 1 GPU machine. For SPG, parameters were inferred using the CNN+GP estimation method at 10kb, running the CNN and GP one time each; hence SPG's increased run time at 10kb relative to other region sizes. Bolded is the best multi-scale search time. Presented RF times are high due to the need for O(n) trees to estimate variance using the Jackknife method.

5.4.3 Identification of cancer driver mutations

We leveraged SPG's ability to model multiple resolutions to search the whole genome of each of the four cancer cohorts for gene drivers, non-coding regulatory drivers, and 50bp windows that may harbor a driver mutation. All significant results are provided as supplementary data tables. We compared our results to those obtained from a previous comprehensive characterization of these cohorts by [51], who used 13 different methods to identify drivers. Our model did not have access to information about gene structure or function unlike the methods used in the previous characterization. Nonetheless, the model's p-values were well calibrated (fig. 5-3c), and we identified 19 genes with a significant excess of missense or nonsense mutations. All 19 genes were previously reported as drivers by [51]. We failed to detect only two known driver genes present in >5% of samples. This performance is on par with state-of-the-art methods specifically designed for driver gene identification [214].

When analyzing non-coding regulatory elements, SPG's p-values were again well calibrated (fig. 5-3c), and it identified all non-coding drivers (n=11) identified by [51]. Moreover, SPG implicated several additional putative non-coding driver elements that had not been previously reported. Examples include 1) the promoter of the gene *MTERFD1* in esophageal cancer ($P = 3.1 \times 10^{-8}$), whose over-expression has been observed in numerous cancers, has been shown to promote cell growth, and decrease clinical survival [274]; 2) an enhancer of *DHX33* in liver cancer ($P = 4.8 \times 10^{-11}$), whose over-expression has been shown to promote cancer development [259]; and 3)

the 5' UTR of *ERN1* in melanoma, which has been linked to cancer therapy resistance [281].

Finally, we performed the first, to our knowledge, genome-wide search for individual driver mutations. All significant genic hits fell within known driver genes whose functions have been experimentally validated including *TPF3*, *BRAF*, *KRAS*, *PIK3CA*, and *CTNNB1*. In addition, SPG identified two recurrent mutations in the genes *GPR98* and *KLB* that had not been previously identified in [51]'s analysis of the data. These mutations are listed as driver mutations in the Catalogue of Somatic Mutations in Cancer [248]. SPG implicated numerous hotspots in the mostly unexplored non-coding genome, including the well-known *TERT* promoter mutation (fig. 5-3d). These results are promising targets for future studies of non-coding drivers in cancer cell lines and organoids.

5.5 Discussion

We introduced an extension of the Poisson-Gamma distribution to model discrete-time, integer-valued stochastic processes at multiple scales. The split-Poisson-Gamma (SPG) model makes several simplifying assumptions including: 1) that the process is quasi-stationary in a small enough region; 2) events are distributed among the discrete units approximately independently; and 3) the behavior of the random variables can be captured by particular parametric distributions. The assumptions are necessary to derive a closed-form posterior distribution. This enables efficient prediction over multiple length-scales without having to re-estimate the model parameters. We additionally proposed a variational inference strategy to reduce input dimensionality and estimate the parameters of the model using a CNN coupled with a GP. Indeed, the use of a CNN+GP to perform variational inference for a distribution of interest may be of use well beyond the SPG framework and discrete stochastic process modeling.

To demonstrate the utility of the SPG, we applied it to model mutation rates in cancer and identify genomic elements that drive tumor emergence. In the case of this application, previous work has established the validity of the above assumptions,

demonstrating that the mutation rate is approximately constant within 50kb regions [214] and that mutations occur approximately independently given each position's trinucleotide context [169]. We demonstrated that the approach is more accurate than other methods on both real and synthetic data. We also demonstrated that multi-resolution prediction enables the identification of both known and novel putative drivers of cancer, including in the non-coding genome, a crucial open problem in genomics [143, 214].

In chapter 6 we will provide additional detail on the utility of the SPG to identify novel discoveries in cancer biology, identify the underlying predictors it uses for its inference, and compare it against well-validated driver detection techniques.

Chapter 6

Genome-wide mapping of somatic mutation rates uncovers drivers of cancer

6.1 Summary

Identification of cancer driver mutations that confer a proliferative advantage is central to understanding cancer; however, searches have often been limited to protein-coding sequences and specific noncoding elements (e.g., promoters) because of the challenge of modeling the highly variable somatic mutation rates observed across tumor genomes. In this chapter, we improve on the SPG to build Dig, a method to search for driver elements and mutations anywhere in the genome. We use deep neural networks to map cancer-specific mutation rates genome-wide at kilobase-scale resolution. These estimates are then refined to search for evidence of driver mutations under positive selection throughout the genome by comparing observed to expected mutation counts. We mapped mutation rates for 37 cancer types and applied these maps to identify putative drivers within intronic cryptic splice regions, 5' untranslated regions, and infrequently mutated genes. Our high-resolution mutation rate maps, available for web-based exploration, are a resource to enable driver discovery

genome-wide.

6.2 Introduction

Neutral (passenger) mutations that do not provide a proliferative advantage to a cell dominate the mutational landscape of tumors [244, 167]. Only a relatively small fraction of mutations are under positive selection [262, 169, 214] due to their ability to drive cancer by promoting cell growth, resisting cell death, or enabling tissue invasion [117]. Because positively selected mutations reoccur across tumors [170], genomic elements (e.g., coding sequence, promoters, enhancers, and lncRNAs) with carcinogenic potential accumulate more mutations than expected compared to the rates at which neutral mutations occur when counted across multiple tumors [196, 73]. Searching for mutational excesses attributable to positive selection to discover driver mutations, genes, and noncoding elements provides crucial insight into the mechanisms of cancer [169, 214, 71, 18, 51, 123, 93, 175].

Because robust identification of mutational excess requires an accurate model of the neutral mutation rate, computational tools that carefully model somatic mutation rates are central to locating additional cancer drivers. This task is made challenging by the highly variable and tissue-specific patterns of neutral mutations across the cancer genome [200, 245]. Existing methods address this challenge by fitting bespoke statistical models of mutation rates to specific regions of the genome [169, 73, 161, 236, 280, 151]. For example, methods designed to identify driver genes model mutation rates specifically within protein-coding sequences by using synonymous mutations as a proxy for neutral mutations [262, 169, 151, 278]. Recent methods designed to identify noncoding cancer drivers train sophisticated machine learning methods such as gradient boosting machines to model mutation rates within a subset of the genome [161, 236, 280] (4% of the genome in a recent pan-cancer analysis of noncoding drivers [214]). Additionally, some models search for driver mutations in unexpected nucleotide contexts [71], in unexpected clusters [247], or by directly (and interpretably) predicting the consequences of variants within the coding sequence of

select genes [185]. Despite this progress, the ability to search for evidence of driver mutations in arbitrary genomic regions remains incomplete: existing methods either are not applicable to most of the genome (e.g., because they operate only within coding sequence), require time-consuming and computationally expensive model training for each set of regions to test in a cancer cohort, or cannot test with base-pair resolution. These limitations contribute to catalogs of cancer driver elements remaining incomplete – particularly in the noncoding genome [276] – hindering precision oncology [169, 18, 99, 253].

In this chapter, we introduce a genome-wide neutral mutation rate model that allows rapid testing for evidence of positively-selected driver mutations anywhere in the genome. Banking on our conceptual progress presented in chapter 5, this approach, is predicated on two key methodological advances: first, we introduce a deep-learning approach to map cancer-specific somatic mutation rates at kilobase-scale resolution across the entire genome. Second, we propose a probabilistic model that uses these maps to test any set of candidate mutations from an arbitrary cancer cohort for evidence of positive selection. Through this framework, our maps enable millions of mutations to be evaluated in arbitrary cancer cohorts in minutes using the resources of a personal computer. We applied our deep-learning framework to map cancer-specific somatic mutation rates for 37 cancer types present in the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset [51], using high-resolution epigenetic assays from healthy tissues as predictive features (well-known correlates of tumor mutation rates at the megabase scale [200, 226]). We then used Dig to identify new coding and noncoding candidate cancer drivers in publicly available whole-genome, whole-exome, and targeted-sequencing cancer datasets. Our mutation maps are publicly available both as an interactive genome-browser and as a standalone software tool for quantifying excess somatic mutations anywhere in the genome in a dataset of interest.

6.3 Results

6.3.1 Testing mutational excess with probabilistic deep learning

To enable rapid evaluation of mutational excess anywhere in the genome, we designed Dig to model somatic mutation rates genome-wide for a given type of cancer. Thus, the distribution of neutral mutations over any set of genomic positions for a cohort of tumors from that cancer type can be looked-up nearly instantaneously. The method employs a probabilistic deep learning model that explicitly captures two central determinants of somatic mutation rate variability [200, 245, 151]: 1) kilobase-scale variation driven by epigenomic properties such as replication timing and chromatin accessibility that broadly impact the efficacy of DNA repair [73]; and 2) base-pair-scale variation driven by the sequence context biases of processes that induce somatic mutations such as APOBEC-driven cytidine deamination and UV light exposure [71, 245, 11, 10]. Kilobase-scale variation is modeled with a custom deep-learning architecture [270] that uses a neural network to predict cancer-specific mutation rates within 10kb regions and a Gaussian Process to quantify the prediction uncertainty, taking as input high-resolution epigenetic assays (and, optionally, flanking mutation counts) (fig. 6-1a, Extended Data fig. 6-5, Methods section 6.5). By strictly partitioning the genome into non-overlapping train, validation, and held-out test sets with five-fold cross-validation (predicting mutation rates in each one-fifth of the genome using a model trained and validated on observed mutations in the remaining four-fifths; Methods section 6.5), the network constructs a kilobase-scale map of the mutation rate genome-wide for a given type of cancer (**??**b). Base-pair variation is subsequently modeled using a generative graphical model that simulates how mutations should be distributed to individual positions in a region according to the nucleotide biases of mutational processes (Supplementary fig. D-1, Methods section 6.5). The marginal distribution over the number of neutral mutations at any set of positions has a closed form solution that depends only on the predicted regional

mutation rate, the prediction uncertainty, and the genome-wide probability that a position is mutated based on its neighboring nucleotides (Methods section 6.5). Thus, once values for these parameters are learned from a training cohort of a given cancer type, the distribution of mutations expected at any set of positions in the genome can be queried for any tumor cohort of the same cancer and used to test for evidence of positive selection by quantifying if excess mutations are observed (fig. 6-1c, Methods section 6.5).

We constructed mutation rate maps and inferred nucleotide mutation biases for 37 cancer types (Supplementary table D.1, table D.2) based on somatic mutations from the PCAWG dataset [51] and 100-bp patterns of 725 chromatin marks in 110 tissues from Roadmap Epigenomics [218], replication timing from 10 cell lines from ENCODE [64], and average nucleotide and GC content of the reference genome. We then benchmarked the accuracy of our somatic mutation rate models using the metric of proportion of variance explained, which we calculated as the square of the correlation coefficient between predicted and observed mutation counts as in previous works [200]. Dig successfully predicted a median of 77.3% (mean: 70.6%; range 22.7-92.3%) of variance in observed single nucleotide variant (SNV) rates in 10kb regions and a median of 94.6% (mean: 91.9%; range: 73.1-98.0%) of variance in 1Mb regions (fig. 6-1b, Supplementary table D.3) (Methods section 6.5) across 16 cancer types for which benchmarking power was sufficient (>1 million mutations and excluding lymphomas, in which activation-induced cytidine deaminase produces extreme outlier mutation counts in locally hypermutated regions). Compared to existing methods designed specifically to analyze tiled regions [191], coding sequence [169, 151], and noncoding elements in which synonymous mutations cannot be used to calibrate mutation rate models [161, 236] (e.g., enhancers and noncoding RNAs), Dig explained the most variation of SNV counts within 10kb regions in 14 of 16 cohorts, of nonsynonymous SNV counts in 16 of 16 cohorts, and of enhancer and noncoding RNA SNV counts in 15 of 16 cohorts, respectively (fig. 6-1d, table 6.1, Supplementary fig. D-2, Supplementary table D.3, table D.4, table D.5). Our approach's accuracy is attributable in part to the ability of the deep-learning network to identify local epigenetic structures

such as active transcription start sites and to associate these structures with mutation rates (Extended Data fig. 6-6, Supplementary Results appendix D.1.1).

This accuracy enabled correspondingly powerful driver identification: in benchmarks testing downstream ability to identify evidence of positive selection (i.e., excess of mutations) within previously-identified driver elements, Dig matched or exceeded the performance of methods tailored towards specific classes of elements [169, 161, 236, 280, 151] in whole-genome and whole-exome sequenced samples (fig. 6-1e, Supplementary fig. D-3, fig. D-4, fig. D-5, table D.6, table D.7, table D.8, table D.9, table D.10, table D.11, appendix D.1). Considering driver genes – for which high-quality databases of known driver genes that can approximate gold standard true-positives exist (Methods section 6.5) – Dig had the highest F1-score (a measure of accuracy) in 24 of 32 PCAWG cohorts (excluding skin and blood cancers as in previous works [236] due to local hypermutation processes) and the most power in 14 of 16 whole-exome cohorts compared to widely used, burden-based driver gene detection methods (fig. 6-1e, Supplementary fig. D-3, fig. D-4, Supplementary table D.6, table D.7) (power was measured as the area under approximated receiver-operator characteristic curves, which could be estimated due to the larger sizes of the exome sequenced cohorts; Methods section 6.5).

Identifying potential driver elements with Dig was 1-5 orders of magnitude faster than existing methods that train new models for every element and cohort analyzed (fig. 6-1f). For example, testing 10^7 observed mutations for evidence of positive selection within 10^5 noncoding elements with Dig completed in $<$90 seconds on a single CPU core compared to between $\sim$ 10 minutes and $>$2 days for other methods. Thus, our method matches or exceeds the power of existing approaches while requiring less runtime and providing flexibility to identify drivers with mutation-level precision genome-wide.

6.3.2 Small mutation sets increase power to identify drivers

Previous searches for noncoding driver elements have concluded that such drivers are likely rare, carried by $<$1% of samples [214]. A power analysis using our model's

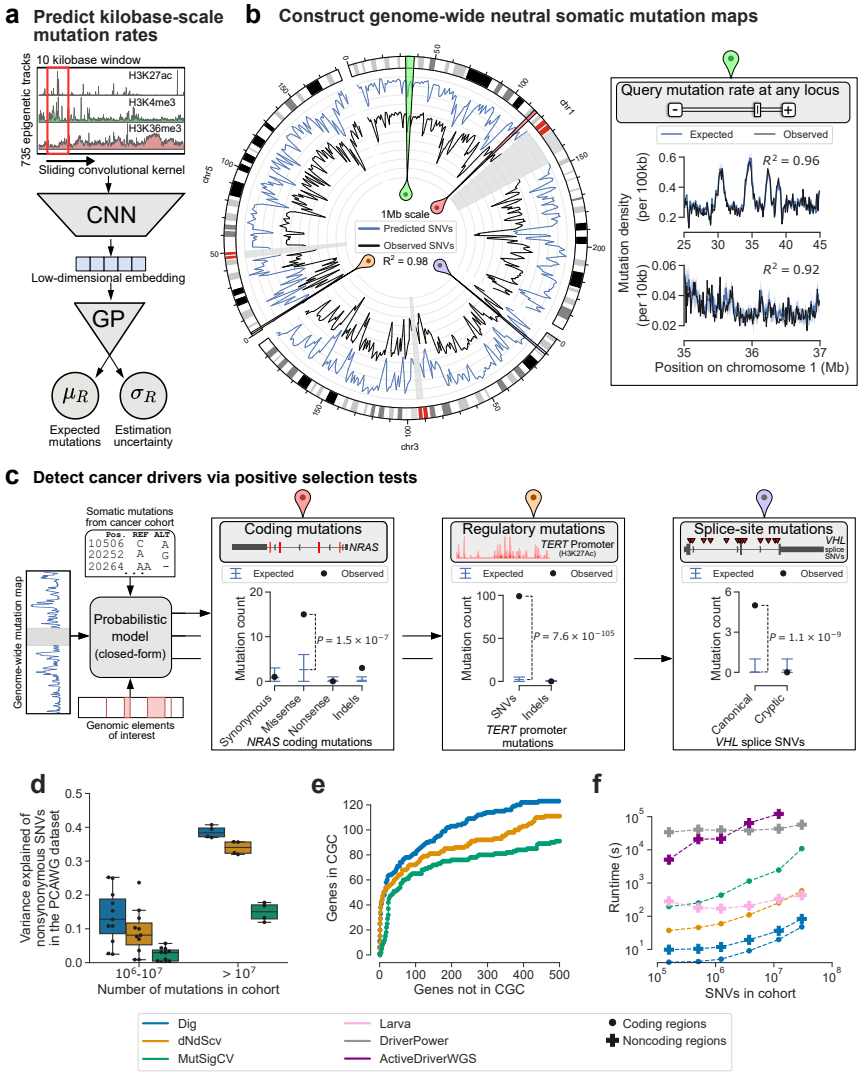

Figure 6-1: Modeling the genome-wide neutral somatic mutation rate and identifying cancer driver elements. **a,** Deep-learning scheme to predict expected number of somatic mutations and prediction uncertainty using epigenetic sequencing of healthy tissue. CNN: convolutional neural network; GP: Gaussian process. **b,** Genome-wide neutral somatic SNV map and observed density of SNVs in 1Mb windows from the PCAWG pan-cancer cohort (n=2,279 samples). Highlighted regions correspond to panels with the matching colored symbol. Inset: region on chromosome 1 modeled at 100kb and 10kb resolution. **c,** Examples of burden tests in the PCAWG pan-cancer dataset (N=2,279 samples) for coding mutations in *NRAS*, noncoding mutations in the *TERT* promoter and splice-site SNVs in *VHL*. Expected is mean with 95% confidence intervals. **d,** Proportion of variance of nonsynonymous SNV count in genes 1-1.5kb in length (n=3,740 genes) in 16 PCAWG cancer cohorts explained by different methods. **e,** Approximate numbers of false-positive and true-positive driver genes identified in the PCAWG pan-cancer cohort by method (across a range of calling thresholds). Numbers approximate because the true set of driver genes is unknown; CGC genes were used as a conservative approximation of true positives (a non-CGC gene may still be a true driver). **f,** Runtime of coding and noncoding driver detection methods. Comparison restricted to SNVs because not all methods support indels ActiveDriverWGS required >2 days to analyze the largest cohort.

	Percent of variance explained in observed SNV count (Pearson R^2 between observed and predicted SNV counts)		
Method	10kb regions	Nonsynonymous SNVs in coding sequence	Enhancers & noncoding RNAs
Dig (this work)	**92.30%**	**39.50%**	**49.00%**
NBR[34]	85.30%		
dNdScv[4]		35.70%	
MutSigCV[21]		17.80%	
Larva[18]			26.40%
DriverPower[19]			47.50%

Table 6.1: Proportion of variance in observed SNV counts in the PCAWG pan-cancer cohort (n=2,279 samples) explained by different methods. To minimize confounding from variation in element length (as longer elements are expected to have more mutations on average than shorter elements), the comparisons were restricted to genes with coding sequence 1-1.5 kb (n=3,740 genes) and to noncoding elements 0.5-1kb in length (n=7,412 elements). A shaded cell indicates that the method did not produce predictions over the associated annotation (NBR was able to analyze a subset of 6,024 enhancers and noncoding RNAs; it explained 1.8% of SNV count variation in those regions).

generative capabilities concurred (Methods section 6.5), indicating the most known noncoding elements (e.g., enhancers) require at least 1-2% of samples to carry driver mutations to have a >90% likelihood of detecting mutational excess at current sample sizes ($\sim 10^2$ for individual cancer types; $\sim 10^3$ for pan-cancer cohorts) (Supplementary fig. D-6. However, by reducing the size of tested elements to encompass only tens to hundreds of positions (as opposed to the thousands of bp spanned by most noncoding elements considered to date, e.g., average enhancer size: 1717 bp, range: 600-30,200 bp) power to identify driver mutations in <1% of samples increased by 20% (Supplementary fig. D-6). To demonstrate Dig's ability to find putative drivers, we thus defined and tested specific sets of mutations with potential functional impact for evidence of selection. The ability to test user-specified sets of specific mutations genome-wide is a unique feature (to our knowledge) of our method.

6.3.3 Quantifying pan-cancer selection on cryptic splice SNVs

Alternative-splicing is increasingly recognized as functionally relevant to cancer [195, 61] and recent studies have associated specific somatic mutations outside canonical splice sites with alternative splicing events observed in expression data [48, 52]. We thus applied Dig to rigorously quantify the extent to which cryptic splice SNVs, which may exist in both exons and introns of a gene (fig. 6-2a), occur in excess of the neutral mutation rate and therefore may function as driver mutations under selection. In tumor suppressor genes (TSGs) from the Cancer Gene Census (CGC) [249], cryptic splice SNVs as predicted by spliceAI [128] (Methods section 6.5) occurred significantly more often than expected under neutrality (648 SNVs observed in 283 TSGs vs. 550 SNVs expected, $P = 2.38 \times 10^{-5}$) (fig. 6-6b, Supplementary table D.12), were primarily enriched in introns (where the majority of such mutations occur), and were biased to occur in sites with high predicted impact on splicing (SNVs with predicted impact Δ score>0.8 exhibited a 1.75-fold enrichment (95% CI: 1.31-2.22 fold), $P = 2.52 \times 10^{-5}$) (fig. 6-22b,c). Overall, intronic cryptic splice SNVs were estimated to account for 4.5% (95% CI: 1.3-7.4%) of excess (potential driver) SNVs in TSGs, similar in magnitude to the 7.4% (5.6-9.7%) attributable to canonical splice SNVs,

whose driver potential is well established [169] (fig. 6-2d) (exonic excess SNV estimates were consistent with estimates from dNdScv; Supplementary fig. D-7). Results were robust to high mutation burden samples (Supplementary fig. D-8) and consistent with an analysis that did not rely on our mutation maps (Supplementary fig. D-9). Neither control genes not in the CGC nor oncogenes in the CGC were enriched for cryptic splice SNVs (Extended Data fig. 6-7, Supplementary table D.12). The lack of enrichment in oncogenes suggests that gain-of-function splice mutations beyond those that induce skipping of MET exon 14 are extremely rare, which may reflect the low likelihood of an intronic splice mutation resulting in the in-frame addition of residues that pathologically activate an oncogene. Conversely, the enrichment in TSGs suggests that cryptic splice mutations are generally inactivating, likely by triggering nonsense-mediated decay of mRNA transcripts or generating a protein with impaired function.

Considering individual genes, seven TSGs in 12 cancer types had a significant burden of intronic cryptic splice SNVs (FDR$<$0.1 for n=283 TSGs in 37 cancers) (Methods section 6.5) (fig. 6-2e, Supplementary table D.13), with patterns of TSG-cancer associations consistent with known tissue specificity of TSGs. Pan-cancer, *TP53* and *SMAD4* – both implicated in numerous cancers – carried an excess of cryptic splice SNVs. In contrast, the hematopoietic-specific TSG *CIITA* and the renal-specific TSG *PBRM1* carried excess cryptic splice SNVs in blood and kidney malignancies, respectively. In further support of these associations, the intronic cryptic splice SNVs observed in these TSGs, the majority (79.3%) of which fell outside annotated splice regions (i.e., $>$20bp from exon-intron boundaries) (fig. 6-2f), had significantly higher predicted impact on splicing than those observed in genes not in the CGC (fig. 6-2c) (mean SpliceAI Δ score=0.55 vs. 0.33; $P < 3 \times 10^{-4}$; Methods section 6.5). Moreover, of the six cryptic splice SNV carriers with available RNA-seq data of sufficient coverage, five had evidence of alternative splicing (fig. 6-2g, Supplementary fig. D-10, Supplementary table D.14, Supplementary Results appendix D.1) as quantified by LeafCutter [155] (Methods section 6.5). Overall, these results provide evidence that intronic cryptic splice SNVs are under positive selection in TSGs and likely act as

driver events in several percent of tumors across multiple cancer types.

Nine genes not in the CGC also had a significant burden of intronic cryptic splice SNVs in six cancers (Supplementary table D.15) at FDR<0.1, of which two genes had a significant burden at the more stringent Bonferroni ($\alpha < 0.05$) correction for 712,600 tests conducted across all genes and cancers. The burdens of four genes were driven by recurrent mutations at a single intronic location per gene (Supplementary table D.16). Implicated genes include *BTG2* in lymphoma, which is involved in the regulation of the G1/S transition of the cell cycle and has recently been implicated as a driver of blood cancers based on mutations in its coding sequence [71], and *ADAM19* in hemopoietic tumors, which has been implicated in the oncogenesis of breast [139], prostate [124], colorectal [275], and ovarian [55] cancers. While the computational prediction of new drivers should be interpreted with caution (Discussion section 6.4), these genes may be promising targets for future experimental studies to investigate their potential tumorigenic properties.

6.3.4 Noncoding candidate cancer driver mutations in 5' UTRs

Hypothesizing that indels could have large effect size on gene expression by disrupting transcription factor binding motifs, we searched promoters (n=19,251) for a burden of indels in the PCAWG pan-cancer dataset (Methods section 6.5). The *TP53* promoter was the only element with a genome-wide significant (FDR<0.1) burden of indels (7 observed vs. 0.54 expected; $P = 9.4 \times 10^{-7}$) (fig. 6-3a), consistent with a previous analysis that used restricted hypothesis testing to boost statistical power [214]. The observed mutations – all deletions significantly larger than expected (fig. 6-3b) (median length = 17bp vs 1bp expected; $P = 7.4 \times 10^{-4}$, one-sided Mann-Whitney U-test) – specifically affected exon 1 of the canonical 5' UTR, disrupted critical sequence elements (transcription start site, *WRAP53* binding sequence [164], internal ribosome entry site [272, 208], and the donor splice region of the multi-exonic 5' UTR) (fig. 6-3a), and exhibited enrichment comparable to cryptic exonic splice SNVs in *TP53*, which are well-characterized cancer drivers [246] (fig. 6-3c). More than half of the mutations (four of seven) within the exon 1 splice region did not alter the canonical

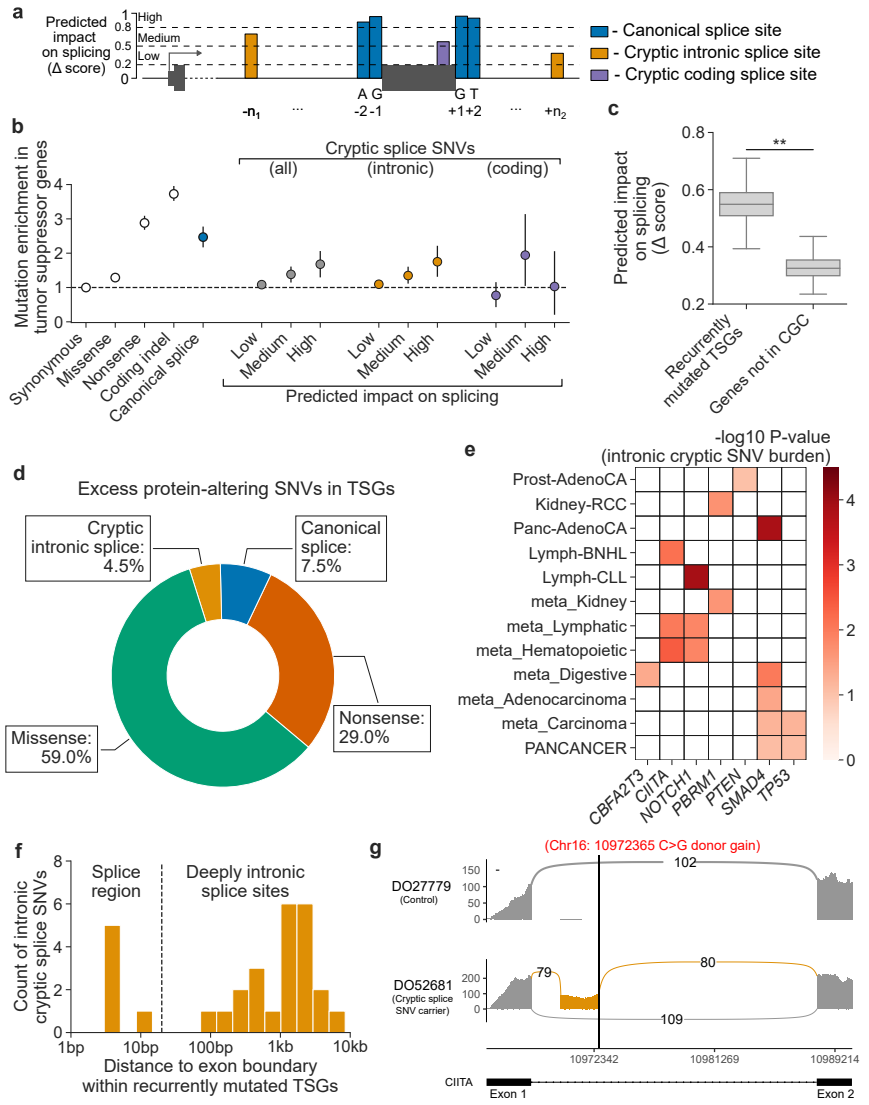

Figure 6-2: Evidence of positive selection on intronic cryptic splice SNVs in tumor suppressor genes. **a,** Schematic of the splice-altering SNVs considered in this analysis. Predicted impact on splicing measured by the SpliceAI Δ score (higher score approximates higher likelihood of altered splicing). **b,** Estimated enrichment (with 95% confidence interval) of observed mutations compared to expected neutral mutations in tumor suppressor genes stratified by variant type and predicted impact on splicing in N=2,279 pan-cancer samples from PCAWG dataset. **c,** Predicted splicing impact (SpliceAI Δ score) for intronic cryptic splice SNVs observed in recurrently mutated TSGs (see e) compared to those observed in genes not in the Cancer Gene Census (CGC) (** indicates bootstrapped $P < 3 \times 10^{-4}$, Methods section 6.5). **d,** Proportion of excess SNVs in TSGs contributed by each protein-altering SNV category. **e,** Known tumor suppressor genes per cancer with a significant burden (FDR<0.1) of predicted intronic cryptic splice SNVs. **f,** Distribution of distance to nearest exon boundary for the intronic cryptic splice SNVs observed in recurrently mutated TSGs. **g,** Pileup of RNA-seq reads in a Lymph-BNHL carrier of a predicted, deeply intronic cryptic splice SNV (labeled in red) in *CIITA* and a control Lymph-BNHL sample, showing the inclusion of a cryptic exon (gold) in the cryptic splice SNV carrier. Arc labels indicate the number of RNA-seq reads that support each exon junction.

splice sites, an unexpected pattern compared to other *TP53* splice regions (fig. 6-3d) ($P = 1.8 \times 10^{-3}$, two-sided Fisher's exact test). The 5' UTR mutation carriers had significantly lower expression of *TP53* than individuals without *TP53* mutations and individuals with predicted functional coding *TP53* mutations (1-2 standard deviation decreases in *TP53* expression compared to non-carriers, $P = 1.2 \times 10^{-4}$, Methods section 6.5) (fig. 6-3e; Supplementary fig. D-11), suggesting that these mutations either directly inhibit *TP53* transcription or result in nonsense mediated decay of the mRNA transcripts. Corroborating these results, seven of 2,399 distinct samples from the Hartwig Medical Foundation [204] showed a similar mutational pattern, with three carrying >10bp deletions and four carrying SNVs in *TP53* exon 1 and its donor splice region (fig. 6-3a).

These results motivated a targeted search for mutational burden in 5' UTRs and their splicing regions across 106 TSGs and 95 oncogenes with multi-exonic 5' UTRs (Methods section 6.5). One additional element, the 5' UTR of *ELF3*, had a significant burden of SNVs (fig. 6-3f) in PCAWG samples (6 observed SNVs vs 0.96 expected; $P = 2.9 \times 10^{-4}$); samples from the Hartwig Medical Foundation displayed a similar enrichment (10 observed vs. 1.5 expected; $P = 3.8 \times 10^{-4}$, Methods section 6.5). In both sets of samples, the enrichment was concentrated within the canonical *ELF3* 5' UTR; surrounding sequences (upstream promoter and intron 1) were not enriched for mutations (fig. 6-3f). The 16 mutations largely altered distinct base pairs within the 5' UTR – although two positions mutated in PCAWG samples were also mutated in the Hartwig samples – suggesting that this 5' UTR might be broadly sensitive to perturbation, possibly by prompting changes in promoter methylation that alter *ELF3* expression [75]. An alternative possibility could be an unmodeled local mutational process or technical artifact in this region [73]; however, a careful analysis did not find evidence for any such features that have explained other noncoding mutational hotspots [214] (Supplementary Results appendix D.1). The small number of carriers and limited availability of transcriptomic assays (only three carriers from PCAWG had RNA-seq data) prevented investigation into the possible function of these 5' UTR mutations. Thus, additional follow-up – particularly experimental assays assessing

the impact of 5'UTR mutations [267] – will be necessary to determine whether the mutational enrichment here represents positive selection or represents a new neutral mutational process.

6.3.5 The shared landscape of common and rare driver genes

Small sample sizes have limited assessment of whether rare coding mutations (which account for most exonic mutations in tumors) act as drivers even in well-characterized driver genes. We increased statistical power in two ways: 1) by analyzing large meta-cohorts of nonsynonymous SNVs from 14,018 whole-exome and targeted-sequencing samples representing ten solid tumor types (median samples per cancer: 1,195; range: 515-3,110) (Supplementary table D.19) (Methods section 6.5); and 2) by considering only activating mutations in oncogenes (obtained from the Cancer Genome Interpreter [247]) and predicted loss-of-function (pLoF) mutations in all other genes. Such analysis has previously been impeded by the exclusion of synonymous mutations from large, publicly available targeted sequencing datasets [54, 273, 209, 217, 131] because existing driver gene detection methods are reliant on synonymous mutations. Dig circumvents this difficulty because model parameters have already been inferred from a separate training cohort.

For each cancer, we first restricted our analysis to "long-tail" genes, which we defined as oncogenes and TSGs not associated with that cancer type in any of three recent, large pan-cancer surveys of driver genes [18, 71, 170]. Dig estimated between 1% and 5% of samples (depending on the cancer) carried activating SNVs in long-tail oncogenes (fig. 6-4a) and 3% to 6.5% carried pLoF SNVs in long-tail TSGs (fig. 6-4b). These rates were significantly higher than expected ($P < 3.78 \times 10^{-9}$ for activating SNVs in all cohorts; $P < 3.10 \times 10^{-4}$ for pLoF SNVs in all cohorts except prostate, $P = 0.056$ for prostate) (Supplementary fig. D-12, Supplementary table D.20, table D.21) (Methods section 6.5). These rates were consistent when we restricted the analysis to only whole-exome sequenced samples, though power to detect positive selection was decreased due to reduced sample size (Supplementary fig. D-13). Considering individual genes, 92 oncogene-tumor pairs not reported in recent pan-cancer surveys

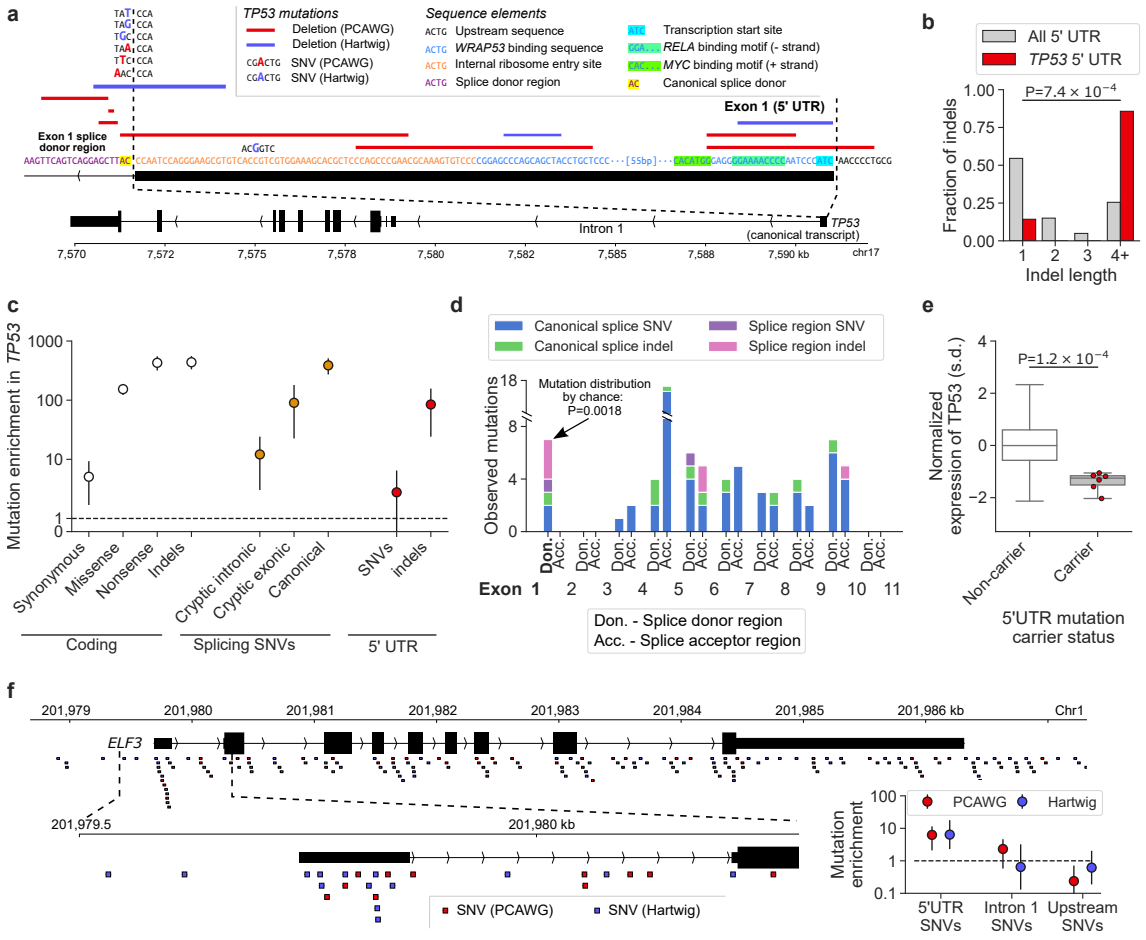

Figure 6-3: Enrichment of somatic mutations in the 5' UTRs of *TP53* and *ELF3*. **a,** Mutations from PCAWG and Hartwig Medical Foundation cohorts observed within exon 1 of the 5' UTR of the canonical *TP53* transcript. DNA sequence from GRCh37 reference genome (+ strand). Mutation types, relevant sequence and regulatory elements as indicated in the legend. **b-e,** Analysis on PCAWG pan-cancer dataset (N=2,279 samples). **b,** Distribution of indel sizes observed within 5' UTRs of genes other than *TP53* (n=3988 indels) and within the *TP53* 5' UTR (n=7 indels). P-value comparing median indel lengths from one-sided Mann-Whitney U-test. **c,** Estimated mutation enrichment relative to the neutral mutation rate (observed / expected neutral mutations) within *TP53* stratified by mutation type and location. Error bars, 95% CI. **d,** Distribution of mutations observed within donor and acceptor splice regions (defined as the 20bp 3' and 5' of an exon, respectively) of the canonical *TP53* transcript. Canonical splice SNVs and indels: mutations altering the two base-pairs immediately adjacent to an exon boundary; splice region SNVs and indels: mutations intersecting the splice region but not the canonical splice sites. The donor splice region of exon 1 of the 5' UTR (shown in a) is bolded. **e,** Expression of *TP53* on standard deviation (s.d.) scale in carriers of *TP53* 5' UTR mutations (n=6) and non-carriers (n=1,205), adjusted for tumor type and copy number in the PCAWG pan-cancer dataset (N=2,279 samples). **f,** SNVs overlapping *ELF3* in the PCAWG and Hartwig Medical Foundation cohorts. Insets: zoom-in of the *ELF3* 5' UTR region and estimated mutational enrichments with 95% CIs within this region.

of driver genes had a significant (FDR<0.1) burden of activating SNVs (fig. 6-4c). 46 TSG-tumor pairs not reported in the pan-cancer surveys had a significant burden of pLoF mutations (fig. 6-4d). The newly identified candidate driver genes were rare compared to driver genes in existing databases (0.28% (interquartile range: 0.14-0.53%) vs 1.3% (interquartile range: 0.59%-3.0%) for newly implicated and known driver genes, respectively, $P = 3.1 \times 10^{-27}$ two-sided Mann-Whitney U-test). Further supporting these predictions, the distribution of activating mutations in a given driver gene was similar in cancers in which the gene is a known, common driver and in cancers in which we newly implicated the gene as a putative rare driver (Extended Data fig. 6-8). For example, the G12, G13, Q61, and A146 positions of *KRAS* accounted for the majority of *KRAS* SNVs in both common and rare scenarios (lung non-small cell tumors: 568/586 mutations; prostate tumors: 12/17 mutations; gliomas: 11/15), and the V600E mutation accounted for the plurality of *BRAF* SNVs in common and rare scenarios despite each gene having dozens of known activating SNVs (52 and 71, respectively). Additionally, carriers of mutations in several predicted rare driver genes exhibited phenotypes consistent with those reported in tumors in which the genes are common drivers (Supplementary Results appendix D.1). For example, CNS tumors with rare pLoF mutations in the DNA mismatch repair genes *MSH2* and *MLH1* exhibited significantly increased global mutation rates across 213 targeted sequenced genes (*MSH2*: mean 30.1 mutations in carriers vs. 3.0 in non-carriers, $P = 3.8 \times 10^{-7}$ one-sided Mann-Whitney U-test; *MLH1*: mean 35.3 mutations in carriers vs. 3.1 in non-carriers, $P = 8.8 \times 10^{-6}$ one-sided Mann-Whitney U-test).

A further 29 gene-tumor pairs had a significant (FDR<0.1) burden of pLoF mutations in genes not in the cancer driver databases for any cancer (Methods section 6.5) (Supplementary table D.22), of which two were significant at the more stringent Bonferroni ($\alpha < 0.05$) correction for the total number of genes tested and six were additionally supported by a nominal ($P < 0.05$) burden of missense mutations. The top hit is the cell polarity gene $PARD3$ in gastroesophageal cancer (9 observed pLoF SNVs vs. 1.1 expected, $P = 1.57 \times 10^{-6}$), which, despite not being in major driver gene databases, is a known fusion partner of the oncogene RET and has been impli-

cated in the tumorigenesis of multiple solid cancers [13]. The ability to distinguish mutational burdens in genes with a low frequency of mutations such as *PARD3* (9 carriers in 827 samples) highlights the increased statistical power our approach can achieve by testing specific sets of mutations in large cohorts for evidence of positive selection.

Our results represent progress toward an unbiased, pan-cancer catalog of driver genes and suggest driver mechanisms are shared across the common and rare driver landscape of solid cancers. However, computational identification of rare driver genes at current sample sizes relies upon small mutation counts, and predictions should be interpreted with care. Experimental characterization of genes' functions in the relevant cancers is essential to confirming their carcinogenic roles.

6.4 Discussion

Dig is a probabilistic deep-learning method that enables rapid tests for evidence of positive selection on genomic elements that can be defined with the precision of individual mutations anywhere in the genome. The strong performance of the method in modeling mutation rates and identifying candidate drivers highlights the power of deep-learning to capture complex cellular processes with data derived from high-throughput sequencing [128, 76, 96, 15, 14, 163]. Specifically, building upon the observation that epigenetics correlate with somatic mutation rates [245], we showed that neural networks applied to a corpus of high-resolution ChIP-seq assays are able to learn nuanced, non-linear associations between local epigenetic structures and patterns of somatic mutations. Moreover, techniques presented here are adaptable to other contexts. For example, quantification of prediction uncertainty by coupling a Gaussian process to the final layer of a neural network may be a practical solution to improve the reliability and interpretability of predictions in other deep-learning settings [121].

The application of our high-resolution mutation rate maps to quantify mutational burdens genome-wide provides a glimpse into the landscapes of rare and noncoding

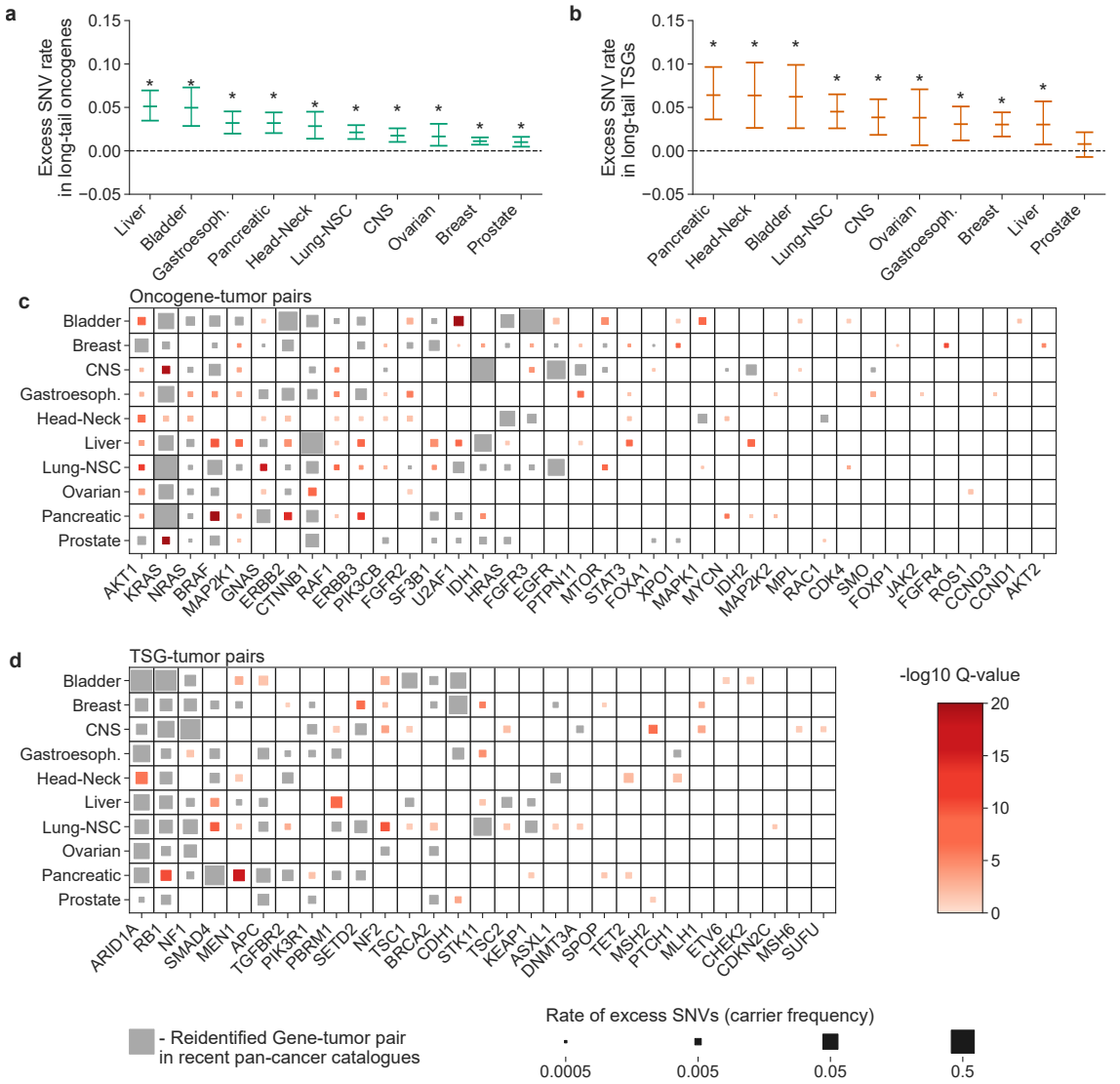

Figure 6-4: Enrichment of protein-altering SNVs in "long-tail" genes reveal a shared landscape of common and rare driver genes. a,b, Estimated mutation rates with 95% CIs of excess oncogenic SNVs in oncogenes (**a**) and predicted loss-of-function (pLoF) variants in TSGs (**b**) that were not previously associated with a given cancer (x-axis) in three large driver gene catalogues [71, 170, 18]. Stars indicate that the burden of oncogenic (pLoF) SNVs was significant in long-tail oncogenes (TSGs) in the cancer type (P-values and number of SNVs per category in Supplementary table D.20, table D.21). **c,d,** Oncogene-tumor pairs and TSG-tumor pairs with a significant burden of oncogenic or protein-truncating SNVs. Gene-tumor pairs previously reported by Dietlein et al. [71], Bailey et al. [18], or Martínez-Jiménez et al. [170] are marked in grey. Pairs that are not present in those catalogues are marked in red with color intensity indicating significance of association. Marker size is proportional to the estimated rate of excess mutations after accounting for cancer-specific neutral mutation rates.

driver mutations that we anticipate will emerge as cancer sequence sample sizes continue to grow. While the driver candidates we report – in cryptic splice sites, 5' UTRs, and rarely mutated genes – occurred at low frequencies individually, our estimates suggest that they collectively contribute to the disease pathology of up to 10% of tumors (summing across the percent of tumors predicted to carry excess mutations in each of these elements). This estimate may be conservative, as several analyses utilized datasets of mutations that are unlikely to be comprehensive (e.g., catalogs of predicted cryptic splice SNVs and known activating SNVs). The quantification of these rare driver events is important in part because it suggests avenues to expand patient treatment options by repurposing therapeutics; a targeted therapy approved for a mutation in one cancer type may prove beneficial to patients with that mutation in other cancer types. Indeed, cancer-agnostic approaches to patient stratification are currently being deployed at some cancer centers [235].

Additionally, current sample sizes are not adequate to uncover infrequent drivers under moderate or weak positive selection. We anticipate that Dig will be particularly useful in uncovering such mutations due to its ability to rapidly evaluate mutations spread over large swaths of the genome. For instance, a preliminary analysis we performed of enhancer networks identified several genes with a burden of enhancer mutations (Supplementary table D.23, Supplementary Results appendix D.1), including *FOXA1*, in which promoter mutations are thought to drive breast cancer by increasing gene expression [215]. A possible approach to increase sample size with existing data is to call somatic mutations in regions flanking coding sequence using off-target reads from large targeted or whole-exome sequenced clinical cohorts.

However, computational prediction alone is not sufficient to establish the causal role of an element or mutation in cancer pathology because an excess of mutations compared to the neutral mutation rate does not definitively prove positive selection. Moreover, recent studies have shown that canonical cancer driver mutations can be present in seemingly healthy tissues [168, 152, 156, 182, 203], adding an additional layer of complexity to interpreting whether or how a mutation causally contributes to a malignant phenotype. Ultimately, experimental validation is necessary to establish

the causal role for a mutation as a driver of cancer. Dig provides a tool for in silico guidance of in vitro and in vivo studies because it enables the prioritization of precise sets of mutations that may act as drivers in both the coding and noncoding genome. These specific sets of mutations can then be evaluated in experimental systems. For example, the predicted cryptic splice mutations that Dig identified as putative drivers could be evaluated as possible drug targets by CRISPR base-editing of cell lines followed by drug screening assays [101]. Thus, we anticipate that deep-learning generally and our tool specifically can improve computational, experimental, and clinical utility of the growing body of cancer genome sequencing data. Similarly, the trend of interpretable models can accelerate novel discoveries in a wide variety of life-science sub-domains.

6.5 Materials and Methods

6.5.1 Sequencing data curation

PCAWG dataset

We obtained somatic SNVs and indels from whole-genome sequencing of 2,583 unique tumors from the ICGC data portal (`https://dcc.icgc.org/`) and dbGaP (project code: phs000178) that previously passed quality control [214]. The somatic mutation calls in this dataset have previously been stringently filtered to remove possible germline calls, false-positive calls due to oxidative DNA damage, and calls with high strand bias [51]. Following procedures described in Rheinbay et al., we grouped samples into 38 individual cancer types and 14 meta-cohorts that combined similar tumor types, including a pan-cancer cohort that included all samples except melanoma and lymphoma tumors (consistent with Rheinbay et al.). We removed samples with reported high microsatellite instability from all cohorts except the pan-cancer cohort and annotated autosomal coding SNVs and indels with their predicted functional impact using a custom annotation method. (We excluded sex chromosomes because the number of observed mutations on the X chromosome depends on the sex composition

of a cohort). For the creation of somatic mutation maps and driver element analysis, we considered cohorts with at least 20 samples and >105 SNVs (Supplementary table D.1). This resulted in a set of 23 individual cancer types and 14 meta-cohorts.

Dietlein et al. dataset

We obtained somatic SNVs and indels from whole-exome sequencing of 11,873 tumors from 28 cancer types that had previously been curated in Dietlein et al. [71] from `http://www.cancer-genes.org/`; the dataset had previously undergone filtering to remove germline calls and due to oxidative DNA damage as described in in Dietlein et al. [71]. We restricted to a set of 8,617 tumor samples from 17 cancer types for which we had a mutation rate model trained on the PCAWG dataset (Supplementary table D.24). We additionally constructed a pan-cancer dataset by merging somatic mutations from all samples excluding melanoma and hematopoietic malignancies as in PCAWG [214]. Coding mutations were annotated for their predicted functional impact as above.

Target sequencing datasets

We obtained somatic SNVs from targeted sequencing of 10 types of solid cancers performed using the IMPACT protocol at the Memorial Sloan Kettering Cancer Institute from cbioportal [54] (`https://www.cbioportal.org/`) (Supplementary table D.19). Possible germline calls had been previously excluded from these datasets. We removed duplicate patients and hypermutated samples with >100 coding mutations in 221 genes common to all whole-exome and targeted sequenced samples (removal of hypermutated samples is common in driver gene detection and has been shown to improve accuracy [169]). Coding SNVs were then annotated for their predicted functional impact in coding sequence as above and merged with SNVs from the whole-exome datasets (after removing hypermutated samples) of the corresponding cancer type to form mega-cohorts with aggregate sample size of 14,018 tumors in 10 cancer types.

Additional filtering of germline mutations

Any mutation occurring in an element with a nominally FDR<0.1 significant burden of mutations was cross-referenced with the gnomAD database v.2.1.1 [140] and excluded if it occurred in gnomAD with an allele count of five or more in any population, unless the mutation occurred primarily in a single population and the carrier was not of that population (this occurred only once; the mutation 1:43804317-C>T was observed in a carrier of European ancestry, but is reported in gnomAD as occurring in Latino/admixed American populations). If the mutational burden of the element did not remain FDR<0.1 significant after exclusion of these possible germline mutations, it was removed from further analysis. This filter was applied to all datasets.

6.5.2 Identification of mutational excess with probabilistic deep learning

Dig consists of two components: 1) a deep-learning module that models approximately constant somatic mutation rates within kilobase-scale regions (e.g., 10-50kb) due to epigenetic features (e.g., chromatin compactness) that vary at this scale5; and 2) a generative probabilistic model that captures the likelihood that a given position is mutated in a cancer cohort, conditioned on its sequence context [71, 11, 10, 191] and the kilobase-scale mutation rate of that cancer type. Intuitively, the kilobase-scale model provides information about how many neutral mutations should be present in a region while the nucleotide context model determines how those mutations should be distributed amongst individual positions.

Modeling kilobase-scale mutation rates with deep-learning

Model architecture The purpose of the deep-learning model is to 1) predict the mutation rate μ_R and 2) quantify prediction uncertainty σ_R^2 conditioned on the epigenetic organization of the region R. The architecture has been previously described [270]. Briefly, the network consists of a convolutional neural network (CNN) that takes as input a high-dimensional matrix of epigenetic assays (see Model input and

output) and projects the matrix into a 16-dimensional vector. Optionally, the CNN also embeds into the 16-dimensional vector the mutation counts observed in the 100kb regions flanking the region of interest. The low-dimensional embedding is then provided as input to a Gaussian process (GP) that predicts the mean and variance of number of mutations in the region. Technical details are provided in Supplementary Methods appendix D.2.

Model input and output The CNN and GP were trained sequentially to predict somatic SNV counts in nonoverlapping 10kb regions by minimizing squared error loss between predicted values and observed counts from the PCAWG dataset for each of 37 cancer types. The network received as input matrices of size 735×100 where each row is an epigenetic feature track and each column is the average track value in nonoverlapping 100bp windows. 723 rows were uniformly processed -log10 P-values for peaks of chromatin markers from 127 tissues [218], 10 rows were replication timings of 10 cell lines from ENCODE [64], and two were the average nucleotide content and average GC content of the human reference genome. The network additionally received as input somatic SNV counts in 100kb regions flanking each 10kb of interest from the relevant cancer in the PCAWG dataset. However, the accuracy of the method over 1Mb regions was benchmarked using networks trained without flanking region counts to avoid any leakage of information between train and test sets.

Model training For each cancer, predictions in each nonoverlapping 10kb region R of the autosome was obtained via the following five-fold cross-validation strategy: bins that passed quality control (Supplementary Methods appendix D.2) were randomly divided into five equal size folds, each containing 20% of the bins. Sequentially, each fold was withheld and a deep-learning model was trained using 80% of the remaining bins and validated over the other 20% of the remaining bins to avoid overfitting (Supplementary Methods appendix D.2). Prediction was then performed over the held-out fold (20% of the genome) and over regions filtered by quality checks. Additional technical details of model training are described in Supplementary Meth-

ods appendix D.2.

Testing mutational burden with a graphical model

Genome-wide likelihood of mutation from sequence context For each cancer, maximum likelihood estimation was used to estimate the genome-wide probability of a mutation in each of 192 possible trinucleotide contexts using SNV counts from the PCAWG dataset. The statistical procedure is described in Supplementary Methods appendix D.2.

Modeling mutation counts over an arbitrary set of positions We conceptualized that mutations arise in a region R with an unknown rate whose possible values are drawn from a distribution defined by the mean and variance predicted by the deep-learning network. As mutations arise they are distributed to individual positions based on the probability that each position in R is mutated based on its sequence context. Let $M_{i,\ aX\to Yb}$ be the number of SNVs of the form $aX \to Yb$ at position i in region R in some cancer cohort of interest. Then under a probabilistic graphical model described in Supplementary Methods appendix D.2, the marginal distribution over a set of possible SNVs in a region is31:

$$\sum_I M_{i,aX\to Yb} \sim \text{NegativeBinomial}\left(\alpha_R, \frac{1}{1 + C_{\text{SNV}} \cdot \theta_R \cdot \sum_I p_{R,aX\to Yb}}\right)$$

where $\alpha_R = \mu_R^2/\sigma_R^2$ and $\theta_R = \sigma_R^2/\mu_R$ (recall μ_R and σ_R^2 are the mean and variance of mutation rate in region R estimated by the deep-learning model); $p_{R,aX\to Yb}$ is the genome-wide probability of a mutation of the form $aX \to Yb$, normalized such that the probability of all possible mutations in R sums to one; and C_{SNV} is a constant scaling factor that accounts for the difference in sample size between the cohort of interest and the training cohort. All parameters in the distribution except C_{SNV} are already estimated from the training cohort. By default, C_{SNV} is calculated as the ratio of the number of observed synonymous SNVs in the target dataset to the number of expected synonymous SNVs in the training cohort across all genes excluding TP53

(in which some synonymous mutations are under positive selection [169]). Thus, once the model has been trained once on the training cohort, calculating the distribution over any set of mutations in a target cohort of interest is essentially reduced to the constant time look-up of parameters. More details on the graphical model including its extension to indels, multiallelic variants, and sets of variants that span multiple regions are described in Supplementary Methods appendix D.2.

6.5.3 Comparison to existing driver detection methods

We compared Dig's performance to that of six existing methods (NBR [191], dNdScv [169], MutSigCV [151], Larva [161], DriverPower [236], and ActiveDriverWGS [280]) over two benchmarks: accuracy of the background mutation rate models and accuracy of driver detection. The six comparison methods were chosen because they are state-of-the-art methods that 1) identify putative driver candidates by searching for mutational excess and 2) are designed to model diverse regions of the genome: tiled regions (NBR), coding sequence (dNdScv and MutSigCV), and noncoding elements such as enhancers (Larva, ActiveDriverWGS, and DriverPower). All methods were run with default parameters.

Comparing background mutation rate models

We compared the variance explained of observed SNV counts between models. Variance explained is the proportion to which a mathematical model accounts for variation in a dataset, which we calculated as the square of the Pearson correlation coefficient between predicted and observed SNV counts as in previous works [200]. To ensure sufficient benchmarking power, we restricted comparisons to 16 cancer types in the PCAWG dataset with >1 million mutations because the variance explained statistic becomes deflated when observed counts are low in a discrete system (Supplementary Methods appendix D.2). Comparisons were performed over nonoverlapping 10kb regions of the genome (Dig vs. NBR), nonsynonymous SNVs in coding sequence (Dig vs. dNdScv vs. MutsigCV), and the noncoding elements enhancers and long & short

noncoding RNAs (Dig vs. Larva vs. DriverPower; ActiveDriverWGS was not included because it does not output its internal estimates of mutation counts). We chose enhancers and noncoding RNAs because they are noncoding elements that all three methods could analyze and are sufficiently far from coding sequence that synonymous mutations cannot be used in general to estimate the neutral mutation rate. To control for confounding from element length (longer elements have more mutations on average than shorter elements), we restricted the analysis to genes 1-1.5kb in length (N=3,740) and noncoding elements 0.5-1kb in length (N=7,412). Additional details of region selection are described in Supplementary Methods appendix D.2.

Comparing driver element identification accuracy

Coding models We compared the sensitivity, specificity, and F1-score (harmonic mean of sensitivity and specificity) for driver gene detection from coding sequence mutations between Dig, MutSigCV, and dNdScv across the 32 PCAWG cancer cohorts (melanomas and hematopoietic cancers were excluded as in previous comparisons [236]). We additionally compared power over the 16 whole-exome sequenced cohorts from Dietlien et al. (excluding hematopoietic cancers as above). Details of both comparisons are provided in Supplementary Methods appendix D.2.

Noncoding models We compared the sensitivity, specificity, and F1-score for driver noncoding element identification from noncoding SNVs between Dig, DriverPower, Larva, and ActiveDriverWGS across the 32 PCAWG cancer cohorts (excluding melanoma and hematopoietic cancers as above). We chose to compare to these three methods because they are recently introduced methods for noncoding driver element identification that rely on neutral mutation models to test for selection. Details are provided in Supplementary Methods appendix D.2.

6.5.4 Power analysis

We conservatively simulated Dig's power to detect driver SNVs at different carrier frequencies across enhancers and noncoding cryptic splice sites under the pan-cancer

mutation map using a Monte Carlo approach described in Supplementary Methods appendix D.2.

6.5.5 Quantifying selection on cryptic splice SNVs

Curation of predicted splice SNVs

From SpliceAI [128], we obtained a list of every possible SNV in the body of 17,816 autosomal genes with predicted impact on splicing (i.e., SpliceAI Δ score) >0.2. Predicted splice-altering SNVs were separated into canonical (altering positions 1 or 2 base-pairs 5' or 3' to an exon boundary) from cryptic splice SNVs (all other SNVs excluding sites that were 5 base-pairs 3' to an exon boundary that had been included in the definition of "essential splice sites" considered by Martincorena et al. [169] – excluded to ensure any enrichment we observed was independent of enrichment reported in that work). SNV positions were assigned based on the Gencode V24 list of basic transcripts. Cryptic splice SNVs were further divided into coding SNVs (defined as synonymous SNVs common to each transcript of a gene) and intronic (defined as SNVs not falling within any coding sequence of any transcript).

Enrichment of coding mutations and splice SNVs in PCAWG

Dig was applied with default settings to the following sets of mutation from the PCAWG pan-cancer cohort in each of 17,815 genes for which we had predicted splice SNVs: synonymous SNVs, missense SNVs, nonsense (stop-gained) SNVs, coding indels, canonical splice SNVs, and cryptic splice SNVs. Mutation enrichment was defined as the ratio of the observed mutations to expected mutations (this statistic is conceptually similar to the selection coefficient reported for coding mutations by dNdScv). P-values for a gene set and mutation type were exactly calculated by convolving the mutation-type specific negative binomial distributions for each gene in the gene set and summing the upper-tail probability that at least the number of observed mutations occurred by chance. We used a Monte Carlo simulation approach to estimate the 95% confidence intervals of enrichment within a set of genes and given

mutation type (Supplementary Methods appendix D.2). To further assess mutational enrichment, we directly compared the rate of mutations in TSGs and oncogenes to the rate in genes not in the CGC (Supplementary Methods appendix D.2). The excess of SNVs in TSGs in the CGC stratified by function (missense, nonsense, canonical splice, and noncoding canonical splice) was calculated as the difference between the number of mutations observed and the number expected. The relative contribution for each functional category was defined as the excess for that category normalized by the sum of the excess across all categories. The 95% confidence interval for the contribution of each category was calculated using a Monte Carlo approach (Supplementary Methods appendix D.2).

Genes enriched for noncanonical cryptic splice SNVs

In each of the 37 PCAWG cohorts, we identified genes with a significant burden of noncanonical cryptic splice SNVs as quantified by Dig. We considered two sets of genes: 1) all TSGs in the CGC (n=283) and 2) all autosomal genes with predicted splice SNVs (n=17,815). The significance threshold was defined per cancer as FDR q-value$<$0.1 corrected for the number of tests (n=283 or n=17,815). We excluded genes where multiple SNVs contributing to the burden were observed in a single sample. We used a bootstrap method to determine whether predicted cryptic splice SNVs observed in TSGs with a significant burden were enriched for high predicted impact on splicing (Supplementary Methods appendix D.2).

Analysis of alternative splicing events in RNA-seq data

We obtained RNA-seq data for 8 samples carrying deep intronic predicted cryptic splice SNVs (i.e., distance to nearest exon boundary $>$20 base-pairs) in TSGs with a significant burden of predicted noncoding cryptic splice SNVs and 41 control samples without a cryptic splice SNV. For each carrier-control pair of the same cancer type, we performed differential splicing analysis using LeafCutter as described by Li et al. [155]. Further details of the analysis are provided in Supplementary Methods appendix D.2.

6.5.6 Quantifying mutational excess in promoters and 5' UTRs

Discovery of elements with a burden of mutations

Dig with default parameters was used to evaluate the PCAWG pan-cancer cohort (excluding hypermutated samples with >3000 coding mutations) for mutational excess within two sets of regions: 1) indel excess within promoters previously defined by the PCAWG consortium [214] (n=19,251) and 2) SNV and indel excess within 5' UTRs of TSGS (n=106) and oncogenes (n=95) in the CGC that spanned multiple exons of the canonical transcripts of genes (as defined by UCSC genome browser for GRCh37); we additionally included the splice regions of the 5' UTRs in our analysis, defined as the 20 base-pairs bordering the start or end of an exon. The significance threshold was defined per cancer as FDR q-value<0.1 corrected for the number of tests (n=19,251 or n=201).

ELF3 5' UTR mutations in the HMF cohort

We downloaded somatic mutations observed in the Hartwig Medical Foundation metastasis cohort [204] from their online data portal (`https://database.hartwigmedicalfoundation.nl/`), excluding skin and hematopoietic tumors. Since we could only download mutations specific to a gene, we did not quantify burden with Dig. Rather, we directly compared the rate of SNVs in the 5' UTR, first intron, and 1kb upstream region of ELF3 to the rate of synonymous mutations in *ELF3* using a two-sided Fisher's exact test.

Analysis of expression levels

We obtained gene expression levels (FPKM) and gene-level copy number estimates from the PCAWG data portal for all tumors for which RNA sequencing was performed. For a gene of interest, we applied a fixed-effects linear regression model to residualize the expression values for gene-level copy number per sample and the interaction between gene-level copy number and the cancer project that originally generated the RNA-seq data. We then normalized the residual expression values to

have mean zero and unit variance across all samples and compared the normalized values between mutation carriers and noncarriers using a two-sided Mann-Whitney U-test.

6.5.7 Driver gene prediction in WES & targeted sequenced samples

Mutational excess in "Long-tail" driver genes

For each of the 10 cancer types for which we compiled SNVs from whole-exome and targeted sequenced cohorts, we assembled a list of known driver genes identified in any of three recent, pan-cancer driver gene discovery efforts [170, 71, 18] (we required genes be discovered with FDR$<$0.1, the significance threshold common across the driver element detection literature) that were also common to all whole-exome and targeted sequenced samples (n=69 oncogenes and n=56 TSGs). For a given cancer, we considered "long-tail" genes to be driver genes that were not on the list of known driver genes for the given cancer (that is, they were driver genes associated with other cancers). Dig was then used to quantify mutational excess in those long-tail genes. Because synonymous mutations were not available from the targeted sequenced samples, we instead used missense mutations with CADD phred score$<$15 to estimate the scaling factor that adapted the somatic mutation maps trained on PCAWG cohort to the meta-cohorts (details in Supplementary Methods appendix D.2). We directly estimated the P-value of the mutational burden long-tail genes by convolving the neutral mutation distributions for each individual gene and calculating the upper-tail probability of at least the number of observed mutations across all genes occurring by chance under the null distribution. We calculated 95% confidence intervals of excess mutations using the same Monte Carlo approach as in our analysis of cryptic splice SNVs. Excess rate per sample was calculated as the number of excess SNVs divided by the number of samples in the cohort for a given cancer type.

Identification of putative driver genes

We used Dig to identify individual genes with an excess of mutations in two cases: 1) in our meta-cohorts, testing 69 oncogenes for an excess of activating SNVs and 56 TSGs for an excess of pLoF SNVs (these were the set of known driver genes common to all whole-exome and targeted sequenced cohorts); and 2) in the exome-sequenced cohorts alone, testing 19,210 autosomal genes for an excess of pLoF SNVs. In each case, significance was defined as FDR q-value$<$0.1 for the number of genes tested.

6.6 Extended Data Figures

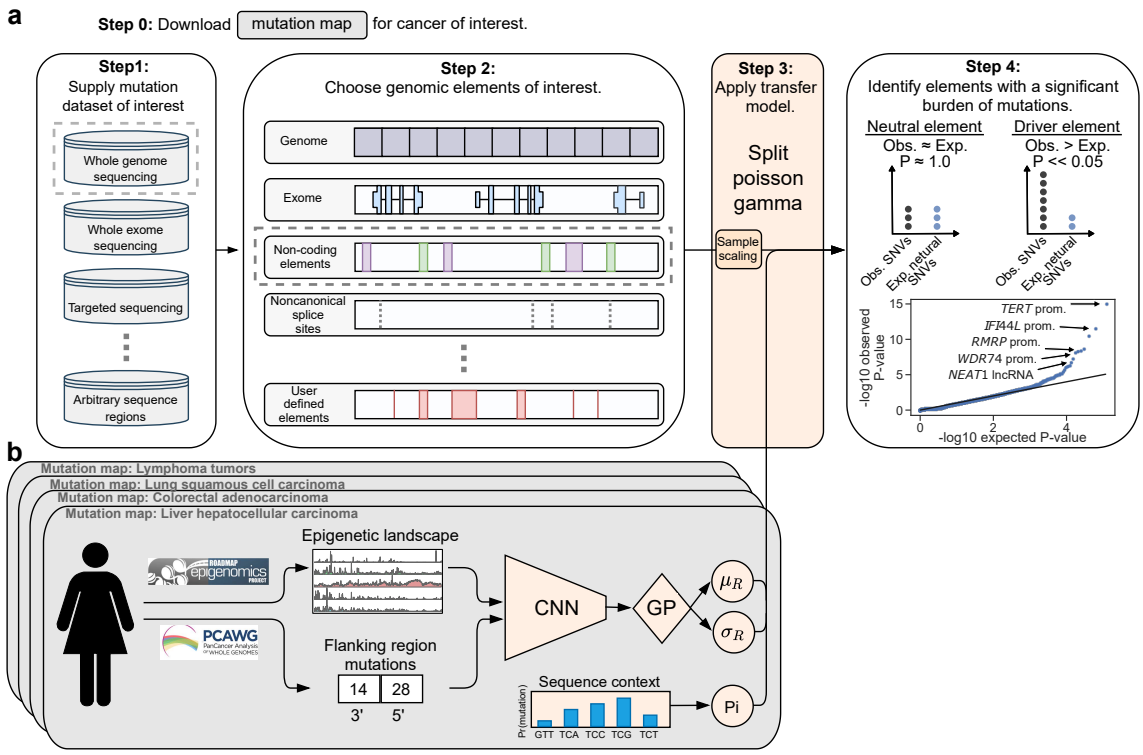

Figure 6-5: Detailed overview of the Dig model. **a,** Dig takes as input somatic mutations (SNVs and/or indels) (Step 1) identified from a cancer cohort sequenced with any methodology and a set of genomic elements of the user's interest (Step 2). The neutral mutation rate from an available neutral somatic mutation map (detailed in panel b) is transferred to the selected SNV dataset via a closed-form probabilistic model (a split-Poisson gamma distribution [270]), that infers only a single scaling parameter at runtime (Step 3); then, a P-value for positive selection is calculated for each element by comparing the number of observed mutations to the number of expected neutral mutations (Step 4). **b,** A neutral mutation map for a particular cancer consists of 1) the mean and variance of the number of neutral mutations in kilobase-scale regions of the genome (default: 10kb) as inferred by a convolutional neural network (CNN) and Gaussian process (GP) based on 735 epigenetic features from the Roadmap Epigenomics dataset and ENCODE (and optionally the number of mutations observed 100kb up- and downstream of the region in the a training cancer cohort dataset); and 2) a sequence context model that provides the genome-wide likelihood of a mutation given its sequence context (default: trinucleotide sequences).

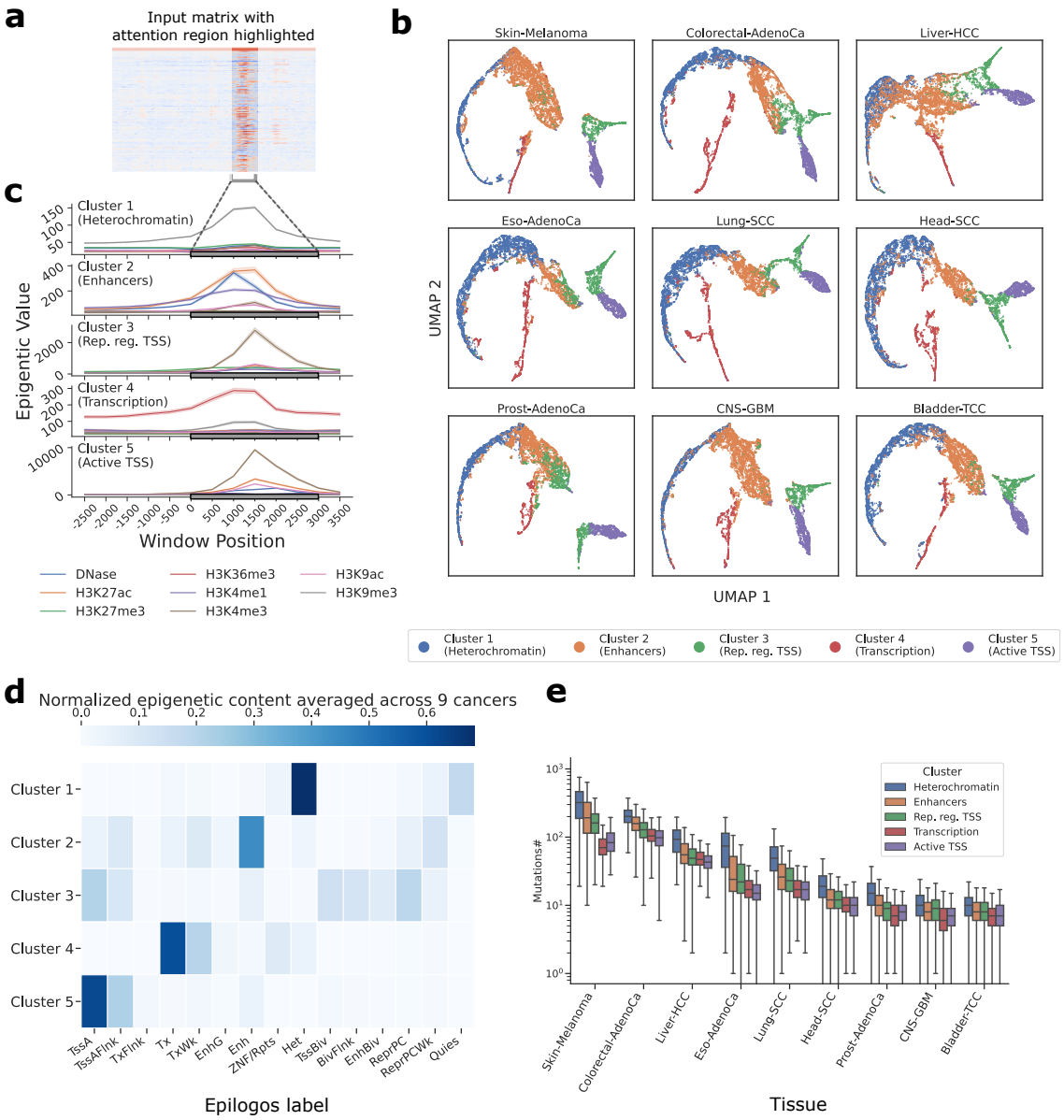

Figure 6-6: Epigenetic input features used by Dig to predict mutation density in nine cancer types. **a,** An example of a feature map across the 735 input features in a 50kb region. The attention column is highlighted. **b,** UMAP visualization of the epigenetic content within attention columns, produced by averaging the same chromatin marks (e.g., H3K27ac) across tissues, for nine types of cancer. The epigenetic content consistently formed five clusters in each cancer type. **c,** An example of the average epigenetic content of each cluster from lung squamous cell carcinoma. Each chromatin mark is the average across tissues with 95% CI. **d,** The epigenetic content of each cluster as determined by epilogos [179], averaged across the nine cancer types. **e,** Boxplots of the number of mutations in regions containing an attention column from a given cluster, stratified by cancer type. Skin-melanoma: N=107 samples, Colorectal-AdenoCa: N=50 samples, Liver-HCC: N=314 samples, Eso-AdenoCa: N=97 samples, Lung-SCC: N=47 samples, Head-SCC: N=56 samples, Prost-AdenoCa: N=199 samples, CNS-GBM: N=39 samples, Bladder-TCC: N=23 samples.

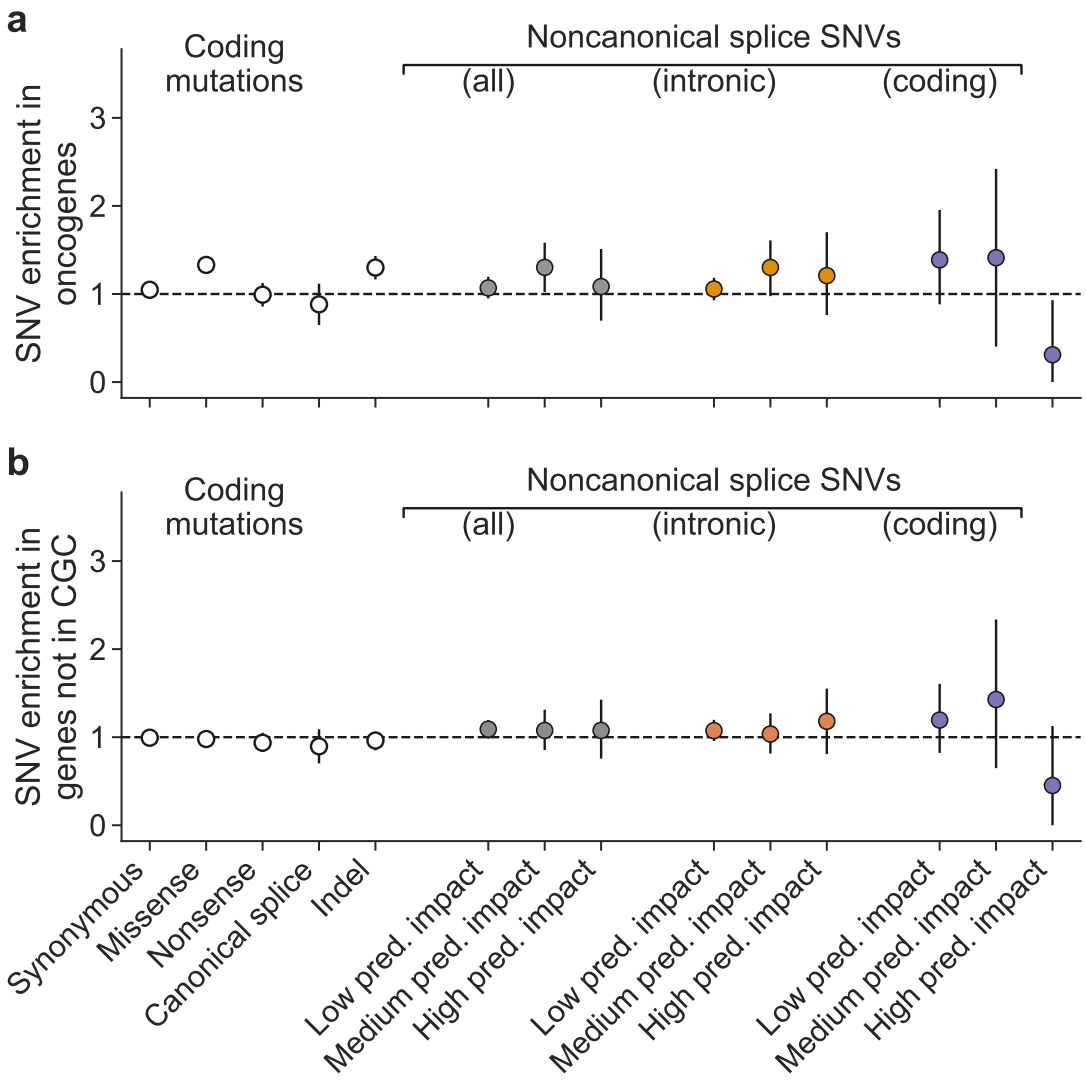

Figure 6-7: Cryptic splice SNV enrichment in oncogenes and genes not in the CGC. Estimated SNV enrichment with 95% CIs as in fig. 6-3b for oncogenes in the CGC, **a**, and 500 randomly selected genes not in the CGC, **b**. Enrichment is not significant in any category after accounting for multiple hypothesis testing except missense mutations and indels in oncogenes, as expected. (N=2,279 samples in each panel; number of mutations per category in Supplementary table D.12).

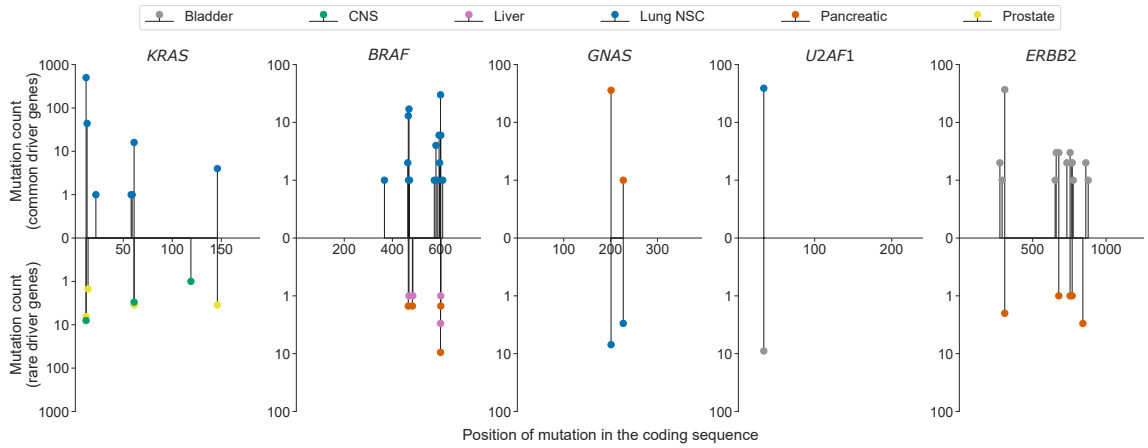

Figure 6-8: Examples of distribution of activating mutations in gene-tumor pairs. Top y-axis: distribution in cancers for which the gene is a known common driver. Bottom y-axis: distribution in cancers for which the gene is a newly proposed rare driver. The genes shown are the five long-tail genes with the highest carrier frequency across the cancer types tested. Color of the ball indicates cancer type.

Chapter 7

Conclusion

This thesis aims to optimize the utility of ML frameworks to life science research. Throughout its pages, we reviewed current drawbacks and potential solutions such as bespoke datasets and tailored computational methods. We highlight two unmet challenges that are hindering the contribution of ML to the study of complex biological systems: 1) lack of general-purpose datasets at scale, and 2) limited interpretability of deep-learning models.

The majority of living systems are simultaneously dynamic and brittle, making it extremely challenging to curate sufficient information to disentangle their inherent noise from the desired signal. Therefore, many researchers still rely on siloed and narrow datasets that are collected independently per study, limit reproducibility, and are unsuitable for ML models. In this work, we use neurolinguistics - the study of language processing in the brain - as a hallmark for a complex system that can greatly gain from multi-purpose datasets at scale. In chapter 3, we describe the curation of a first-of-a-kind multimodal treebank, the AMMT. In chapter 4, we augment the AMMT with aligned intracranial neural signals to study how the brain process different POS during passive listening. We found that by using naturalistic stimuli we were able to collect a sufficient amount of data in highly restrictive settings. The breadth of the curated data enabled the investigation of a variety of different questions, control for a plethora of confounding factors, and application of deep-learning models. Specifically, we demonstrated the critical nature of language context that

differentially modifies the neural pattern and latency evoked by nouns and verbs. We also identified a tightly-connected network of brain areas that anticipates, analyzes, and transmits the POS of an incoming word.

Unlike digital ML tasks (e.g. image object recognition or Netflix video recommendations), life sciences have an additional requirement to benefit from computational methods. Their processing needs to be humanly explainable. Notably, deep-learning models are notorious black-boxes, providing almost no insights into their decision-making process or certainty of their outcome. A biologist needs to know *how* two proteins interact, a chemist needs to know *where* a functional group will generate the desired response, and a physician needs to know *why* a patient will react to a drug as predicted. For the most part, answers to these questions exceed the need for superior performance. In chapter 5, we developed an interpretable and probabilistic deep-learning approach to efficiently model discrete non-stochastic processes at multiple resolutions. We apply this framework to model somatic mutational patterns genome-wide in seconds instead of hours or days. In chapter 6, we extended this method to identify somatic mutations that putatively drive cancer all across the genome. The certainty estimation of the output allows us to identify observed anomalies with respect to expected values, down to a base-pair resolution, and the interpretation of the input implies what functional regions govern these expectations. The analysis of 37 cancer types revealed that cryptic splice mutations, 5' UTR mutations, and mutations that occur infrequently in genes may all contribute to the development and progression of cancer. Moreover, a limited set of local chromatin states explains nearly all variance of regional mutation rates.

Chapters 3-6 included sections discussing the specific contributions, limitations, and future directions of the above. We will not repeat those points here. Rather, we conclude by reflecting on major open challenges and the future prospects of machine learning in human biology.

This work adds to the growing body of literature demonstrating the ability of ML approaches to solve complex problems in biology [66, 3, 252, 134]. However, there is great value to be gained from an even tighter integration of the fields. Progress

can be made by carefully incorporating human knowledge into computational models. Indeed, language models with embedded knowledge of amino acid code are solving fundamental problems in predicting protein folding and protein properties [122, 279, 134, 17] and architectures that explicitly reproduce the underlying molecular networks of a cell are enabling interpretability without loss of predictive power [162, 146, 74]. Yet, such examples are sparse and require tailored solutions.

Moreover, ML models trained over static datasets can only obtain so much. Active-learning paradigms provide robust feedback between computation and experimentation – computational predictions are tested in the lab and results are used to improve the computational model – and therefore yield higher experimental efficiency and substantially improve the accuracy of computational models. Recent work has begun to demonstrate the utility of this approach [121], and we expect such research to prove highly fruitful in the coming years.

The future of computation and life sciences will be intricately linked. Computation will be needed to solve key challenges in biology and the complexity of biology will drive innovation in the computational sciences. Inherent data limitations of living systems will force ML models to be more flexible to smaller and noisier datasets. Such enhanced potential will then give birth to novel branches of computation frameworks. This work and others merely reveal the tip of the computational biology iceberg, there are many more puzzles of life to be assembled by learning machines.

Appendix A

Supplementary Information Related to Chapter 3

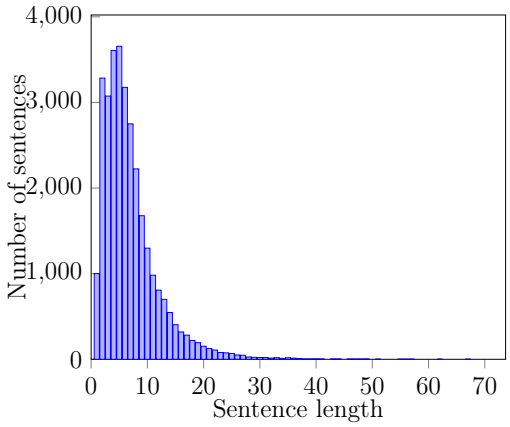

Figure A-1: Distribution of sentence lengths in AMMT. Most sentences are quite short. The mean sentence length is 6.97 words long. Compare to standard corpora derived from written sources like the English Web Treebank (15.33 words/sentence) long and the Penn Treebank (23.73 words/sentence in the test set).

POS	Count
ADJ	9829
ADP	12464
ADV	13688
AUX	18965
CCONJ	3746
DET	12984
INTJ	6275
NOUN	25457
NUM	1835
PART	7202
PRON	36370
PROPN	8679
PUNCT	30301
SCONJ	2140
SYM	10
VERB	28139
X	6

Dependencies	Count
nsubj	25050
advmod	14003
obj	12825
det	12325
case	11274
aux	9286
cop	7830
obl	6653
mark	5693
amod	4958
xcomp	4306
nmod:poss	3996
discourse	3912
cc	3682
compound	3335
conj	3322
vocative	3134

Table A.1: The distribution of POS tags (left), and the most common dependencies (right). There is a long tail of dependencies.

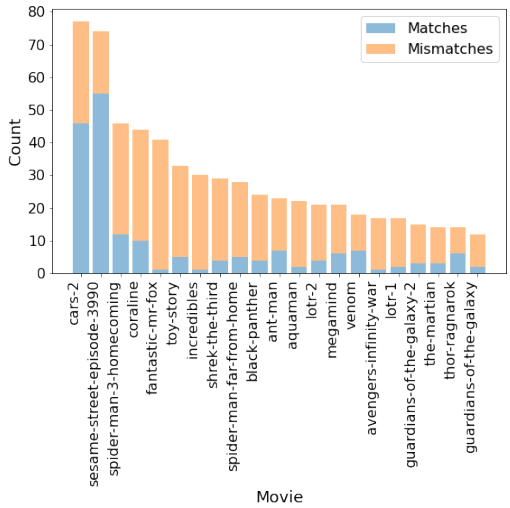

Figure A-2: COCO classes noun-object agreements per movie (sorted by number of nouns). All nouns corresponding to one of the 80 COCO classes (orange) vs their corresponding objects in the video during the noun utterance (blue) per movie.

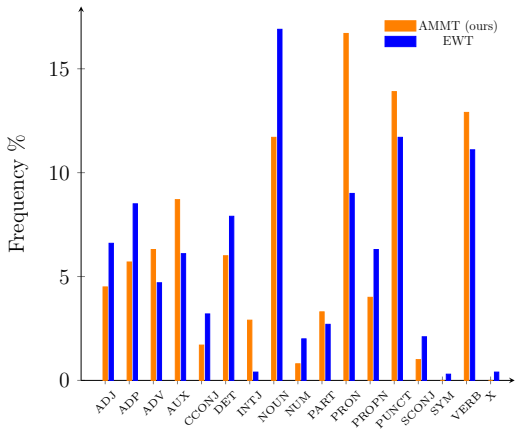

Figure A-3: Comparing POS frequency in EWT, a treebank derived from text on the web, and AMMT, our new benchmark derived from spoken language. Among many differences, note that in AMMT, nouns are much less common and pronouns are far more common.

Movie	Year	IMDb ID	Time (s)	Sentences	Tokens	Types	Rating	Frames
Ant-Man	2015	tt0478970	7027	1412	9846	1956	PG-13	168507
Aquaman	2018	tt1477834	8601	1003	7218	1563	PG-13	206251
Avengers: Infinity War	2018	tt4154756	8961	1372	8479	1780	PG-13	214884
Black Panther	2018	tt1825683	8073	1139	7571	1628	PG-13	193590
Cars 2	2011	tt1216475	6377	1801	11404	2060	G	152920
Coraline	2009	tt0327597	6036	933	5428	1251	PG	144743
Fantastic Mr. Fox	2009	tt0432283	5205	1162	8457	1892	PG	124815
Guardians of the Galaxy 1	2014	tt2015381	7251	1104	8241	1799	PG-13	173878
Guardians of the Galaxy 2	2017	tt3896198	8146	1180	9332	1839	PG-13	195341
The Incredibles	2003	tt0317705	6926	1408	9369	1966	PG	166085
Lord of the Rings 1	2001	tt0120737	13699	1424	10538	2011	PG-13	328502
Lord of the Rings 2	2002	tt0167261	14131	1620	11017	2085	PG-13	338861
Megamind	2010	tt1001526	5735	1351	8833	1748	PG	137525
Sesame Street Ep. 3990	2016	tt13725852	3440	718	4218	804	TV-Y	103096
Shrek the Third	2007	tt0413267	5568	999	7192	1586	PG	133520
Spiderman: Far From Home	2019	tt6320628	7764	1705	12004	1988	PG-13	186180
Spiderman: Homecoming	2017	tt2250912	8008	1993	12258	2107	PG-13	192031
The Martian	2015	tt3659388	9081	1421	11360	2210	PG-13	217762
Thor: Ragnarok	2017	tt3501632	7831	1471	9651	1806	PG-13	187787
Toy Story 1	1995	tt0114709	4863	1240	7194	1545	G	116614
Venom	2018	tt1270797	6727	1301	7859	1527	PG-13	161313

Table A.2: Name, unique identifier (IMDb ID), and statistics for the 21 movies from which AMMT is derived. Movies were selected to be appropriate for most ages enabling a wide range of experiments. Movies are not randomly sampled; they were selected for their verbose scripts and subjects entertainment during experiments. For more on IMDb identifiers, see `https://developer.imdb.com/documentation/key-concepts#imdb-ids`

Appendix B

Supplementary Information Related to Chapter 4

B.1 Supplementary Methods

B.1.1 Cortical surface extraction and electrode visualization

For each subject, pre-operative T1 MRI scans without contrast were processed with FreeSurfer's `recon-all` function with `-localGI`, which performed skull stripping, white matter segmentation, surface generation, and cortical parcellation [69, 89, 86, 85, 83, 84, 88, 133, 149, 219, 222, 228, 67, 87, 116, 241, 229, 211, 210, 212]. iELVis [109] was used to co-register a post-operative fluoroscopy scan to the preoperative MRI. Electrodes were manually identified using BioImageSuite [132], and then assigned to one of 74 regions (according to the Destrieux atlas [70]) using FreeSurfer's automatic parcellation. The alignment to the atlas was manually verified for each subject.

We excluded a total of 66 electrodes from two subjects from all analyses and plots due to tumors and lesions: one subject had 43 electrodes on the border of and within a tumor, and another subject had 23 electrodes on the border of and going through a lesion (suspected from prior surgery). For depth electrodes in the white matter, if they were within 1.5 mm of the gray-white matter boundary, they were projected to the nearest point on that boundary, and were labeled as coming from that region (for

the purposes of region significance analyses). This procedure is very similar to the post brain-shift correction methods used for electrocorticography electrodes [271]. For solely visualization purposes, all electrodes identified to lie in the gray matter or on the gray-white matter boundary were first projected to the pial surface (using nearest neighbors), and then mapped to an average brain (using Freesurfer's fsaverage atlas) for the visualizations shown in Extended Figures fig. 4-8.

B.1.2 Audio transcription and alignment

The audio track of each movie was first annotated by commercial services (`Rev.com` and `HappyScribe.com` depending on the movie) and manually corrected by trained annotators. A custom tool was developed to refine the alignment via an auditory spectrogram of 4 seconds at a time and slowed-down audio track. Annotators were instructed to adjust the onset and offset of every word to align with the spectrogram and their perception of when the word started and ended. The audio annotation tool automatically played the audio segment corresponding to each word to allow annotators to verify their work. As the audio was played a line marked the location of the audio sample in the spectrogram in real time.

Since speech recognizers often misused or missed critical punctuation marks, these were inserted by annotators manually. Sentences were then manually segmented. Annotators were instructed not to use abbreviations, even if they are common. Annotators marked audio segments that consisted of overlapping speech or signing. These were removed from the dataset. All foreign language was marked and removed from the dataset. Annotators were instructed to transcribe literally, i.e, contractions were used in the transcript only when spoken as such. Similarly, foreshortened words, e.g., goin' vs going, were transcribed as such when used by speakers. Cardinal numbers were spelled out. Longer numbers were spelled out as spoken, including conjunctions such as "and". All overheard words were transcribed, even when they could not easily be localized on the spectrogram, for example, short words such as "to" can sometimes be heard but no specific segment of the spectrogram seems to correspond uniquely to such words. In this case annotators were asked to mark their onset and offset as

they heard the words. Transcripts are as spoken, without correction, even when the speaker erred omitting a word or using a word inappropriately.

B.1.3 Part of speech tagging

We used a state-of-the-art syntactic parser, Stanford NLP Group's Stanza qi2020stanza, to parse every sentence. POS tags were recorded for every word. The homonym set was initially constructed automatically by taking every word which occurred as both a noun and a verb. This set was then manually inspected against the original audio track to ensure that homonym pairs were pronounced the same, removing any that were not.

B.1.4 Confounding features

Attributes unrelated to whether a word is a noun or a verb may still be correlated with part of speech. Models which naively analyze neural data may be decoding part of speech from the neural activity evoked by these correlated but undesirable attributes. We create an extensive array of such features and include them in our analysis to hone in on part-of-speech distinctions specifically. While these confounds are undesirable in our analysis, they may serve as objects of study in their own right. We extracted potentially-confounding features from the video and audio tracks, and from the transcript itself. Specifically, we extract 33 scalar features and 2 vector features that were included in the analyses (see Extended Figures table 4.1). We also extract 4 string features for general purposes of this work as well as the service of additional works.

The visual scene scalar features were extracted from the middle frame presented during a word utterance via OpenCV 4.4.0 [37]. Brightness was quantified as the average pixel HSV value channel. Flow vectors were computed as dense optical flow over grey-scale frames via the OpenCV `calcOpticalFlowFarneback` function (pyramid scale 0.5, 5 levels, window size 11, 5 iterations, pixel neighborhood of 5, and smoothing of 1.1). Number of faces per-frame was estimated via the OpenCV

`CascadeClassifier` function with the Haar cascade frontal face default classifiers over gray-scale frames (scale factor: 1.1, minimum neighbours: 4). The first frame of every word utterance was mean-normalized and than passed through a pretrained ResNet-50 object detector (Torchvision 0.6.1) to compute a visual vector image embedding (size 2,048) as the last feature layer of the model.

The auditory scalar features were collected with the Python Librosa package (0.7.2) [172], an open source audio analysis library. Sound intensity and mean frequency of the audio track during word utterance were estimated, as well as their change relatively to the preceding $500ms$ window. The average intensity of the audio segment was computed in two ways, first, as the root-mean-square (RMS) (`rms` function, frame and hop lengths 2048 and 512 respectively) of that segment, and second, as the magnitude of the Mel-spectrogram. Magnitude and pitch were extracted using Librosa's `piptrack` function over a Mel-spectrogram (sampling rate 48,000 Hz, FFT window length of 2048, hop length of 512, and 128 mel filters). Auditory vector embeddings were computed as the flattened log-Mel-spectrogram of the $500ms$ word utterance window (size $128 \times 47 = 6016$).

Surprisal was quantified as the negative-log word probability. Word probabilities were estimated by four different language models, ranging from word frequency in the the BLLIP corpus [56] to transformer models. GPT-2 probabilities were computed via GPT-2 large using the Hugging Face Transformers 3.0.0 library [265]. Word particle surprisal were combined by summation. LSTM (layers: 2, dropout: 0.2, input/output dimensions: 200) probabilities were pretrained on BLLIP. N-gram [42] probabilities (N=5) were computed via the Python KenLM language model [120] (`full_scores` function (beginning of sentence: true, end of sentence: false), pre-trained over BLLIP.

Word syllable and phoneme numbers were estimated via the Python Syllables package (0.1.0) and the Python NLTK package (3.4.4) [31] (Carnegie Mellon Pronouncing Dictionary Corpus Reader) respectively. All Universal Dependency features were inferred using the standard English model of the Stanza Natural Language Processing toolkit [206] and then manually corrected via a single trained annotator over the course of a year (see Methods appendix B.1.3).

B.2 Supplementary Figures

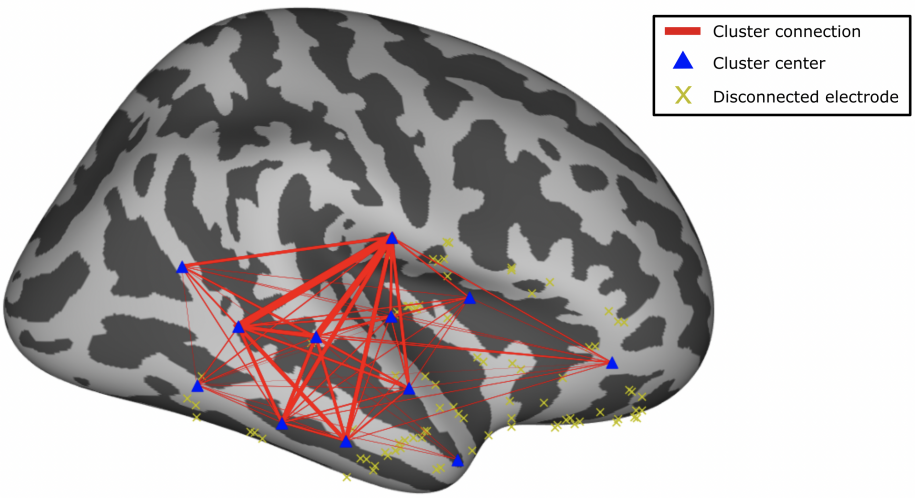

Figure B-1: The language connectivity map for a single subject's electrodes. A pair of electrodes are connected if they are significantly more correlated when the subject is hearing speech vs non-speech. This highlights functional connectivity specific to language processing rather than other faculties. Connected electrodes are spatially clustered and represented by their cluster centers (blue triangles). Rad line width corresponds to the cumulative number of connections between two clusters.

B.3 Supplementary Tables

Subject	Age	Gender	Movies	Time (h)	# Sentences	# Words	# Unique lemmas	# Electrodes	# Probes	Was included
1	19	M	71819	6.14	4054	29468	5908	154	13	Yes
2	12	M	234891721	15.49	9092	60958	12243	162	47	Yes
3	18	F	51112	9.50	4845	32959	6156	134	12	Yes
4	12	F	101315	5.06	3758	25394	5300	188	15	No
5	6	M	7	1.45	1162	8457	1892	156	12	Yes
6	9	F	61320	8.02	3524	21455	4544	164	12	No
7	11	F	513	3.36	3152	20237	3808	246	18	No
8	4	M	14	0.96	718	4218	804	162	13	Yes
9	16	F	1	1.95	1412	9846	1956	106	12	Yes
10	12	M	516	3.93	3506	23408	4048	216	17	Yes

Table B.1: All subjects language, electrodes and personal statistics. Columns from left to right are the subject's ID and information (age and gender), the the IDs of the movies they watched (corresponding to Extended Figures table B.2), the cumulative movie time (hours), number of sentences, number of words (tokens) and number of unique words (types), as well as the number of probes the subject had and their corresponding number of electrodes. Right most column indicates which subjects were included in the analyses of this study.

#	Movie	Year	Time (s)	# Sentences	# Words	Unique words	Nouns	Unique nouns	Verbs	Unique verbs
1	Antman	2015	7027	1412	9846	1956	1370	712	1538	581
2	Aquaman	2018	8601	1003	7218	1563	1066	517	1094	508
3	Avengers: Infinity War	2018	8961	1372	8479	1780	1081	608	1294	485
4	Black Panther	2018	8073	1139	7571	1628	1084	544	1199	506
5	Cars 2	2011	6377	1801	11404	2060	1576	737	1649	563
6	Coraline	2009	6036	933	5428	1251	759	407	817	353
7	Fantastic Mr. Fox	2009	5205	1162	8457	1892	1240	690	1240	490
8	Guardians of the Galaxy 1	2014	7251	1104	8241	1799	1101	615	1235	521
9	Guardians of the Galaxy 2	2017	8146	1180	9332	1839	1210	623	1368	533
10	Incredibles	2003	6926	1408	9369	1966	1234	659	1545	582
11	Lord of the Rings 1	2001	13699	1424	10538	2011	1470	681	1480	595
12	Lord of the Rings 2	2002	14131	1620	11017	2085	1593	760	1587	631
13	Megamind	2010	5735	1351	8833	1748	1183	610	1340	496
14	Sesame Street Ep. 3990	2016	3440	718	4218	804	716	233	674	211
15	Shrek the Third	2007	5568	999	7192	1586	989	568	1072	418
16	Spiderman: Far From Home	2019	7764	1705	12004	1988	1442	660	1755	555
17	Spiderman: Homecoming	2017	8008	1993	12258	2107	1591	795	1794	569
18	The Martian	2015	9081	1421	11360	2210	1781	826	1686	630
19	Thor: Ragnarok	2017	7831	1471	9651	1806	1183	604	1440	546
20	Toy Story 1	1995	4863	1240	7194	1545	1039	561	1015	388
21	Venom	2018	6727	1301	7859	1527	892	509	1200	427

Table B.2: Language statistics for all movies. Columns from left to right are the movie's ID, name, year of production, length (seconds), number of sentences, number of words (tokens), number of unique words (types), number of nouns, number of unique nouns, number of verbs and number of unique verbs.

Appendix C

Supplementary information related to Chapter 5

In this appendix, we provide detailed information on:

1. The data used in this work including its origin and all preprocessing steps.

2. Additional method details including:

    - A derivation of the closed-form marginal distribution of the graphical model presented in main text.
    - Architecture and training details of all models.
    - How the genome-wide search for driver mutations was performed.

The appendix also includes an analysis of the sensitivity of negative binomial regression to detect well-known drivers genome-wide and additional figures that provide context to results presented in the main paper.

C.1 Supplementary Materials and Methods

C.1.1 Data

Epigenetic tracks

We obtained 733 $-log_{10}$(P-value) chromatin tracks representing the epigenetic organization of 111 human tissues from Roadmap Epigenomics [218] (see Appendix table "predictor_track_descriptions.csv"). These tracks measure the abundance of a particular chromatin mark genome-wide, with smaller (more significant) p-values reflecting a greater abundance of the chromatin mark at a genomic position. Chromatin marks are chemical modifications of histones, the proteins used to package DNA within a cell. We additionally obtained 10 replication timing tracks from the ENCODE consortium. Replication timing assays measure the relative time at which each position in the genome is replicated during cell division. For non-overlapping regions R of predefined size and location (see main text for more details), we extracted the signal for each epigenetic track using 100 bins per region with pybbi [2]. We additionally calculated the average nucleotide content in each window by assigning each nucleotide a numeric value between 1 and 5 and taking the average across a bin (N [unspecified nucleotide] = 1, A = 2, C = 3, G = 4, T = 5), and we calculated the GC content as the percent of G and C nucleotides in a bin, resulting in a total of 735 epigenome tracks per region. The mean values for each region were calculated as the mean chromatin signal for each track in the region.

Mutation count data

We downloaded somatic single-base substitution mutations identified in the ICGC subset of the Pan-Cancer Analysis of Whole Genomes Consortium cohorts of esophageal adenocarcinoma, skin melanoma, stomach adenocarcinoma, and liver hepatocellular carcinoma. These data are freely available for download from the International Cancer Genomics Consortium data portal (see fig. C-1). We excluded mutations on the sex chromosomes (X and Y) because males and females carry different sets of these

chromosomes, leading to differential mutation patterns. We summarized the data as mutation counts per window for window sizes of 50bp, 100bp, 500bp, 1kb, 5kb, 10kb, 25kb, 50kb, 100kb, and 1Mb.

Restriction to regions of high mappability

High-throughput genome sequencing works by randomly reading millions of short sequences of nucleotides (36-150 bases in length) from a target genome. These "reads" are then mapped to the human reference genome to reconstruct the target. A challenge is that short sequences of k nucleotides (kmers) can occur multiple times in the genome. This results in ambiguous mappings for some reads and thus a degradation of data quality in regions composed of many kmers that occur multiple times across the genome. Following previous work [200], we removed regions of the genome with low quality data by calculating a mappability score for each region. Mappability scores reflect how many times a particular kmer occurs in the genome and have been pre-computed for the human reference genome GRCh37. We required that a region's average mappability score based on 36mers (e.g. average across all sequences of 36 nucleotides in the region) be >70%, reflecting that all 36mers in the region be >70% unique. The majority of the genome passed this threshold; for 10kb regions, for example, >75% of the genome passed this threshold. We chose to measure mappability with 36mers because this was the length of read used to generate the Roadmap Epigenomics sequencing data.

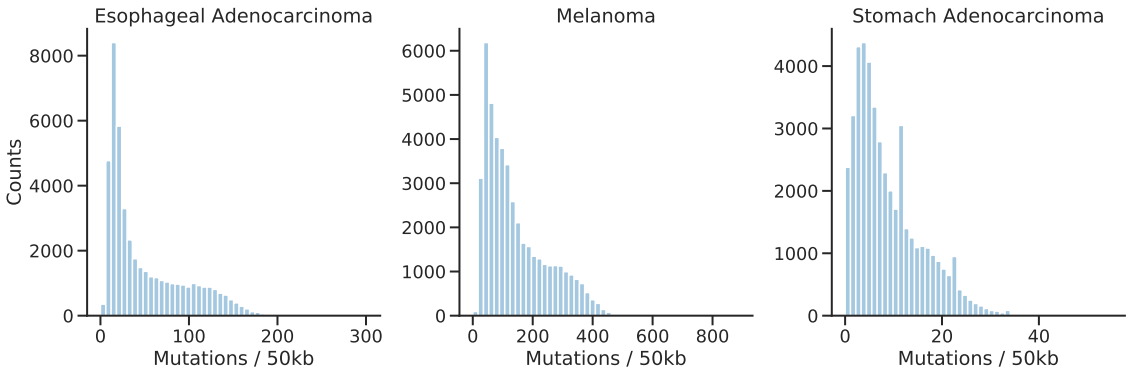

Figure C-1: Distribution of mutation counts in 50kb windows tiled across the genome with 36mer uniqueness >70% (see section C.1.1) for esophageal adenocarcinoma, skin melanoma, and stomach adenocarcinoma. Of note: esophageal adenocarcinoma has a highly skewed distribution, skin melanoma has high mutation counts relative to the other cancers, and stomach adenocarcinoma has low mutation counts relative to the other cancers.

Synthetic data simulation

We generated synthetic datasets for each of the cancers in order to have datasets with known mean and variance rate parameters. To generate the datasets, we used a k-nearest-neighbors strategy to identify the 500 nearest neighbors for each region. The mean and variance for that region were then taken to be the empirical mean and variance calculated from the 500 nearest neighbors. The number of "observed" mutations was then randomly sampled from a binomial defined by the mean and variance parameters. It is important to note that these datasets are purely derived for the purpose of comparing methods over datasets with a known ground-truth. They do not reflect mutation patterns in the real datasets. The specific steps to generate the simulated data were:

1. Generate vectors of the mean values for each of the 735 tracks (733 epigenetic tracks, GC content track, and average nucleotide content track) in 50kb regions of the genome with 36mer uniqueness >70%.

2. Perform ordinary least-squares (OLS) regression of the mean vectors against the observed number of mutations in each 50kb window for that cancer.

3. Scale each value in the feature vectors by its corresponding coefficient from

OLS and compress the weighted mean vectors to 50 components using Principal Components Analysis (capturing >94% of the variance for each cancer).

4. For each region R, perform k-nearest-neighbor clustering with Euclidean distance to identify its 500 nearest neighbors in the PC space. Define the mean μ_R and variance σ_R^2 of the mutation rate in R to be the mean and variance of the KNN cluster.

5. For region R, randomly draw a new "observed" number of mutations from a negative binomial distribution defined using the associated mean and variance. Specifically, $X_R \sim NB(\alpha, 1/(\theta+1))$ where $\alpha = \mu_R^2/\sigma_R^2$ and $\theta = \sigma_R^2/\mu_R$

We created two versions of the simulated data, one in which all regions in the genome were used to estimate the rate parameters and one in which rate parameters were estimated separately within independent train and test subsets. Results were qualitatively indistinguishable.

C.1.2 Graphical model derivation

Here we derive the closed form negative binomial distribution presented in the main text as the graphical model marginal distribution over events at some unit i in a region R. We use the following notation:

- M_i: # mutations observed at pos i (observed)
- p_i: genome-wide probability of observing a mutation at the nucleotide context of i (inferred)
- $\tilde{p}_i$: normalized probability of observing a mutation at i in region R (inferred)
- λ_R: the background mutation rate in region R (unobserved)
- X_R: # background mutations in region R (unobserved)
- μ_R: the expected background mutation rate in region R (inferred)

- σ_R^2: the variance of background mutation rate in region R (inferred).
- η_R: covariates associated with the behavior of the stochastic process within R (observed)

As presented in the main text and main Figure 1, the graphical model implies the factorization

$$Pr(M_i, X_R, \lambda_R | \alpha_R, \theta_R, \tilde{p_i}; \eta_R) = Pr(M_i = k | X_R, \tilde{p_i}; \eta_R) \cdot Pr(X_R = x | \lambda_R; \eta_R) \cdot Pr(\lambda_R | \alpha_R, \theta_R; \eta_R) \tag{C.1}$$

where

$$\alpha_R = \mu_R^2 / \sigma_R^2$$
$$\theta_R = \sigma_R^2 / \mu_R.$$

Since η_R is a given in each equation, we suppress it for notational ease.

To marginalize out X_R, we note that

$$Pr(M_i = k | \lambda_R) = \sum_{x=k}^{\infty} Pr(M_i = k | X_R, \tilde{p_i}) \cdot Pr(X_R = x | \lambda_R)$$

is equivalent to a split Poisson process [97]. Thus

$$Pr(M_i = k | \lambda_R) = \text{Possion}(M_i = k; \tilde{p_i} \lambda_R). \tag{C.2}$$

We now marginalize out the unknown rate parameter λ_R.

$$\begin{aligned} P(M_i = k | \tilde{p_i}, \alpha_R, \theta_R) &= \int_0^{\infty} P(M_i = k | \lambda_R; \tilde{p_i}) P(\lambda_R | \alpha_R, \theta_R) \mathrm{d}\lambda_R \\ &= \int_0^{\infty} \frac{(\tilde{p_i} \lambda_R)^k}{k!} e^{-\tilde{p_i} \lambda_R} \frac{1}{\Gamma(\alpha_R) \theta_R^{\alpha_R}} \lambda_R^{\alpha_R - 1} e^{-\lambda_R / \theta_R} \mathrm{d}\lambda_R \\ &= \frac{{\tilde{p_i}}^k}{k! \Gamma(\alpha_R) \theta_R^{\alpha_R}} \int_0^{\infty} \lambda_R^{\alpha_R + k - 1} e^{-\lambda_R (\tilde{p_i} + 1/\theta_R)} \mathrm{d}\lambda_R. \end{aligned}$$

Making the substitution $t = \lambda(\tilde{p_i} + 1/\theta_R)$ and noting that the resulting integrand

is an unnormalized gamma distribution, we have:

$$\begin{aligned} P(M_i = k|\tilde{p_i}, \alpha_R, \theta_R) &= \frac{\tilde{p_i}^k}{k!\Gamma(\alpha_R)\theta_R^{\alpha_R}}\Gamma(\alpha_R + k)\left(\frac{1}{\tilde{p_i} + 1/\theta_R}\right)^{\alpha_R+k} \\ &= \frac{\Gamma(\alpha_R + k)}{k!\Gamma(\alpha_R)}\left(\frac{\tilde{p_i}\theta_R}{\tilde{p_i}\theta_R + 1}\right)^k \left(\frac{1}{\tilde{p_i}\theta_R + 1}\right)^{\alpha_R} \\ &= \mathrm{NB}\left(M_i = k; \alpha_R, \frac{1}{\tilde{p_i}\theta_R + 1}\right). \end{aligned}$$

C.1.3 Overview of parameter estimation procedure

Estimation of regional rate parameters: As training data, we use a set of input matrices $\{\eta_R;\ R \in \mathcal{T}\}$ and associated mutation counts $\{X_R;\ R \in \mathcal{T}\}$. First, a CNN is trained to take η_R as input and predict X_R as output, using mean squared error loss. The final 16-dimension feature vector of the trained CNN is then used as input to train a Gaussian process to predict the mutation count X_R and the associated estimation uncertainty by maximizing the likelihood of the observed data. The mean and variance output by the GP were used as estimates for μ_R and σ_R^2.

Estimation of time-averaged event probabilities: the time-average probability of an event at p_i was estimated based on it's trinucleotide composition, n, t, n' where n is the nucleotide at $i-1$, t is the nucleotide at i and n' is the nucleotide at $i+1$ in the reference genome. We first counted every occurrence of n, t, n' in the human genome and then counted the number of times the middle nucleotide of the 3mer was mutated across the genome. The maximum likelihood estimate of p_i is then the ratio of the number of observed mutations of the 3mer divided by the total occurrences of the 3mer.

C.1.4 Regional parameters estimation methods

To compute a model's R^2 accuracy to μ_R and σ_R^2 for regions R of size S, the genome was divided into non-overlapping contiguous segments of size S. To assure high data quality, any region with mappability score $< 70\%$ was excluded from further analysis. The remaining windows (accounting for more than 75% of the genome) were randomly

divided into train and test sets in an 80–20 split respectively. The test set was held-out and served solely for evaluation purposes. The train set was then divided into train and validation sets by another 80–20 split respectively (train set = 64%, validation = 16%, and test = 20% of the considered regions with mappability score $< 70\%$, see appendix C.1.1).

Gaussian process feature vector generation

All networks were independently trained for 20 epochs with a batch size of 128 samples and using the Adam optimizer to minimize mean squared error loss to either the true mutation count (CNN and FCNN) or input tensor (AE). After training the model parameters using the train set, predictions over the held-out test set were computed by 1) extracting the last 16-dimensional feature layer (middle feature layer for AE) for all sets over the best performing model over the validation set across all epochs (according to the validation accuracy); 2) training multiple GPs (typically 10) to predict mutation counts using the 16 dimension feature vectors of the train set as input (see appendix C.1.4 for details); 3) taking the mean μ_R and σ_R^2 of all 10 runs over the test set as the ensemble prediction of the model. All neural network models were implemented in Pytorch [199].

1. **Convolutional neural network (CNN):** The CNN contains 4 convolutional blocks with 2 batch normalized convolutional layers and ReLU activation. The first block transformed the input tensor from 735×100 to 256×50 with 256 channels and a double stride. The other blocks are ResNet-style residual blocks that maintain their input dimension to facilitate residual connections, with 256, 512, and 1024 channels respectively. Between each of the 3 residual blocks there is a double stride (ReLU activated and batch normalized) convolutional layer, which divides the tensor length by two and doubles its height with additional channels. The output of the last residual block is flattened and passed through 3 fully-connected layers. The first two are ReLU activated and reduce the dimensionality of the tensor to 128 and 16 dimensions respectively. The last uses linear functions to reduce the tensor to a single cell holding the output

of the regression. This forces a linear relation between the regression output and the last feature layer, thus simplifying the function the GP needs to learn, which we found empirically improves the GP's accuracy.

2. **Fully-connected neural network (FCNN):** The FCNN has an architecture similar to the CNN's 3 fully-connected layers but with an input space of the mean epigenetic vector (735 dimensions). Thus, the FCNN is computationally similar to the CNN, but operates on the mean vector instead of the full matrix as an input. The FCNN is designed to demonstrate maximum performance possible when reducing the input tensor to an averaged feature vector.

3. **Autoencoder neural network (AE):** The encoder of the AE used the same architecture as the CNN, excluding the last linear fully connected layer. The decoder has a mirror architecture with the same number of parameters but differs in the internal design of the convolutional blocks. Convolutional layers were replaced by 1D transpose convolutional layers with no batch normalization and no residual connections. The AE was designed to demonstrate the predictive power of a feature embedding that was not optimized to a specific task but produced in a way comparable to the CNN.

4. **Other dimensionality reduction methods:** PCA was computed using the Python Scikit-learn package with default settings and UMAP was computed via Python's umap-learn package [173] with 20 nearest neighbours and Euclidean distance. Both methods were computed over the entire training set (80%) with no validation set and reduced the mean epigenetic vector dimensionality (735 dimensions) to 16, just like all other models. Prior to processing, we log-transformed the epigenetic data as we found this improved prediction accuracy downstream.

Gaussian process

We implemented a sparse, inducing-point Gaussian process [250] with a radial basis function kernel using Python's GPyTorch package [98]. The GP was optimized with

2000 inducing points using the Adam optimizer for 100 steps. All features were mean-centered and standardized to unit variance prior to training. For each dataset, we ran the GP ten independent times and calculated the ensemble mean of the mean and variance predictions from each of the individual runs. We took these ensemble predictions as the mean and variance for each region.

Alternative models

We implemented previously proposed alternative methods [200, 191, 169] for the estimation of μ_R and σ_R^2 without the use of GP. These methods use the mean epigenetic vector as an input.

1. **Random forest (RF):** RF regression was implemented via the Ensemble Methods module in the Python Scikit-learn package, with a maximum depth of 50 trees. Since RF does not directly compute a variance, we implemented the Jackknife method as described in [258] (we have compared our implementation to [201] and found them highly correlated). Wager et al. suggests that the number of estimators, i.e., trees, must be linearly related to the number of samples to obtain reasonable estimates of the variance. We chose to have one tenth as many estimators as samples in an attempt to keep running time within reasonable limit for datasets of smaller region sizes. Even so, for 10kb regions (containing approximately 300K regions), RF required >24 hours to train.

2. **Negative binomial regression (NBR):** As described in section 3.3.2 of the main text, NBR directly specifies the variance as $\sigma_R^2 = \mu_R(1 + \beta\mu_R)$, where β is an overdispersion parameter. When $\beta = 0$ NBR reduces to Poisson regression, also widely used in the community. NBR was implemented via the discrete module in the Python statsmodels package [227] with the Broyden–Fletcher–Goldfarb–Shanno optimization algorithm and 1k maximum iterations. Epigenetic predictors were log-transformed and reduced to 20 principle components, following the field-standard [169] in both train and test sets. When used to compare against the GM we also included the expected number of mu-

tations based on the sequence context model (see main paper section 3.2) as an exposure term in the model as in previous work [191, 169].

3. **Binomial regression (BR):** Following a previous study [26] that suggested multinomial regression to model multiple types of mutations, we also considered binomial regression (as the binary version of multinomial regression applicable to our simple counts data) as a method to model mutation rates at high resolution. BR was implemented via the generalized linear module in the Python statsmodels package [227]. As in previous work [191, 169], we included the expected number of mutations based on the sequence context model (see main paper section 3.2) as an exposure term in the model. As with NBR, the epigenetic predictors were log-transformed and reduced to 20 principle components for both train and test sets following state-of-the-art recommendations [169].

C.1.5 Empirical variance estimation

For real data, the true variance in mutation counts of a region is unknown. Thus to estimate variance empirically for a given model, we used the following approach:

1. For a region in the test set, perform k-nearest neighbors clustering with Euclidean distance to identify the 500 regions in the train set that are most similar to the region of interest based on the model's feature embedding. For all models, a feature embedding of 16 dimensions was used.

2. Calculate the empirical variance as the variance of the KNN cluster.

Since feature embeddings are model-specific, we calculated an empirical variance estimate per model. The feature-vector embeddings for models specified in section C.1.4 were the feature vectors used as input to the GP. Models specified in section C.1.4 do not create or require comparable feature vectors and therefore were not considered in the main paper results. However, to measure the ability of these methods to estimate empirical variance (Fig. C-4), we computed their feature vectors by 1) taking the dot product of the model parameters and the input data mean vectors and 2) reduced

these scaled vectors to 16 dimensions via a PCA reduction (explaining 80%-95% of the variance across the different region scales). For RF, we took the model parameters to be the feature importance weights derived from the trained forest and for NBR, we used the model coefficients as the parameters.

C.1.6 Performing a genome-wide search for cancer driver mutations

For each cancer, the background mutation rate parameters were estimated across the genome using 5-fold cross validation in 10kb, 25kb and 50kb regions. While the model is robust to choice of 10kb, 25kb or 50kb region size (fig. C-2), the 25kb and 50kb models include some additional regions of the genome due to the the mappability threshold (see section C.1.1). To analyze the largest possible subset of the genome, we performed our analysis iteratively: we first searched for drivers using regions accessible via the 10kb model; we then searched additional regions not accessible by the 10kb model in the 25kb model and then in the 50kb model. To search for drivers, we applied our probabilistic model to estimate the mutation count distributions in 50bp regions across the genome, and we then searched for 50bp regions with significantly more observed mutations than expected under the null distribution of our model. We controlled false-discovery rate at the 0.05 level using a Bonferroni-corrected p-value threshold of $P<1e\text{-}9$.

To compare our hits with known cancer drivers, we tabulated the recurrent driver mutations reported by PCAWG that were present in our dataset, including in the *TERT promoter*, a well known non-coding driver. While most recurrent driver mutations are activating mutations (e.g. cause a gain of cellular function), we also found recurrent mutations in the tumor suppressor genes *TP53* and *SMAD4*. Recurrent mutations in a single position are far less likely in tumor suppressor genes because any deleterious mutation can act as a potential cancer-causing mutation. For example, *TP53* had 6 genome-wide significant 50bp regions, consistent with its status as a crucial tumor suppressor that can be knocked-out with many different mutations (see

table C.1). Methods specialized to discover driver genes are necessary to find tumor suppressor genes in general [151, 187, 169].

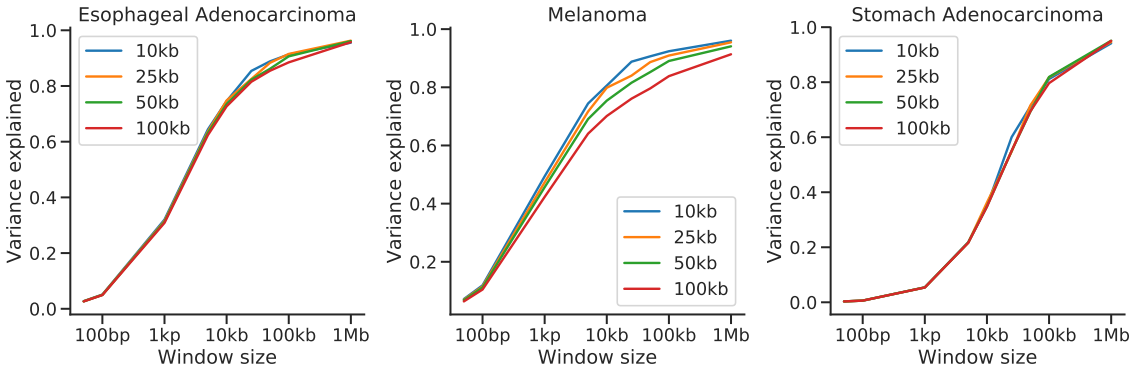

Figure C-2: Model robustness to region size. We tested the robustness of our GM estimates to the choice of the scale of region R over which μ_R and σ^2_R were inferred with the CNN+GP. Here we show our GM's Pearson R^2 accuracy to the observed number of mutations over a range of sizes for different choices of initial region size S. Melanoma shows a slight decrease in performance at larger scales, suggesting local chromatin structure more strongly influences mutation rates in this cancer.

C.1.7 Environment and compute time

A benchmark run at 10kb scale with 10 GP reruns takes 2-3 hours on a single 24 Gb Nvidia RTX GPU, with 8 CPU cores and 756GB RAM. Thus, a full 5-fold of the entire genome takes 10-15 hours. Due to the model's robustness to scale, this time may be significantly reduced without drastic loss of accuracy by using larger region scales (e.g. only 30-40 minutes for 50Kb regions, fig. C-2). Importantly, after completing the CNN+GP training, projections to lower or higher scales via the GM require no additional training.

C.2 Supplementary Results

C.2.1 Negative Binomial Regression does not detect well-known drivers genome-wide

Negative binomial regression is the only other method that has been used to perform an unbiased genome-wide search for driver mutations [191, 214]. We thus evaluated how the sensitivity of NBR to detect driver mutations genome-wide compared with the sensitivity of our method. While all known melanoma drivers present in >3 samples were found by the GM by projecting down to only 1kb scale, NBR at 1kb fails to detect *TERT*, the only known common non-coding driver mutation, yielding a p-value that was an order of magnitude less significant than the genome-wide significance for this scale. Similarly, while the GM detects all known esophageal adenocarcinoma drivers by projecting down to 100bp, NBR over 100bp fails to detect *KRAS*, an important genic driver of esophageal cancers, again yielding a p-value that was an order of magnitude less significant than the genome-wide significance threshold for 100bp. Note: we presented results at 50bp in the text to highlight our model's ability to search in arbitrarily small regions, but all known drivers for esophageal adenocarcinoma are also detected in a search over regions of 100bp.

C.2.2 Convolutional neural network outperforms other dimensionality reduction alternatives for a Gaussian process

We first evaluated the methods for regional rate first and second moment inference, μ_R and σ_R^2, using our simulated datasets. We calculated accuracy as the Pearson R^2 of the estimated mean and variance to the simulated ground-truth mean and variance. CNN+GP, FCNN+GP, NBR and RF accurately inferred μ_R, with $R^2_{\mu_R} > 0.95$ for all three datasets (fig. C-3a). However, PCA+GP, UMAP+GP, and AE+GP consistently under-performed (fig. C-3a left), suggesting supervision when creating feature vectors is critical for the GP downstream performance.

The CNN+GP and FCNN+GP outperformed the other models when estimating

the simulated variance (fig. C-3a, right), suggesting the ability to represent arbitrary functions is important for learning uncertainty in a complex dataset. This conclusion is strengthened by the observation that UMAP and AE enabled relatively accurate variance estimation despite mediocre performance over the mean. Importantly, the clusters used for the simulated data were computed from mean epigenetic vectors; thus our CNN architecture (receiving an input in matrix form) was at a disadvantage. Nonetheless, the CNN+GP most accurately learned both μ_R and σ_R^2 across all three simulated datasets (Fig. C-3a), with slight improvement over the FCNN+GP.

To further compare the approaches, we applied the GP coupled models to estimate real mutation counts from the three cancers on multiple scales. Models were compared by their R^2 to the observed mutations over the test set and to an empirical variance based on the model's own feature vectors (fig. C-3b) (see Appendix). The CNN+GP outperformed the FCNN+GP model over observed mutation counts and empirical variance estimation for all three cancer types. Additionally, the performance advantage of the CNN appeared to grow as window size and observed mutation counts increased. This suggests that local epigenetic patterns play an appreciable role in setting mutational processes and indicates that our model is well-designed to leverage the recent growth in genomics corpus sizes.

C.2.3 Existing whole-genome regression models are time inefficient at multi-resolution search

All existing regression models (RF, NBR, BR) require retraining for each desired scale. A requirement that becomes computationally challenging at finer resolutions (e.g. >1.5h for NBR at 100bp). To provide an estimate of the differences between existing methods and our SPG, we performed a multi-scale time analysis presented in . However, it does not include scales <100bp, such as 50bp used in this work to detect driver hot-spots. A log-log transform of the scale against the run-time () exposes a polynomial relation between the the window size and time (for small enough scales where the compute power is not governed by the machine's memory

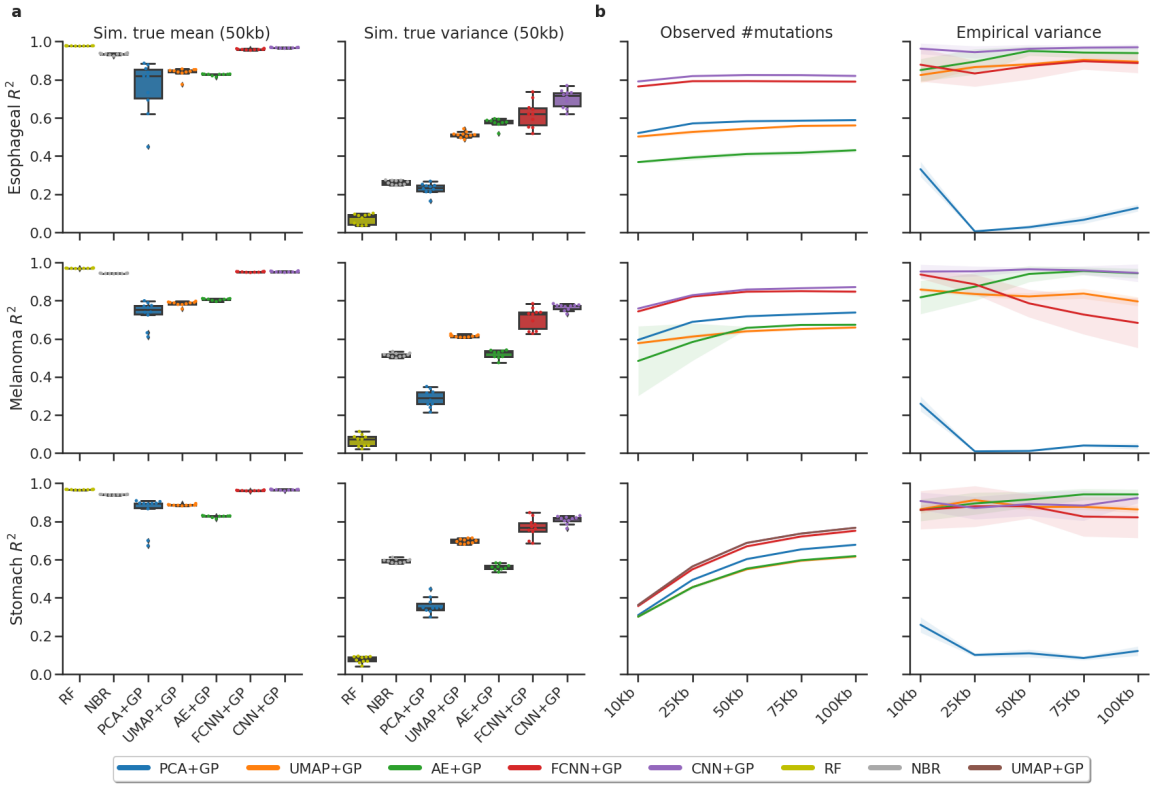

Figure C-3: μ_R and σ^2_R estimation accuracy over three cancer types: esophageal adenocarcinoma (top row), skin melanoma (middle row), and stomach adenocarcinoma (bottom row). **a.** R^2 accuracy of all models with respect to simulated μ_R (left) and σ^2_R (right) at 50kb. **b.** R^2 accuracy of GP-based models to observed number of mutations (left) and empirical variance (right) across scales in real data.

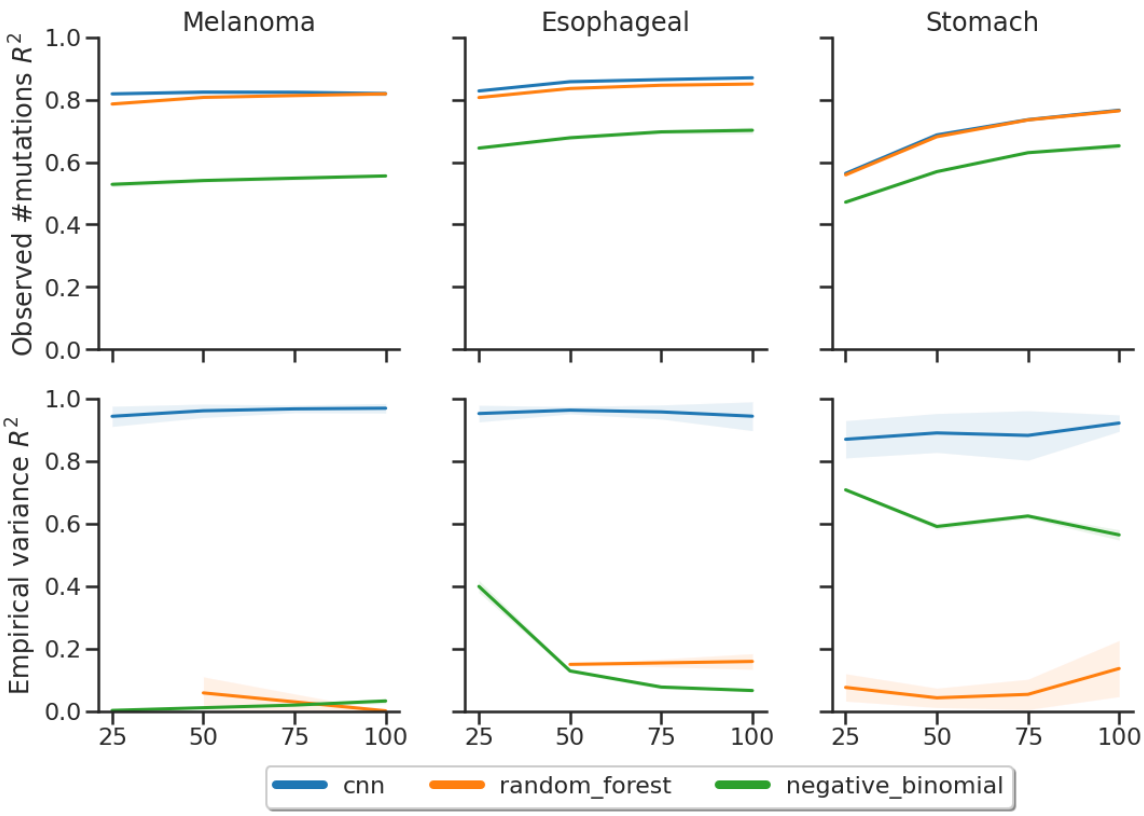

Figure C-4: NBR and RF R^2 accuracy (with CNN+GP as a reference) to observed number of mutations (top row) and empirical variance (bottom row) in real data and across multiple scales for each cancer type: melanoma (left), esophageal adenocarcinoma (middle) and stomach adenocarcinoma (right). Due to the Jackknife method requirement that the number of RF estimators be linear with respect to the number of samples, estimating RF variance at scale <50kb was computationally infeasible (with >8,000 estimators).

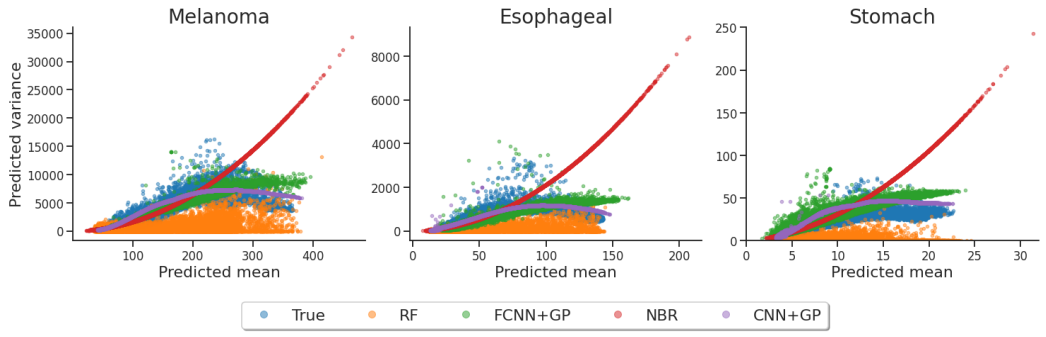

Figure C-5: Mean (μ_R) vs variance (σ^2_R) at 50kb for the ground-truth simulated data (blue) and predictions for each model across all cancer types: melanoma (left), esophageal adenocarcinoma (middle), stomach adenocarcinoma (right). NBR significantly over estimates σ^2_R in high mutation count regions because of its strict quadratic relation to the predicted mean. RF consistently underestimates σ^2_R. FCNN+GP is accurate in low to medium mutation count windows, but overestimates σ^2_R with respect to the CNN+GP in high mutation count regions.

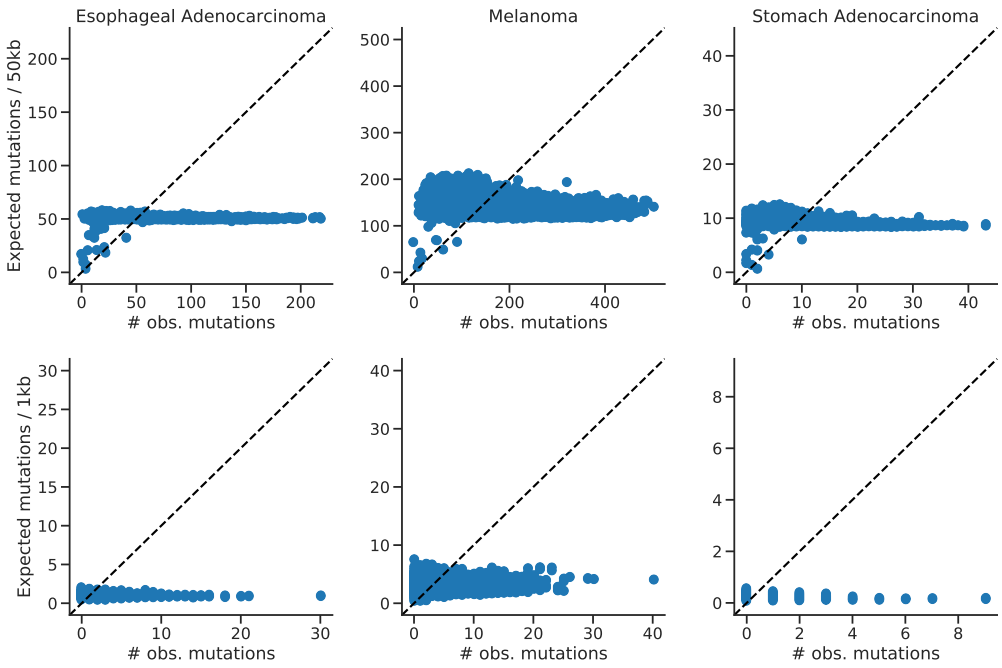

Figure C-6: Number of observed mutations versus number of expected mutations based on the sequence context model alone in 50kb and 1kb regions with mappability >70% across the three cancers. Sequence context explains <10% of variance at 50kb and <1% of variance at 1kb scales for all cancers.

and system operations). Extending this relation to a scale as small as 50bp run-time is as high as 1.5h for BR and 2.5h for NBR. Making the overall run-time for a typical multi-resolution scan of 50bp, 100bp, 500bp, 1kb, 10kb over 2h for BR and over 4h for NBR, while the SPG run-time remains under 1h.

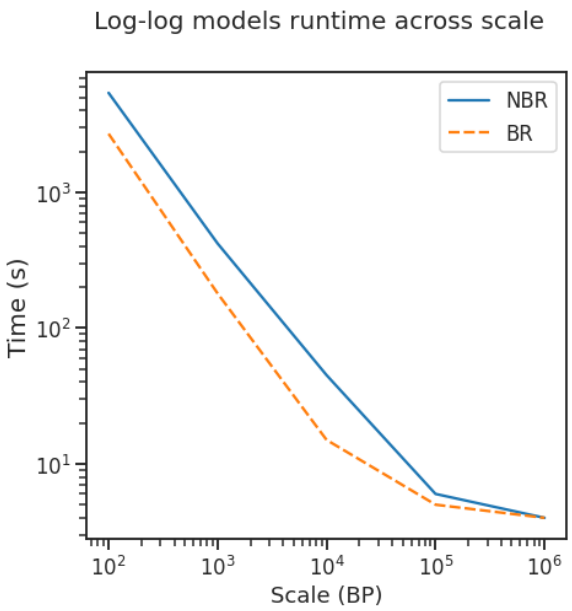

Figure C-7: Log-log run-time of current whole-genome regression methods that require retraining per desired scale. Run-time increases polynomially with scale beyond a threshold (where memory operations dominate the computation of the method).

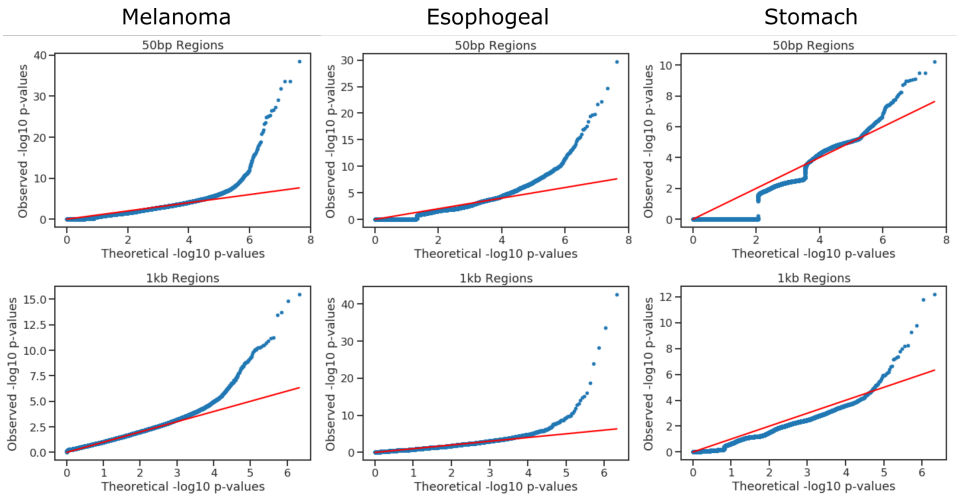

Figure C-8: -log_{10}(P-value) quantile-quantile (qq) plots for expected vs observed number of mutations in 50bp and 1kb windows using our graphical model with rate parameters estimated in 10kb regions for each cancer. Under a properly calibrated null model, p-values generated from the null distribution are uniformly distributed between zero and one. QQ plots thus provide a qualitative assessment of the accuracy of a model's null distribution: the observed p-values should closely match the expected p-values from a uniform distribution (red line) except at extremely small p-values where observations from the alternate model should be found. The step-like nature of the qq-plot for stomach adenocarcinoma in 50bp regions is because the null distribution is discrete (negative binomial) and the dataset has relatively few mutations; thus each 50bp bin can have only one of a few possible mutation counts (typically between 0 and 5).

Chrom	*Start*	*End*	*Observed*	*Expected*	*p-value*
17	7577500	7577549	13	0.0141	$\mathbf{1.75 \times 10^{-30}}$
17	7577100	7577149	10	0.0153	$\mathbf{7.43 \times 10^{-23}}$
17	7577550	7577599	8	0.00856	$\mathbf{3.72 \times 10^{-20}}$
17	7578400	7578449	8	0.0147	$\mathbf{2.76 \times 10^{-18}}$
17	7578500	7578549	6	0.0129	$\mathbf{6.16 \times 10^{-14}}$
17	7578200	7578249	6	0.0146	$\mathbf{1.28 \times 10^{-13}}$

Table C.1: All 50bp windows with significant recurrent mutations in the *TP53* gene from genome wide driver search in esophageal adenocarcinoma.

Appendix D

Supplementary information related to Chapter 6

D.1 Supplementary Results

D.1.1 Insights into mutation rate prediction accuracy from feature maps

To gain insight into which specific epigenetic features the deep-learning model utilized to achieve its high prediction accuracy over mutation counts, we leveraged an approach that highlights input features important to the model's performance (feature maps, Supplementary Methods appendix D.2). Averaging chromatin marks of the same type (e.g., H3K27ac) across tissues revealed that the network learned to focus on localized epigenetic structures (avg. size 1526 bp; 95% CI: 1512-1540 bp) corresponding to known functional elements: transcription start sites, regions of active transcription, enhancers, repressive regulatory states, and heterochromatin to make predictions within kilobase-scale regions (Extended Data fig. 6-6). This behavior was consistent across numerous cancers (Extended Data fig. 6-6). The functional epigenetic structures that the network learned to recognize associated with observed somatic mutation rates in ways consistent with known epigenetic correlates of mutation rates [245] (Extended Data fig. 6-6). For example, regions of closed chromatin

exhibited high mutation rates while those of active transcription exhibited relatively low mutation rates. These results add to the growing evidence that deep-learning models can implicitly learn biological structure when trained to directly predict function from sequence [59, 15, 122].

D.1.2 Comparison of cancer driver detection methods

Because our approach identifies driver candidates by testing for selection, we compared its accuracy to other methods that also test for selection. We first compared our method's ability to identify driver genes in the PCAWG dataset against MutSigCV [151] and dNdScv [169], two widely used methods created specifically to identify genes under positive selection. Following previous works [71, 236], we used the Cancer Gene Census (CGC) [249] as a conservative approximation of the true-positive rate and found our method matched or exceeded the F1-score (a joint measure of sensitivity and specificity) of the other methods in 24 of 32 PCAWG cohorts (excluding hematological and skin malignancies [236]) (uniquely highest score in 13 cohorts; tied for highest in 11 cohorts) (Supplementary fig. D-3, Supplementary table D.6). We additionally calculated the receiver-operator curves for the top 600 genes identified by each method in the PCAWG pan-cancer cohort and found Dig systematically identified more true-positive drivers and fewer false-positives than the other methods (Supplementary fig. D-3), a pattern that we also observed when we additionally compared the methods across whole-exome sequenced (WES) cohorts [71] (Supplementary fig. D-4). We additionally found that Dig's ability to accurately recall noncoding drivers previously identified in the PCAWG dataset was comparable to that of three other burden-based non-coding driver detection methods, Larva [161], ActiveDriverWGS [280], and DriverPower [236] (Supplementary fig. D-5, Supplementary table D.8, table D.9, table D.10, table D.11), although this analysis was biased against Dig because the other three methods were used to generate PCAWG's own set of noncoding drivers.

D.1.3 Additional details on alternative splicing analysis with LeafCutter

Of the eight predicted cryptic splice SNV carriers for which we obtained RNA-seq data (Methods section 6.5), two carriers were discarded due to insufficient coverage either at the gene of interest (DO222305, median coverage of *CIITA* of 17 reads) or globally (DO9074, median depth of coverage of 33). Of the remaining 6 carriers, 4 had clear evidence of alternative splicing: LeafCutter [155] reported a splicing cluster containing the predicted splice SNV with significantly different usage ($P < 0.05$) between the carrier and at least a majority (4 of 6) of the control pairs (Supplementary table D.14). We further investigated the remaining 2 predicted cryptic splice SNV carriers and observed that one had some evidence of alternative splicing in the raw junction file. This carrier (DO52675) had evidence of differential splicing that was not reported by LeafCutter. Specifically, by manually annotating the junction files produced by Regtools [65] with the introns defined in ENSEMBL, we observed that the carrier used an alternative site consistent with the predicted splice SNV in approximately 10% of transcripts, while the controls utilized this site in approximately 1% of transcripts. The remaining carrier (DO33392) sample did not have evidence of alternative splicing upon manual review. This may be due to the mis-spliced transcripts undergoing nonsense mediated decay; however, we did not have statistical power to evaluate this hypothesis.

D.1.4 Investigation of mutational burden in *ELF3* 5' UTR

The PCAWG consortium previously carefully reviewed noncoding mutational hotspots in the PCAWG dataset [214] and cataloged several reasons for excess mutations that were unrelated to positive selection: activation-induced cytidine deaminase (AID) activity in lymphomas, impaired nucleotide excision repair (NER) at transcription factor binding sites in melanomas, activity of endogenous apolipoprotein B mRNA-editing enzyme catalytic subunit (APOBEC) family deaminases, particularly in the in the loop region of predicted hairpin structures, and systematic short-read mapping

inaccuracies leading to artefactual mutation calls. We examined whether any of these processes could be responsible for the observed enrichment of SNVs in the 5' UTR of *ELF3*.

In our analysis of the 5' UTR of *ELF3*, we specifically excluded hematopoietic tumors and melanomas, so neither AID nor NER likely account for the observed elevated mutation rate. To investigate the possible role of APOBEC at the 5' UTR of *ELF3*, we obtained the results of the ABOPEC analysis performed by the PCAWG consortium in which each observed mutation was annotated for whether it could be attributed to APOBEC. Of the six SNVs observed in the *ELF3* 5' UTR, only one was annotated as occurring in a context targeted by APOBEC; however, the sample in which that mutation occurred was not significantly enriched for APOBEC mutations of that kind nor did the mutation occur within a cluster as would be expected if it were due to APOBEC mutagenesis. We thus do not believe APOBEC likely explains the mutational excess in the EFL3 5' UTR. We next examined the gnomAD database [140] which both cataloged population polymorphic germline genetic variation and noted regions of the genome where mapping artefacts were present. The 5' UTR of *ELF3* was not annotated as a region with mapping artefacts by gnomAD. Moreover, of the 16 somatic mutations observed in the PCAWG and Hartwig datasets, only one affected a position also affected by a germline SNP (the canonical splice site chr1:201979836, although the mutation itself is different). The germline SNP was rare (2 alleles observed in >30000 haplotypes). Moreover, the six mutations in the PCAWG dataset were observed in five different cancer types and the ten mutations in the Hartwig dataset were observed in seven different cancer types. Thus, the enrichment cannot be attributed to a mutational process specific to one cancer type. Finally, the mutation enrichment was specific to the canonical 5' UTR of *ELF3*; enrichment was not observed in surrounding regions as was noted by PCAWG for several lncRNAs. In summary, we were unable to explain the mutation burden observed in the 5' UTR of *ELF3* by processes that had been previously noted to increase mutation rate independent of positive selection.

D.1.5 Functional correlates of mutations in rare driver genes

We investigated the functional consequences of rare mutations in three genes with known phenotypes when they act as common drivers: *MSH2* (CNS tumors), *MLH1* (CNS tumors), and *SF3B1* (liver tumors). *MSH2* and *MLH1* encode DNA mismatch repair proteins25; inactivation of these genes increases the spontaneous mutation rate in cells [148]. Thus, carriers of pLoF mutations in these genes are expected to have elevated mutation rates compared to non-carriers. Consistent with this expectation, CNS tumors with rare pLoF mutations in both *MSH2* and *MLH1* exhibited significantly increased mutation rates relative to non-carriers across 213 targeted sequenced genes (*MSH2*: mean 30.1 mutations in carriers vs. 3.0 in non-carriers, $P = 3.8 \times 10^{-7}$ one-sided Mann-Whitney U-test; *MLH1*: mean 35.3 mutations in carriers vs. 3.1 in non-carriers, $P = 8.8 \times 10^{-6}$ one-sided Mann-Whitney U-test). Further supporting the potential driver role of *MSH2* in CNS tumors, the gene also exhibited a significant burden of missense mutations (18 observed vs. 5.3 expected, $P = 2.5 \times 10^{-5}$), and missense *MSH2* carriers also exhibited a significantly elevated mutation rate (mean 35.4 mutations in carriers vs. 3.0 in non-carriers across 213 targeted sequenced genes; $P = 3.7 \times 10^{-12}$, one-sided Mann-Whitney U-test). The mutation rate between pLoF and missense *MSH2* carriers was not statistically distinguishable (P=0.27). *MLH1* did not carry a significant burden of missense mutations in CNS tumors, though this may reflect a lack of statistical power.

SF3B1 encodes a protein involved in the splicing of pre-mRNA molecules. Activating mutations in this gene have previously been associated with increased rates of alternative 3' splice site usage and exon-skipping events [138]. One liver tumor with a rare activating mutation in *SF3B1* had been characterized with RNA-seq. Based on a quantitative accounting of the alternative splicing events in this sample from Kahles et al. [138], the carrier was in the 89th percentile for number of alternative 3' splice events amongst TCGA liver samples (40th of 368 samples) and in the 88th percentile for exon skipping events (43rd of 368 samples), exhibiting more than a standard deviation increase in both types of events relative to the mean across liver

samples. More samples are required to achieve the statistical power necessary to conclude that SF3B1 activating mutations in tumors in which *SF3B1* is rarely mutated alter splicing systematically.

D.1.6 Preliminary analysis of enhancer networks

An analysis of the SNV and indel burden in enhancers (obtained from Nasser et al. [190]) of 725 CGC genes using Dig with default settings revealed 36 enhancers with significant (FDR<0.1) mutational burdens. To coarsely filter regions potentially affected by unmodeled local hypermutation processes, we required that observed mutations each occur in a unique sample. This filter reduced the number of enhancers to ten (Supplementary table D.23). Two enhancers (for *LEPROTL1* and *SRGAP3*) contained recurrent mutations (LEPROTL1: 8:29952919-G>A (n=7), 8:29952921-C>A,G,T (n=5); *SRGAP3*: 3: 8486222-G>C,T (n=6)); however, it is possible that these mutational hotspots could result from APOBEC mutagenesis or mapping artefact [214]. Carriers of mutations in several enhancers demonstrated significant ($P < 0.05$) or nearly-significant ($P < 0.1$) differences in expression compared to non-carriers (not corrected for multiple hypothesis testing). For example, carriers of mutations in the *NCOR2* enhancer (12:125422682-125425761) had a nearly significant decrease in expression ($P = 0.078$). However, expression did not always change in a direction consistent with the known or predicted function of the gene in tumorigenesis. For example, carriers of indels in the *MSI2* enhancer (17:54992281-54993673) had decreased *MSI2* expression ($P = 0.0081$) based on carrier tumors from kidney, rectum, and ovary; however, *MSI2* is a known oncogene in hematopoietic cancers. More follow-up analysis will be necessary to determine whether the mutational enrichment constitutes positive selection or unaccounted for neutral mutational processes.

D.2 Supplementary Methods

D.2.1 Technical details of Dig's deep-learning framework

Deep-learning network architecture

Convolutional neural network The CNN architecture is as follows: it contains 4 convolutional blocks with 2 batch normalized convolutional layers and ReLU activation. The first block reduces the 735×100 input tensor to 256×50 with 256 channels and a double stride. The following blocks are ResNet-style residual blocks which maintain their input dimension to facilitate residual connections with 256, 512, and 1024 channels respectively. Between each of the 3 residual blocks there is a double stride (ReLU activated and batch normalized) convolutional layer, which reduces the tensor length by half and doubles its height with additional channels. The output of the last residual block is flattened (and optionally concatenated with the two-flanking region counts) and passed through 3 fully connected (FC) layers. The first two FC layers are ReLU activated and reduce the dimensionality of the vector to 128 and 16 dimensions respectively. The last FC layer performs the final regression that predicts the SNV count in the 10kb region via a linear function. The CNN architecture was implemented in PyTorch [199].

Gaussian process The Gaussian process is a sparse, inducing-point GP [250] with a radial basis function kernel that takes as input the final 16-dimensional feature vector of the trained CNN and non-linearly predicts both the mean and variance of the neutral mutations in the associated 10kb region. The GP architecture was implemented in GPyTorch [98].

Deep-learning model training

Filtering of 10kb regions To avoid training the model over regions with inaccurate mutation counts due to technical noise, we removed regions likely to contain spurious mutation counts, defined as windows where less than 50% of the 36mers

uniquely mapped back to that region or regions in the top 99.99th percentile of mutation counts.

Model training The CNN and GP were trained sequentially. First, the CNN was trained for 20 epochs with a batch size of 128 samples, using the Adam optimizer to minimize mean squared error loss to the observed mutation counts in each training window. For training, the input data was additionally divided via an 80-20 split into training data and validation data (thus for each fold, 64% of the genome was used for training, 16% for validation, and 20% for held-out prediction). To avoid overfitting the data, the epoch from which the trained CNN was selected was determined by highest validation R-squared accuracy to observed counts in the validation-set across all CNN epochs. Once the CNN was trained, the final 16-dimension feature vector for each training window was passed as input to the Gaussian process which was trained to predict the observed mutation counts in each training window by minimizing a multivariate normal loss function with the Adam optimizer. The GP was optimized with 400 inducing points for 50 iterations. Due to the inherent variability in gradient-based optimization, we ran the GP five independent times and calculated the ensemble average of the mean and variance predictions from each of the individual runs on the held-out set of regions. These ensemble predictions were then used as the mean and variance estimates for each 10kb region. For each fold, we also predicted mean and variance of mutation counts in windows filtered prior to training. The ensemble average across all GP runs and all folds were used as the mean and variance estimates for these regions.

Some random initializations of the GP would fail to converge (defined as a decrease in R-squared accuracy of more than 0.03 compared to the final accuracy of the trained CNN). When this occurred, the GP was restarted up to 3 times to achieve a successful convergence. If after 3 attempts, the GP had not successfully converged, the number of inducing points was reduced by 100 and the GP given another 3 attempts to converge. This process continued until successful convergence or a reduction to zero inducing points. If a GP failed to converge in all 12 attempts, the CNN was

reinitialized to generate a new set of feature vectors.

D.2.2 Technical details of Dig's probabilistic graphical model

We derived a probabilistic method to estimate a distribution over the number of SNVs and indels observed at a set of positions in a dataset of interest given the kilobase-scale estimated mutation rate μ_R and estimation uncertainty σ_R^2 along with the sequence context likelihood estimates. We refer to this method as Dig.

Passenger model for indels and multi-nucleotide variants

The indel model is identical to that of the SNV model with two exceptions. First, we assume a uniform distribution of indels independent of sequence context, as has been assumed in previous works [169]. Thus in the negative binomial distribution above, $\sum_I p_{R,aX\to Yb}$ is replaced by the uniform mutation probability $|I|/|R|$ where $|\cdot|$ denotes the total number of genomic positions in I and R. The uniform assumption of indels could readily be replaced with a probability distribution based on indel type, size and homology [10], but we do not pursue that extension here. Second, the scaling factor for indels, C_{indel}, is estimated as the ratio of the number of indels observed in the target dataset to the number of expected indels in the training dataset across the coding sequence of all genes not in the Cancer Gene Census. We treat multi-nucleotide variants (MNVs) as indels.

We tested estimating μ_R and σ_R^2 independently for SNVs and indels using separate deep learning models for the two types of mutations. We found that direct estimation of these parameters for indels resulted in a less accurate indel model than using the SNV estimates as a proxy for indel estimates. We suspect this is due to the fact that indels occur an order of magnitude less frequently than SNVs and thus there are too few observed indels in the training cohort for the deep-learning model to build an accurate prediction function. As sample sizes become larger, we expect that directly training a deep-learning model to predict indels will yield more accurate predictions.

Extension to mutations spanning multiple kilobase-scale regions

We take two approaches to extend the above passenger models to account for sets of mutations that span multiple kilobase-scale regions.

- Approach 1: approximate the distribution across the regions by extending the variational estimation of α_R and θ_R. Specifically, let $R' = \{R_1, \ldots, R_n\}$ be the set of regions in which a set of mutations occur. Then we estimate $\mu_{R'} = \sum_{i=1}^{n} \mu_R$ and $\sigma^2_{R'} = \sum_{i=1}^{n} \sigma^2_R$, and $\alpha_{R'}$ and $\theta_{R'}$ are then estimated as above from $\mu_{R'}$ and $\sigma^2_{R'}$.

- Approach 2: exactly estimate the distribution across the mutation set by convolving the distributions arising from the subset of mutations in each $R_i \in R'$.

Approach 1 is computationally efficient and accurate so long as the mutation rate estimates across $\{R_1, \ldots, R_n\}$ are sufficiently similar. Thus approach 1 is preferred when $R\prime$ is composed of a small number of contiguous (or nearly contiguous) regions and is the default implemented algorithm. When $R\prime$ is composed of regions with highly variable mutation rates, approach 1 is likely to either over- or under-estimate the passenger mutation rate, leading to improperly calibrated p-values. In this case, approach 2 will provide accurate estimates but requires more computation due to the convolution operation.

Testing mutational burden across a set of candidate mutations using an existing mutation map

The steps to estimate selection using Dig are as follows:

User steps:

1. Download a mutation map for the cancer matching the cancer type of the dataset of interest.

2. Provide the mutation dataset of interest and define the set I of possible mutations. I can be defined as any set of genomic intervals (contiguous or noncontiguous) or any set of possible SNVs anywhere in the genome.

Software steps:

1. The mutation likelihoods $p_{R,\ aX\to Y,b}$ and p_{indel} are calculated as described above for each mutation set. The nucleotide sequence of R is extracted from the reference genome.

2. The SNV and indel scaling factors are estimated for the cohort of interest

3. The p-value of the number of SNVs and indels observed in the cohort of interest for each mutation set are calculated using the negative binomial distributions defined above as the null models. In this work, we calculated the P-value as the upper-tail probability of the observed mutation count, applying a mid-P correction to account for the discrete data.

4. The p-values for the SNVs and indels are combined via Fisher's method.

For this study, we used the mutation maps trained using both epigenetic tracks and flanking mutation counts to test for burdens of mutations. These are also the maps we have made publicly available.

D.2.3 Associating epigenetic structure to mutation density with feature maps

To investigate the underlying features the deep learning model considered when predicting mutation rates, we added another layer of computation between the input epigenetic matrix and the CNN to serve as feature maps. Feature maps are a tool used in computer vision tasks to detect which regions of an image the model uses to perform prediction [239]. We used this technique to evaluate which epigenetic patterns the CNN exploited to predict mutation rates. To reduce the potential for noise, we applied this technique to input matrices encoding 50kb regions.

Feature map generation

An additional two-layered network was added between the input matrix and CNN to force the model to attend to the subset of most salient input sub-regions and compute the feature maps. In the attention augmented CNN, the input matrix was first passed through two convolutional layers preserving the input dimensionality (stride length 1, kernel sizes 5 and then 3) with ReLU activations. Subsequently, the output of the two layers was passed through a row-wise Softmax function that had the effect of making most entries in the matrix close to zero with sparse values close to one. The resulted "feature map" matrix was then element wise multiplied with the original input and passed on to the downstream CNN. This had the effect of setting most entries in the original epigenetic matrix to near zero, thus forcing the CNN to rely only on the small subspace of the input that was not zeroed out. The optimization process compels the feature maps to attend to the features of the input matrix most relevant for the prediction process.

Extraction of epigenetic content of feature maps via dimensionality reduction and clustering

While the feature maps have the theoretical ability to attend to any regions of the input matrix, in practice we found they almost always attended to a large set of epigenomic features (rows) in a small set of contiguous columns (genomic positions), zeroing out most values outside of these columns. We extracted and summarized the epigenetic content of each of these attention columns through the following approach: 1) in each 50kb window, we extracted the largest contiguous set of columns such that each column contained at least 10 cells with a non-zero entry. This contiguous set of columns was defined as an "attention super-column". 2) Each attention super-column was reduced to an 8-dimensional vector by averaging together tracks of the same epigenetic type per column (DNase, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, and H3K9me3) and taking the maximum value across each row. 3) The vectors were normalized and projected into a two-dimensional subspace for

clustering and visualization via a Uniform Manifold Approximation and Projection (UMAP) transformation (Python UMAP-learn package; # neighbors: 30, minimum distance: 0, # components: 2). 4) Spectral clustering (Python Sklearn package) was applied to the two-dimensional subspace to identify attention super-columns with similar epigenetic content. The clustering consistently identified five distinct clusters across cancer cohorts (Extended Data fig. 6-6).

Connecting feature map epigenetic clusters to functional annotations

To determine whether the attention super-columns in a cluster represented a functional epigenetic structure, we extracted the average Epilogos [179] signature vector per attention super-column and examined whether the Epilogos signatures were consistent within a cluster. Epilogos is a summary of the functional epigenetic states across 111 tissue types as inferred by the ChromHMM method.

Connecting feature map epigenetic clusters to mutation rate

We extracted the mutation count for the 50kb window in which each attention column occurred. We computed the mean and standard deviation of mutation counts across the attention column clusters.

D.2.4 Additional details about the comparison of mutation rate models

Deflation of variance explained statistic in low count scenarios

In discrete stochastic systems, random stochasticity of events when event rate is low results in deflation of the variance explained statistic. The characteristic arises because a discrete system generally has a fractional expected value but observations must take on integer values. Thus, even if a model perfectly predicts the expected value, it will explain relatively little variance if the difference between the fractional expected value and possible observed values is of similar magnitude to the possible observations (e.g., expected value of 0.5 versus possible observed values of 0 or 1).

Intuitively, for a discrete process with event rate <1, the expected value will be a real value between zero and one but the observed count will be an integer (0, 1, 2, etc.); thus, the true expected value will explain relatively little variance of observed data because the observed values almost always deviate substantially from the expected value.

Tiled regions

We compared the variance explained (square of the Pearson correlation coefficient) in SNV counts within 10kb windows tiled across the genome between Dig and NBR [191]. NBR is, to our knowledge, the only method that has been previously used to build passenger mutation rate models in kilobase-scale regions tiled across the genome. However, code for running the NBR method is not currently publicly available. For each cancer, the NBR model was trained on the same regions used to train our deep-learning model (excluding regions with 36mer mappability <50% and regions in the top 99.99th percentile of mutation count). The regions excluded from training were also excluded when calculating the variance explained statistic. We also assessed the variance explained of SNV counts in 1Mb regions by our method and NBR (restricted to 1Mb regions with >50% 36mer mappability). To estimate the expected mutation count in each 1Mb region, we summed together the estimates of each non-overlapping 10kb window within the 1Mb region.

Coding sequence

We compared the variance explained in nonsynonymous SNV counts between Dig and two widely used methods that generate nonsynomous SNV passenger mutation models: MutSigCV [151] and dNdScv [169]. Both MutSigCV and dNdScv utilize the synonymous mutations observed in each gene to estimated gene-specific passenger mutation rates. Variance explained was evaluated over the coding sequence of 3,740 genes that were 1) common to all three methods; 2) between 1kb and 1.5kb in length; and 3) not in the CGC. The length restriction was imposed to prevent coding sequence length from artificially inflating variance explained since the number of mutations in

a gene strongly correlates with its length.

Noncoding regulatory elements

We compared the variance explained in SNV counts between Dig and two other methods that estimate passenger mutation rates in noncoding regulatory elements: DriverPower [236] and Larva [161]. DriverPower is optimized to estimate mutation rate within a set of regulatory elements predefined by the authors of the software; this set of elements is not easily changed. We thus evaluated variance explained in a set of 7,412 noncoding regulatory elements (enhancers, lncRNAs, and sncRNAs) between 0.5kb and 1kb in length that could be modeled by DriverPower. The length restriction was again implemented to prevent inflation of variance explained due to variance in element length. While Larva can predict mutation rate within genomic intervals, it cannot natively provide a prediction for elements that are composed of multiple, non-contiguous intervals. To circumvent this, we divided each element evaluated by DriverPower into its constituent intervals, produced a prediction for each interval separately with Larva, and summed the predictions across regions composing a single element.

D.2.5 Details about the comparison of driver element detection methods

Comparison of driver gene detection methods

We compared the sensitivity, specificity, and F1-score (harmonic mean of sensitivity and specificity) for driver gene detection from coding sequence mutations between Dig, MutSigCV, and dNdScv across the 32 PCAWG cancer cohorts (melanomas and hematopoietic cancers were excluded as in previous comparisons [236]). We chose to compare to these two methods because they are widely used driver gene detection methods that rely on neutral mutation models to test for selection. An FDR significance threshold of 0.1 was applied for all methods and cohorts. A true-positive driver gene was defined as any gene in the Cancer Gene Census (CGC) [249] that was

detected as FDR significant by any of the methods in a given cohort. A false-positive was defined as any gene identified as FDR significant that was not in the CGC. Each method was applied to the same set of 16,794 genes. Both SNVs and indels were used to identify potential driver genes. We additionally compared power over the 16 whole-exome sequenced cohorts from Dietlien et al. (excluding hematopoietic cancers as above). The larger cohort sizes enabled the approximation of receiver-operator characteristic curves for the methods. The curves were approximated because genes in the CGC were used as a proxy for true-positives (that is, a gene not in the CGC may still be a true-positive driver but would be counted as a false-positive in this analysis). Because of the approximated nature of these curves, we visualized the results as false-positive counts vs true positive counts rather than the standard false-positive vs true-positive rates, following precedent from Dietlein et al. The power of a method was quantified as the area under these approximated receiver-operator characteristic curves.

Comparison of noncoding driver element detection methods

We compared the sensitivity, specificity, and F1-score for driver noncoding element identification from noncoding SNVs between Dig, DriverPower, Larva, and ActiveDriver-WGS across the 32 PCAWG cancer cohorts (excluding melanoma and hematopoietic cancers as above). We chose to compare to these three methods because they are recently introduced methods for noncoding driver element identification that rely on neutral mutation models to test for selection. An FDR significance threshold of 0.1 was applied for all methods and cohorts. A true-positive driver element was defined as any element previously identified by PCAWG as carrying a burden of mutations [214] that was detected as FDR significant by any of the methods in a given cohort. A false-positive was considered any FDR significant element that was not previously identified by PCAWG as having a burden of mutations. This comparison was conservative (biased against our approach) for two reasons: 1) The other three methods were previously applied to the PCAWG dataset to generate the set of putative driver elements that we then used as a gold standard for the same samples; and 2) we re-

stricted the analysis to SNVs because not all methods we compared to could accept indels. Indeed, our approach is the only approach that models SNVs and indels independently; the other approaches either do not model indels or model indels and SNVs as a single category.

D.2.6 Constructing a genome-browser of genome-wide mutation rate estimates

We used Dig to estimate mutation rates in every non-overlapping regions of size 100bp, 250bp, 500bp, 1kb, 2.5kb, 25kb, 50kb, 100kb, 250kb, 500kb and 1Mb tiled across the genome (excluding assembly gaps in the GRCh37 reference genome) for 37 PCAWG cancer types. These predictions were used to construct data structures that can be interactively visualized by HiGlass [141].

D.2.7 Details about power analysis

We conservatively simulated Dig's power to detect driver SNVs at different carrier frequencies across enhancers and noncoding cryptic splice sites under the pan-cancer mutation map using the following Monte Carlo approach.

For a given sample size and carrier frequency of driver mutations:

1. For each element, randomly draw a mutation rate parameter from the gamma distribution defined by mean and variance estimated by the kilobase-scale model.

2. For each element, estimate the scaling factor as the target sample size divided by the pan-cancer sample size (n=2,279) and randomly draw an observed number of mutations from a Poisson distribution with rate parameter equal to the sampled rate multiplied by the scaling factor and by the probability of an SNV in the element.

3. For each element, randomly sample the number of driver mutations from a Poisson distribution with rate parameter equal to the target sample size multiplied by the carrier frequency.

4. Count the number of elements for which the sum of the background mutations and driver mutations exceeded the Bonferroni-corrected $\alpha < 0.05$ threshold under Dig's negative binomial null mutation distribution for each element. Divide the count by the total number of tested elements to estimate a detection likelihood.

5. Repeat steps 1-4 one thousand times and average the detection likelihoods across all simulations.

D.2.8 Additional details about quantifying selection on cryptic splice SNVs

Monte Carlo method for estimating confidence intervals of mutational enrichment.

Mutation enrichment was defined as the ratio of the observed mutations to expected mutations. We used the following Monte Carlo simulation approach to estimate the 95% confidence intervals of enrichment for a given set of genes and given mutation type.

1. For each gene, estimate the enrichment coefficient as the number of observed mutations divided by the number of expected mutations. A small pseudocount of 1×10^{-16} was added to the numerator and denominator to prevent the enrichment from being identically zero when no mutations were observed in a gene. (This would lead to a degenerate Poisson distribution in step 3). For each gene, randomly draw a Poisson rate parameter from the gamma distribution defined by the mean and standard deviation estimates of the kilobase-scale mutation rate map.

2. For each gene, randomly draw a number of "observed" mutations from a Poisson distribution with rate parameter equal to the simulated rate parameter multiplied by the enrichment coefficient and the likelihood of the mutation type

occurring within the gene. Conceptually, this mutation count is simulated under the hypothesis of positive selection on the mutations within the gene.

3. Estimate a simulated enrichment by summing the number of simulated mutations across all genes in the set and dividing by the expected number of mutations under the null model of no enrichment.

4. Repeat steps 1-4 one thousand times and define the boundaries of the 95% confidence interval as the lower 2.5th percentile and upper 97.5th percentile of the simulated enrichments.

Additional quantification of mutation enrichment in TSGs and oncogenes

To gain additional confidence in the accuracy of our mutation enrichment estimates, we directly compared the mutation rate in genes not in the CGC to TSGs and oncogenes in the CGC using a two-sided Chi-squared test for a two-by-two contingency table. This approach recapitulated the enrichment patterns we observed using Dig. However, the Chi-squared test does not account for global mutation rate differences between genes not in the CGC and genes in the CGC; thus, the precise estimates in Supplementary fig. D-9 are unlikely to be accurate.

Identification of individual TSGs enriched for noncanonical cryptic splice SNVs

In each of the 37 PCAWG cohorts, we identified TSGs in the CGC with a significant burden of noncanonical cryptic splice SNVs under the null model estimated by our method. The significance threshold was defined per cancer as FDR q-value$<$0.1 corrected for the number of tested TSGs (n=283). We excluded one significant gene, *PRDM1*, from further analysis because the observed excess mutations were attributable to a single sample.

Quantification of the pan-cancer contribution of cryptic splice SNVs to TSG driver SNVs

We calculated the excess of SNVs in TSGs in the CGC stratified by function (missense, nonsense, canonical splice, and noncoding canonical splice) as the difference between the number of mutations observed and the number expected. The relative contribution for each category was defined as the excess for that category normalized by the sum of the excess across all categories. The 95% confidence interval for the contribution of each category was calculated using the Monte Carlo approach described above for enrichment with the following modifications:

- In step 3: for each gene, the number of neutral mutations was also simulated from a Poisson distribution with rate parameter equal to the gamma-simulated rate parameter multiplied by the probability of a mutation occurring in the gene. Conceptually, this mutation count is simulated under the hypothesis of neutral selection on the mutations within the gene.

- In step 4: the excess for each gene is calculated as the difference between the number of mutations simulated under positive selection and the number simulated under neutral selection. The total excess for each mutation category is summed across all genes and the relative contribution calculated as above.

Enrichment of predicted splicing impact in noncoding cryptic splice SNVs observed in significantly burdened TSGs

We used a bootstrap method to calculate a p-value for the null hypothesis that non-canonical cryptic splice SNVs observed in the genes with a significant burden of cryptic splice SNVs had a predicted impact on splicing similar to the predicted impact of cryptic splice SNVs observed in genes not in the CGC. We calculated the median of the Δ scores randomly resampled from the observed cryptic splice SNVs in the TSGs and observed cryptic splice SNVs in genes not in the CGC ten thousand times (the number of SNVs sampled from the non-CGC set was equal to the number observed in the TSG set). We estimated the p-value as the number of times the resampled

median of the non-CGC cryptic splice SNVs exceeded the resampled median of the cryptic splice SNVs observed in the TSGs.

Analysis of alternative splicing events in RNA-seq data

We obtained RNA-seq data for 8 samples carrying deep intronic predicted cryptic splice SNVs (i.e., distance to nearest exon boundary >20 base-pairs) in TSGs with a significant burden of predicted noncoding cryptic splice SNVs. This represented all such carriers with available RNA-seq data. We downloaded the STAR aligned BAM files for each donor and six randomly selection non-carriers from the same cancer cohort, and we used bedtools bamtofastq to convert these reads into FASTQ files for de novo alignment. We then ran olego [268] with the default junction database and max edit distance of 4 (flag `-M 4`) on each FASTQ file. Olego is specifically designed for increased sensitivity to de novo splicing in RNA-seq reads. The de novo aligned sam files were then converted to bam files, sorted, indexed, and processed for junctions by Regtools [65] for downstream analysis (input parameters: `-a 8 -m 50 -M 50000`). For each of the carrier-control pairs, we performed differential splicing analysis using LeafCutter as described by Li et al. [155]. The introns in each pair were clustered using the `leafcutter_cluster_regtools.py` script, requiring a single split read to support a junction and assuming a maximum intron length of 500Kb (input flags `-m 1 -o -l 500000`). Differential splicing was then evaluated using the `leafcutter_ds.R` script using the Gencode v19 exons provided with the software. When a gene had more than one transcript available, we used the canonical transcript as annotated in UCSC genome browser. We considered a predicted splice SNV to have strong supporting evidence if LeafCutter reported a splice cluster containing the predicted splice SNV that had significantly different usage between carrier and control ($p < 0.05$) in the majority of the carrier-control pairs. If LeafCutter did not report a cluster containing the predicted splice SNV, we additionally examined the raw junction files from Regtools. We considered a predicted SNV to have some supporting evidence if junctions supporting the prediction were observed in the raw junction files. Two of the eight samples were discarded due to insufficient coverage of

the gene of interest (Supplementary table D.14).

D.3 Supplementary Figures

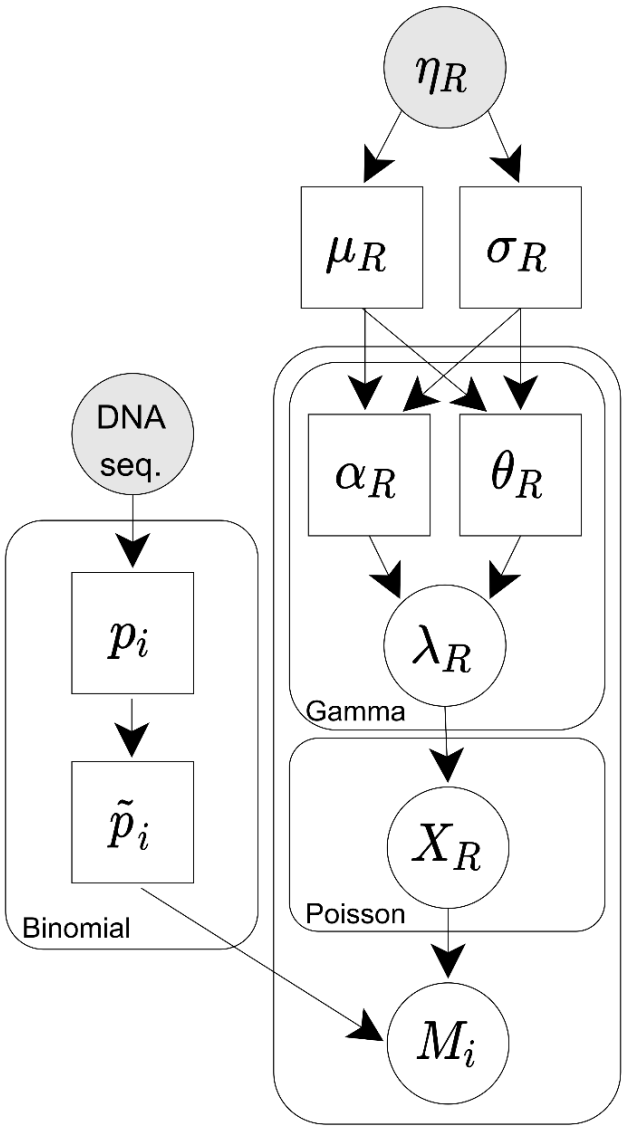

Figure D-1: Plate diagram of the probabilistic model that Dig uses to model the number of neutral mutations (M_i) in an element of interest. η_R: observed data used as input to Dig's deep-learning model (chromatin modifications and, optionally, flanking mutation counts) to estimate regional neutral mutation parameters for region R. μ_R and σ_R: mean and standard deviation estimates of the neutral mutation rate in region R. α_R and θ_R gamma distribution shape and scale parameters, respectively. λ_R gamma-distributed mutation rate parameter for region R. X_R poisson-distributed mutation count in region R. DNA seq.: the DNA sequence from the human reference genome. p_i: genome-wide likelihood of a mutation in a given DNA context centered at position i. $\tilde{p}_i$: likelihood of mutation based on sequence context centered at position i normalized such that $\sum_{i \in R} \widetilde{p}_i = 1$. See Methods for additional details.

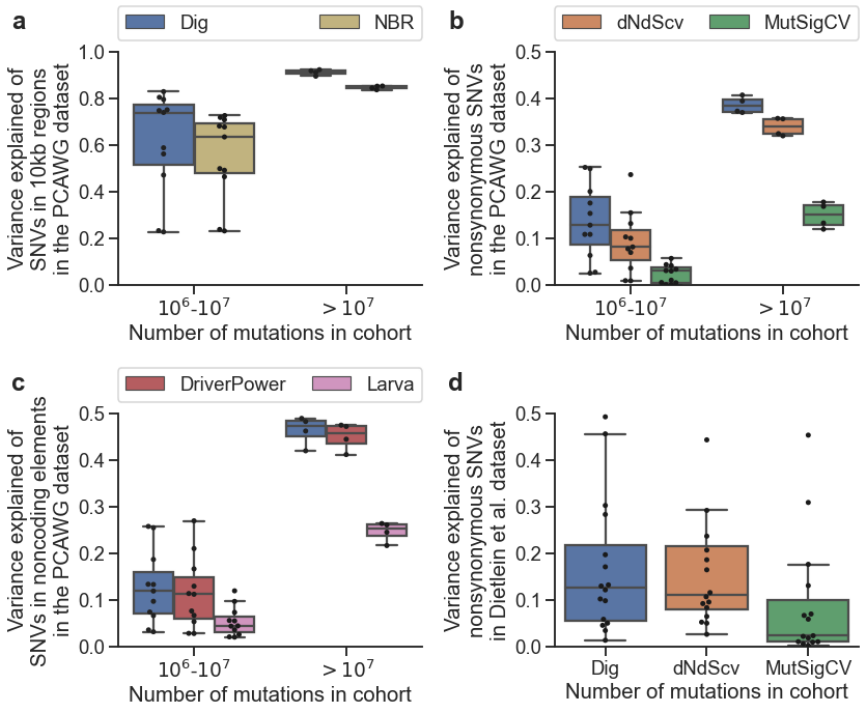

Figure D-2: Comparison of variance explained of SNV counts across methods, annotations, and cohorts. a, Variance explained of SNV count in 10kb regions tiled across the genome by Dig and NBR [191] in N=16 PCAWG cancer cohorts with >1 million SNVs (excluding hemopoietic tumors, for which NBR failed to converge). Regions in which <50% of 36mers are unique are excluded as are regions in the 99.99th percentile of mutation count. b, Variance explained of nonsynonymous SNV count in genes 1-1.5kb in length (n=3,740 genes) in N=16 PCAWG cancer cohorts. c, Variance explained of SNV count in enhancers and noncoding RNAs (long and short) 0.5-1kb in length (n=7,412 noncoding elements) in 16 PCAWG cancer cohorts. d, as b for 16 whole-exome sequenced cancer cohorts from Dietlein et al. [71].

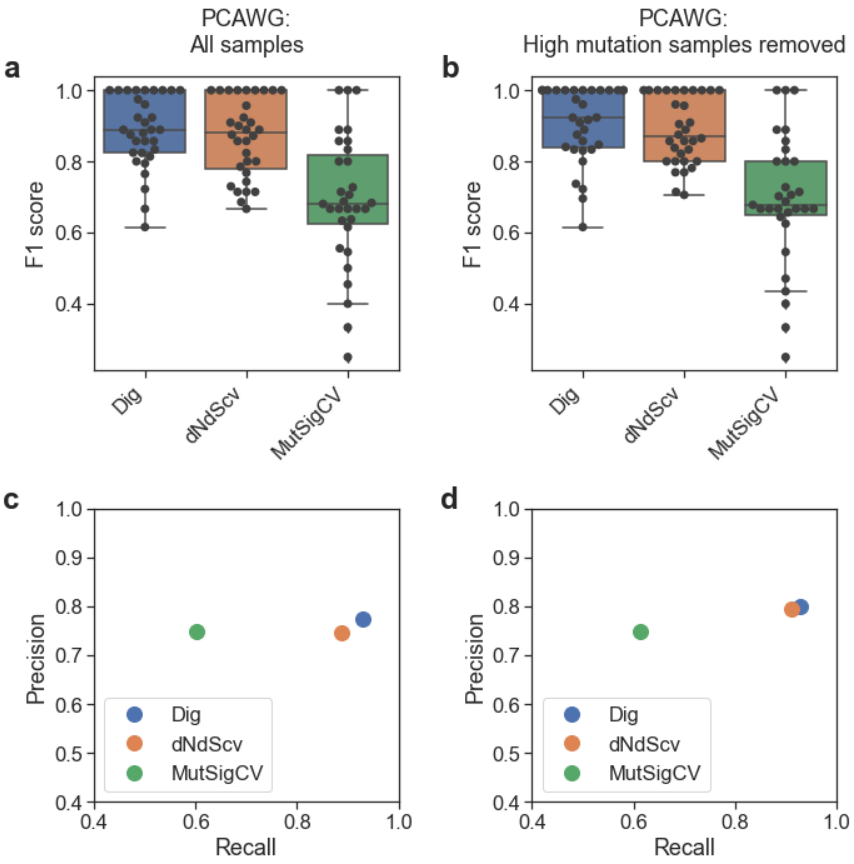

Figure D-3: Precision-recall comparison of gene driver methods in the PCAWG cohort. a,b F1-score (harmonic mean of precision and recall) in N=32 PCAWG cohorts (melanoma and hematopoietic tumors were excluded as in previous work [236]) across 16,794 genes common to the three methods. Precision and recall were calculated using genes in the Cancer Gene Census as a conservative true positive set. a, All samples. b, Excluding samples with >3000 coding mutations and restricting the total number of mutations per sample per gene to 3 (default filtering options for dNdScv). c,d Recall and precision measured across all N=32 PCAWG cohorts for c, all samples and d, samples with <3000 coding mutations.

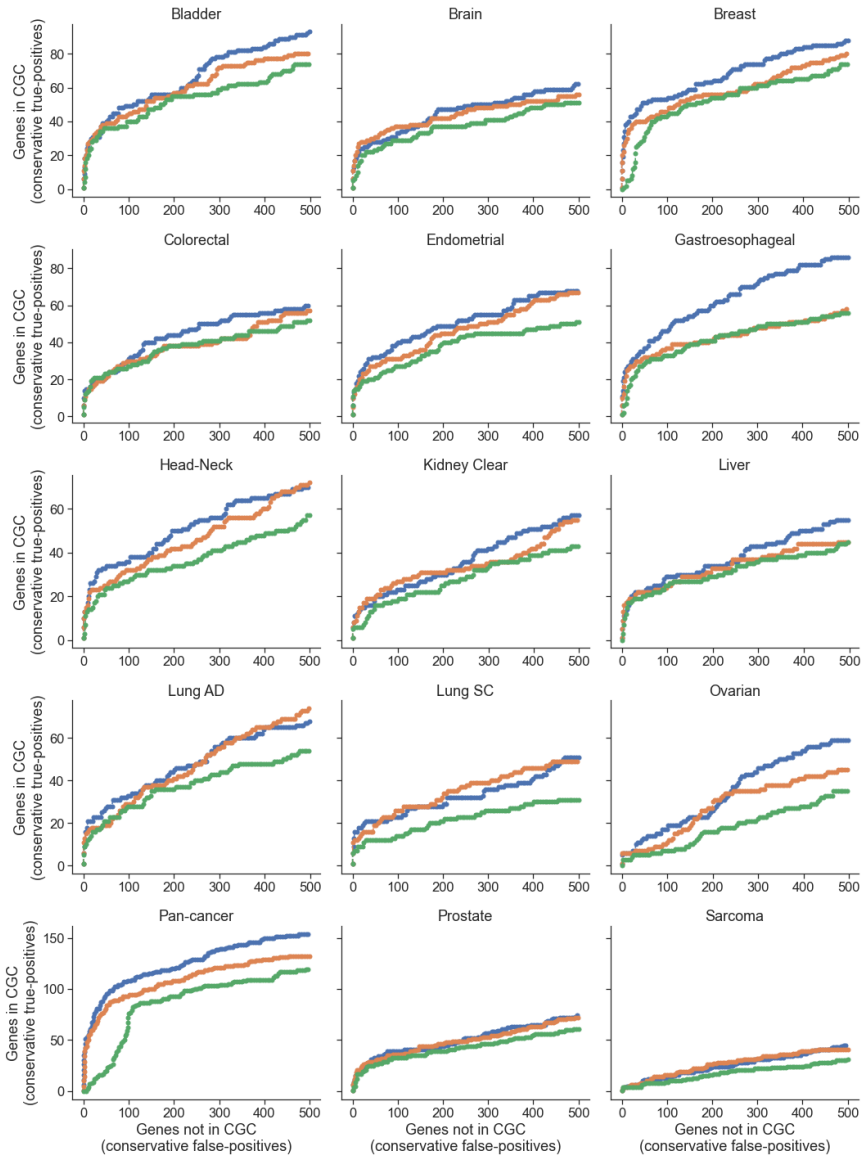

Figure D-4: Approximate number of false-positive and true positive driver genes identified from 15 whole-exome sequenced cohorts from Dietlein et al. [71]. The numbers are approximate because the full set of driver genes is unknown; we therefore used genes in the CGC as a conservative approximation of true positives (since a non-CGC gene may still be a true driver). The MutSigCV model produced mis-calibrated p-values for the pan-cancer cohort, suggesting that its model assumptions may have been violated by the large cohort of heterogeneous cancer types.

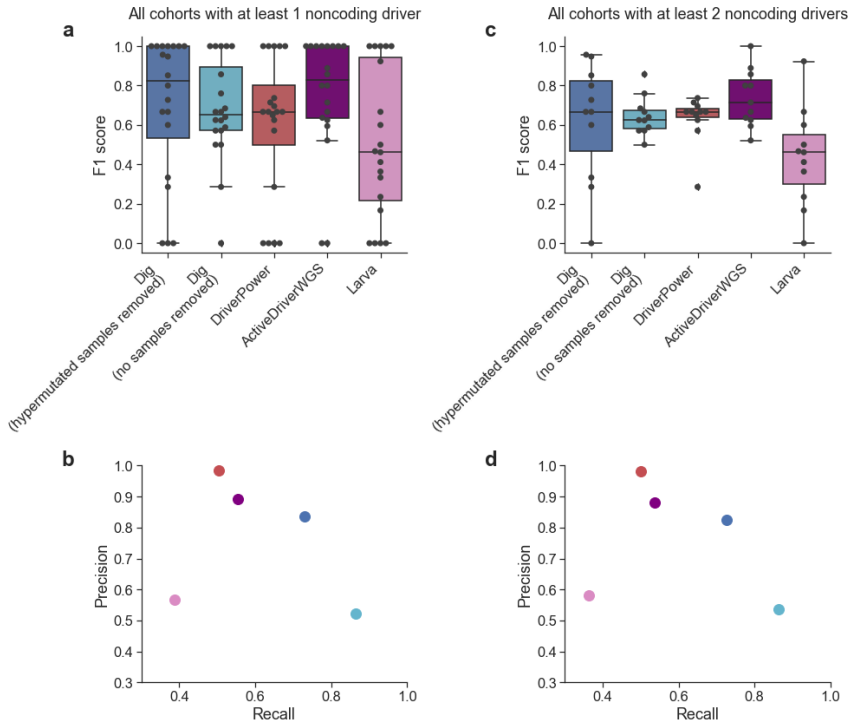

Figure D-5: Precision-recall comparison of noncoding driver detection methods in the PCAWG dataset. a, F1-score across 95,231 noncoding elements as defined in Rheinbay et al. [214] in PCAWG cancer cohorts with at least one identified noncoding driver (n=20 cohorts). The performance of Dig was also evaluated when removing samples with >1000 SNVs across all elements and restricting the total number of SNVs per sample per element to 3. DriverPower and Larva do not have built-in filtering options. ActiveDriverWGS was run with default filtering which removes any sample with >30 SNVs per megabase. b, Recall and precision by method combined across the cohorts in a. c,d, as in a and b but restricting to n=11 cohorts with at least two identified noncoding drivers.

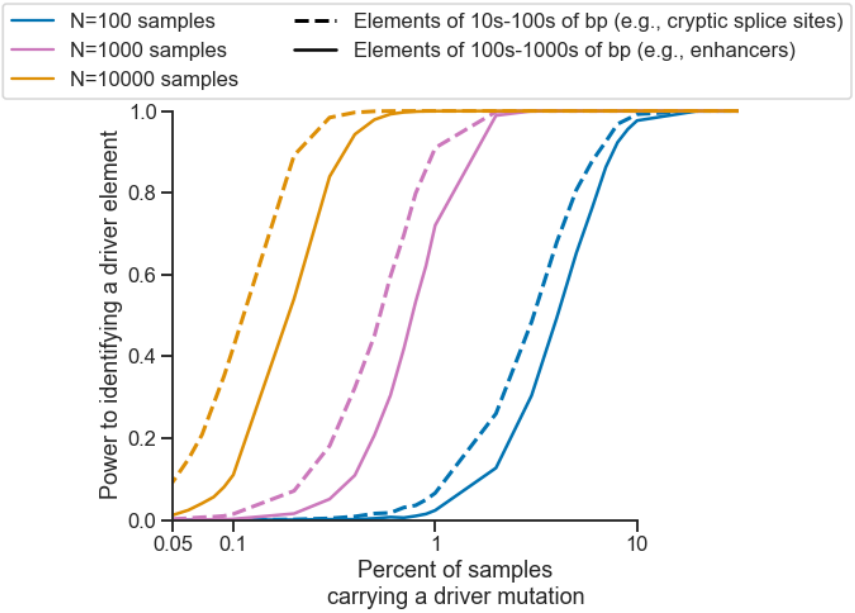

Figure D-6: Simulated power to detect driver elements in a pan-cancer cohort by sample size and by size of the elements being tested. The simulations were performed based on cryptic splice sites in 15,000 genes and 15,000 enhancers.

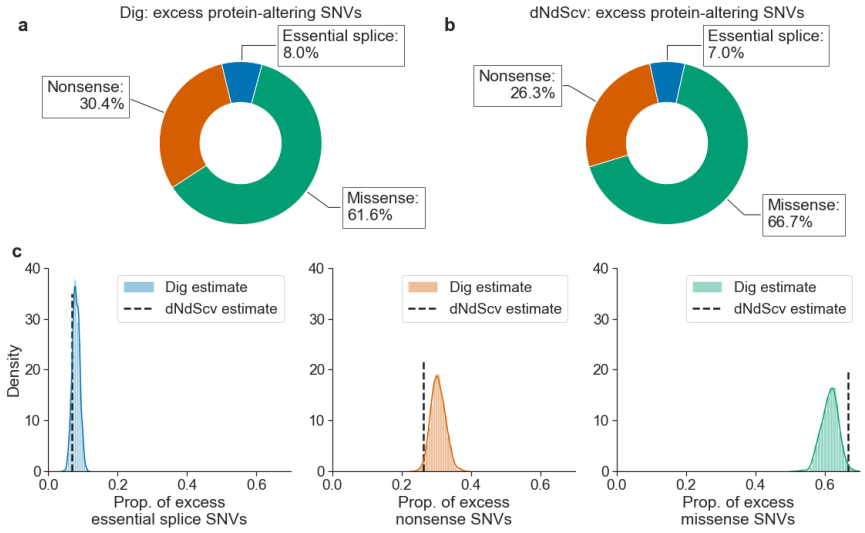

Figure D-7: Proportion of excess protein-altering SNVs in TSGs as estimated by Dig, a, and dNdScv, b. c, Distribution of proportion of excess SNVs as estimated using a Monte Carlo simulation approach based on Dig (Methods section 6.5) with the corresponding dNdScv estimate indicated with a black dashed line. Essential splice SNVs include SNVs at canonical splice sites (see fig. 6-3a) and SNVs 5 bp 5' of an exon start, which dNdScv also considers in its analysis of splice mutations.

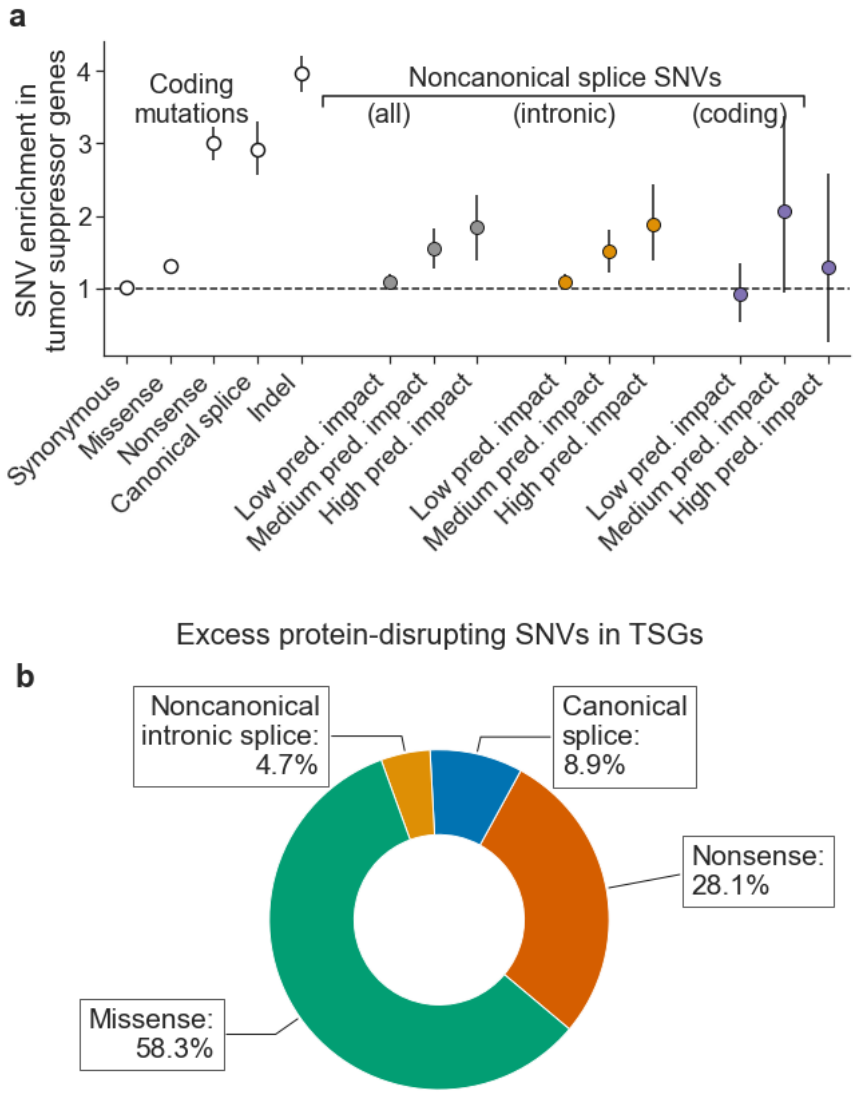

Figure D-8: SNV enrichment (with 95% CI) and excess analysis excluding samples with >3000 coding mutations. a, as in fig. 6-3b but excluding samples with >3000 coding mutations (default filtering criterion in dNdScv) (N=2,271 samples). b, As in fig. 6-3e but excluding samples with >3000 coding mutations.

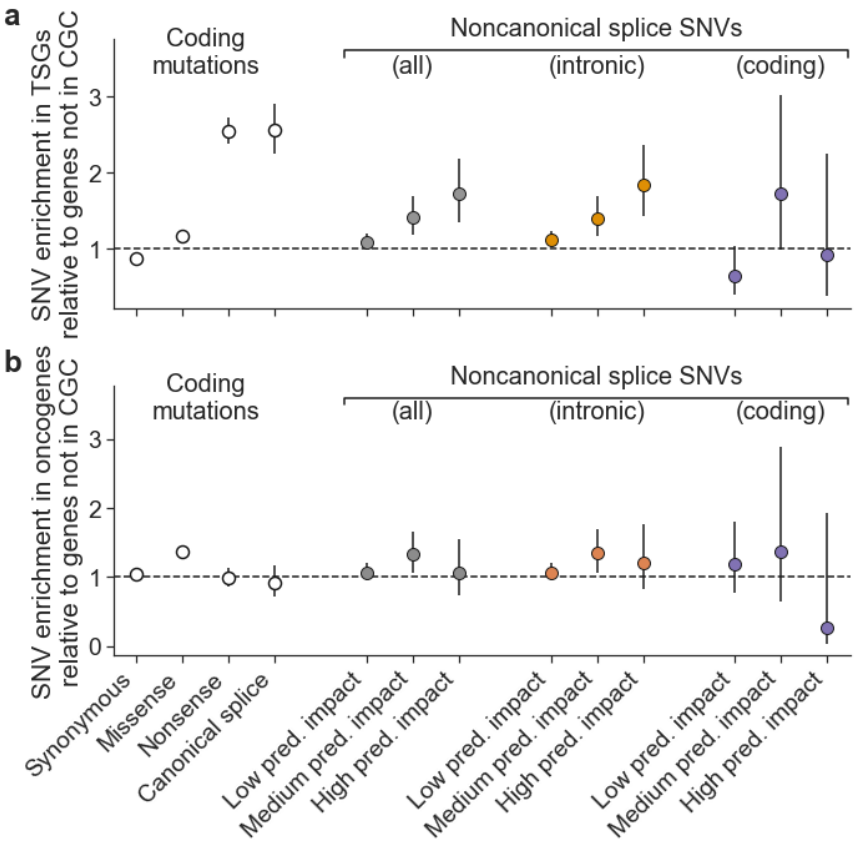

Figure D-9: Estimated SNV enrichment with 95% CI in tumor suppressor genes (TSGs), a, and oncogenes, b, with enrichment calculated with respect to the number of observed mutations in genes not in the Cancer Gene Census (CGC). Enrichment is calculated as the rate of SNVs of a given type observed in TSGs (oncogenes) relative to the rate of SNVs of the same type observed in genes not in the CGC. (N=2,279 samples in both panels).

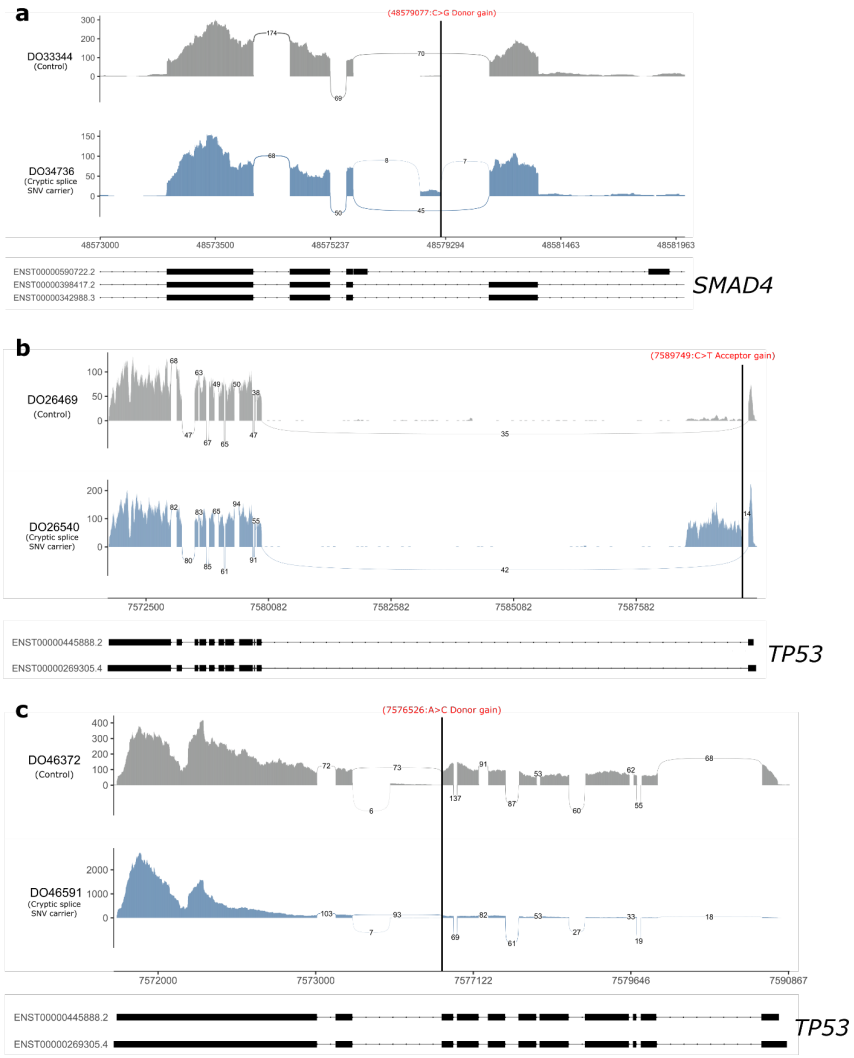

Figure D-10: Additional predicted cryptic splice SNV carriers in which LeafCutter identified strong evidence of alternative splicing. The location of the predicted cryptic splice SNV is marked with a thick black vertical line and labeled in red. a, *SMAD4* cryptic splice carrier. b,c *TP53* cryptic splice SNV carriers.

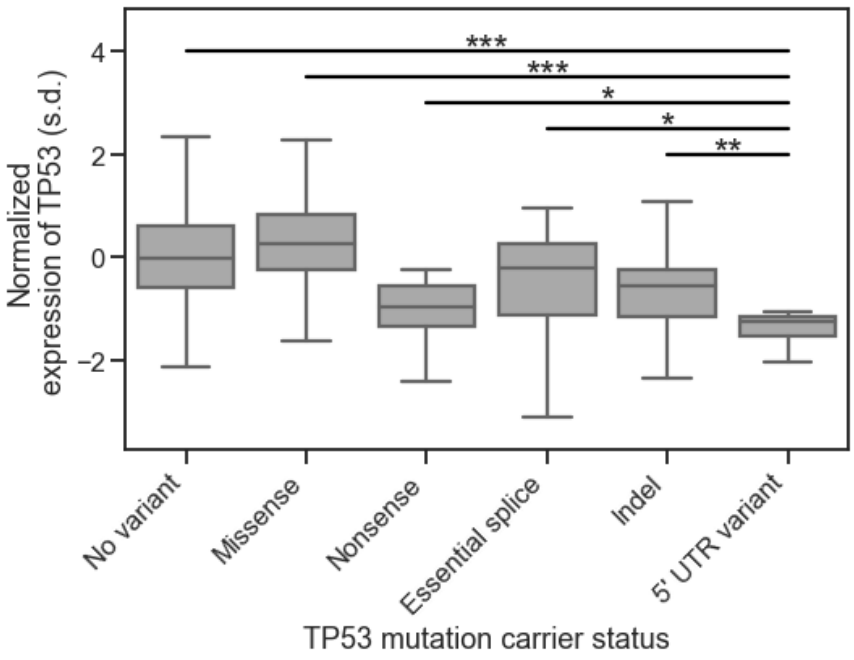

Figure D-11: Normalized expression of *TP53* stratified by the type of mutation individuals carry in *TP53*. P-values comparing expression of 5' UTR variant carrier to other carrier categories: 5' UTR vs no variant: $P = 1.2 \times 10^{-4}$; 5' UTR vs. missense: $P = 3.3 \times 10^{-5}$; 5' UTR vs. nonsense: $P = 0.023$; 5' UTR vs. essential splice: $P = 0.011$; 5' UTR vs. coding indel: $P = 8.5 \times 10^{-3}$. All p-values by one-sided Mann-Whitney U-test.

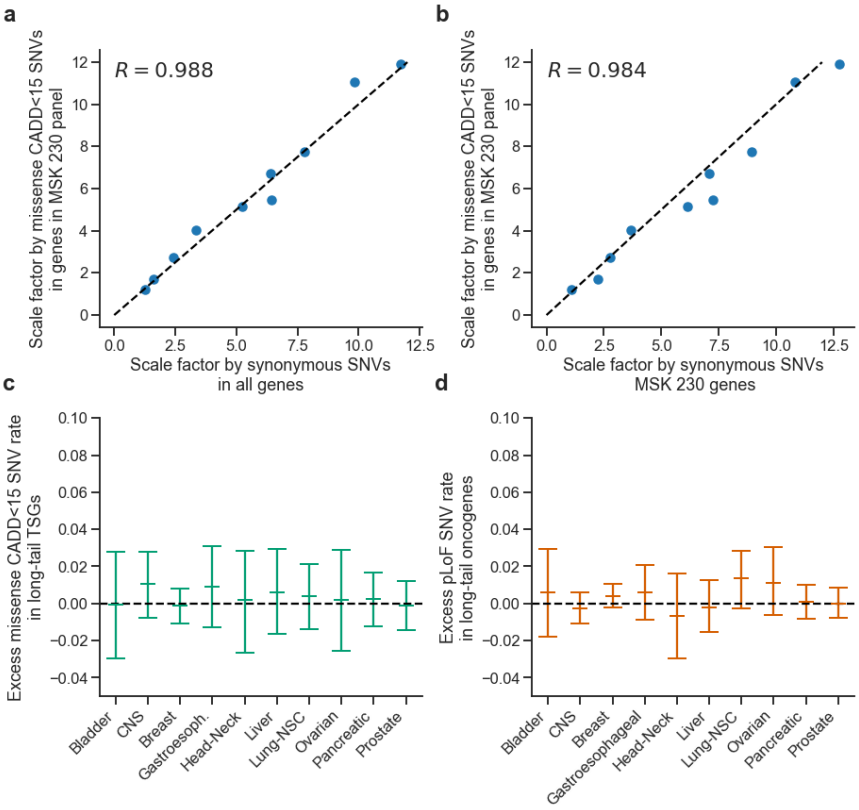

Figure D-12: Evaluation of neutral mutation model for ten solid cancer megacohorts. Using whole-exome sequenced samples, we compared the accuracy of estimating the scaling factor based on missense SNVs with CADD phred<15 observed in genes in the MSK IMPACT 230 targeted sequencing panel (the approach used for analyzing the megacohorts, see Methods) to the scaling factor estimated using synonymous mutations observed in all autosomal genes (Dig's default method), a, and using synonymous mutations observed in genes in the MSK IMPACT 230 targeted sequencing panel, b. c, The estimated rate of excess missense SNVs with CADD phred<15 (with 95% CI) in tumor suppressor genes in the MSK IMPACT 230 targeted sequencing panel. The burden of missense SNVs with CADD phred<15 is not significant in any cancer type. d, The rate of excess pLoF SNVs in oncogenes (with 95% CI) in the MSK IMPACT 230 targeted sequencing panel. The burden of pLoF SNVs is not significant in any cancer type.

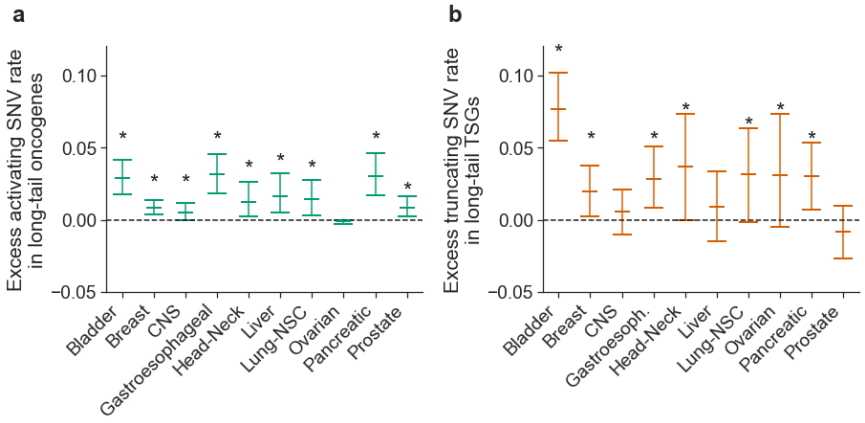

Figure D-13: Estimated excess activating SNV rate in oncogenes with 95% CIs, a, and excess pLoF SNV rate in TSGs with 95% CIs, b, as in fig. 6-4 but with analysis restricted to whole-exome sequenced samples only. Asterisks indicate the burden of SNVs is significant in the given cancer type. Error bars are larger than in 6-4a,b because sample size is smaller.

D.4 Supplementary Tables

PCAWG cancer code	MSI samples in cohort	Number of samples (excluding MSI high)	Number of SNVs (excluding MSI high)	Number of indels (excluding MSI high)
Adenocarcinoma_tumors	TRUE	1622	22927110	1346913
Biliary-AdenoCA	FALSE	34	421403	135990
Bladder-TCC	FALSE	23	486025	17926
Bone-Leiomyo	FALSE	34	172618	9343
Bone-Osteosarc	FALSE	41	147296	7854
Breast-AdenoCa	FALSE	195	1305975	89681
Breast_tumors	FALSE	208	1396392	92707
CNS-GBM	FALSE	39	478194	19565
CNS-Medullo	FALSE	141	184168	21725
CNS_tumors	FALSE	287	726664	46066
Carcinoma_tumors	TRUE	1847	26554707	1527041
ColoRect-AdenoCA	TRUE	50	8270011	161029
Digestive_tract_tumors	TRUE	792	17188295	883504
Eso-AdenoCa	FALSE	97	2557599	145549
Female_reproductive_system_tumors	TRUE	378	3012549	221936
Glioma_tumors	FALSE	146	542496	24377
Head-SCC	FALSE	56	835990	41955
Hematopoietic_tumors	FALSE	235	1437483	103016
Kidney-RCC	FALSE	143	857733	133974
Kidney_tumors	FALSE	186	932040	139803
Liver-HCC	FALSE	314	3600337	252980
Lung-AdenoCA	FALSE	37	1231550	63730
Lung-SCC	FALSE	47	1960685	109184
Lung_tumors	FALSE	84	3192235	172875
Lymph-BNHL	FALSE	105	1164199	88542
Lymph-CLL	FALSE	90	205996	11659
Lymph_tumors	FALSE	197	1395113	101257
Ovary-AdenoCA	FALSE	110	911513	88368
Panc-AdenoCA	FALSE	232	1410855	163903
Panc-Endocrine	FALSE	81	239462	12277
Pancan	FALSE	2279	28682089	2004228
Prost-AdenoCA	FALSE	199	607393	57541
Sarcoma_tumors	FALSE	95	349465	19414
Skin-Melanoma	FALSE	107	11585437	314368
Squamous_tumors	FALSE	121	2902110	157191
Stomach-AdenoCA	TRUE	65	928090	60175
Uterus-AdenoCA	TRUE	40	591830	34734

Table D.1: Information about the 37 PCAWG cancer cohorts used to train Dig's mutation rate models.

PCAWG cancer cohort	Fold_1 R2	Fold_2 R2	Fold_3 R2	Fold_4 R2	Fold_5 R2
Adenocarcinoma_tumors	0.918957259	0.920199532	0.92084372	0.920420119	0.921227
Biliary-AdenoCA	0.159543393	0.153712927	0.161466432	0.15548508	0.158615
Bladder-TCC	0.093561652	0.096111685	0.09370767	0.10493435	0.090787
Bone-Leiomyo	0.100839814	0.09198088	0.097773089	0.097283233	0.099921
Bone-Osteosarc	0.102080291	0.107975532	0.10504885	0.108352047	0.111192
Breast-AdenoCa	0.233392222	0.232596857	0.222005497	0.222182137	0.230129
Breast_tumors	0.23160267	0.236206516	0.232948879	0.229368443	0.235223
CNS-GBM	0.155941469	0.152852124	0.160001526	0.161626354	0.160865
CNS-Medullo	0.056017201	0.054874819	0.055408597	0.055214318	0.055461
CNS_tumors	0.200499694	0.208184835	0.204590346	0.203489125	0.207013
Carcinoma_tumors	0.930802911	0.927844493	0.929195139	0.930959549	0.930235
ColoRect-AdenoCA	0.806928846	0.809537932	0.809242211	0.809217844	0.806461
Digestive_tract_tumors	0.913153946	0.914786675	0.915964141	0.917274978	0.915462
Eso-AdenoCa	0.84146191	0.83870089	0.837969342	0.840333921	0.836239
Female_reproductive_system_tumors	0.475950291	0.477638573	0.475720539	0.471543677	0.470669
Glioma_tumors	0.177448169	0.181617708	0.181398039	0.184360074	0.185118
Head-SCC	0.415764933	0.422772844	0.414905789	0.421471756	0.412209
Hematopoietic_tumors	0.66365421	0.660048014	0.663345927	0.661911568	0.661729
Kidney-RCC	0.197549456	0.195892232	0.200420277	0.198640356	0.202251
Kidney_tumors	0.214503216	0.209528507	0.211909113	0.21112298	0.212555
Liver-HCC	0.753755021	0.750048604	0.753049178	0.749070396	0.754676
Lung-AdenoCA	0.589564947	0.592223691	0.587637085	0.593390371	0.590048
Lung-SCC	0.751722084	0.744283421	0.744044097	0.74499904	0.750165
Lung_tumors	0.795927779	0.797426594	0.796048109	0.795781177	0.797582
Lymph-BNHL	0.619589997	0.62002137	0.62140508	0.618096388	0.615534
Lymph-CLL	0.226229235	0.220440701	0.225749969	0.234055611	0.237688
Lymph_tumors	0.661000575	0.66246853	0.658772919	0.669499312	0.656566
Ovary-AdenoCA	0.314582108	0.316631599	0.316777848	0.318607898	0.315259
Panc-AdenoCA	0.564556965	0.558164325	0.565755976	0.568496612	0.561917
Panc-Endocrine	0.086544892	0.087967663	0.091137666	0.082849311	0.094436
Pancan	0.928776305	0.927468551	0.929467186	0.929231796	0.928408
Prost-AdenoCA	0.302719943	0.308234709	0.287893416	0.314358688	0.302626
Sarcoma_tumors	0.184687828	0.180720697	0.180044256	0.183628744	0.176017
Skin-Melanoma	0.912902822	0.912437702	0.914483541	0.913165657	0.913739
Squamous_tumors	0.765511687	0.770300171	0.771727313	0.770406053	0.77486

Table D.2: Variance explained of SNV counts in 10kb regions in held-out test data per fold.

	Dig (mutation map method)							
	Epigenetics only		Epigenetics & autoregression		NBR			
PCAWG cancer cohort	1Mb	10kb	10kb		10kb		N_SAMPLES	N_SNVS
Bone-Osteosarc	0.762397502	0.097705721	0.105403829		0.07356512		41	147296
Bone-Leiomyo	0.632934419	0.070315813	0.094698034		0.072018018		34	172618
CNS-Medullo	0.62325403	0.046471993	0.054585715		0.03784277		141	184168
Lymph-CLL	0.011062646	0.000119605	0.218881093		0.179932359		90	205996
Panc-Endocrine	0.719018688	0.077896134	0.088192437		0.082441717		81	239462
Sarcoma_tumors	0.799670003	0.160428289	0.180047275		0.148143625		95	349465
Biliary-AdenoCA	0.832524155	0.141677143	0.157383322		0.155395946		34	421403
CNS-GBM	0.690874734	0.12210819	0.157085493		0.139083746		39	478194
Bladder-TCC	0.724027393	0.079330139	0.093148161		0.105067891		23	486025
Glioma_tumors	0.715322665	0.141919626	0.180075218		0.159451746		146	542496
Uterus-AdenoCA	0.83918933	0.225759519	0.258828798		0.23972845		40	591830
Prost-AdenoCA	0.799613419	0.237597235	0.301137401		0.256896811		199	607393
CNS_tumors	0.764176435	0.170915811	0.200078732		0.173431748		287	726664
Head-SCC	0.911758088	0.382609641	0.416539007		0.392665114		56	835990
Kidney-RCC	0.767899271	0.147843571	0.197716808		0.198247594		143	857733
Ovary-AdenoCA	0.791652276	0.248288033	0.315226757		0.288544809		110	911513
Stomach-AdenoCA	0.932541527	0.576731284	0.627379688		0.550732866		65	928090
Kidney_tumors	0.802218929	0.159346194	0.210508499		0.209764135		186	932040
Lymph-BNHL	0.03779398	0.001040949	0.618649649		3.03E-06		105	1164199
Lung-AdenoCA	0.898850542	0.519971276	0.588167241		0.491606212		37	1231550
Breast-AdenoCa	0.745866918	0.173171051	0.22704735		0.231369184		195	1305975
Lymph_tumors	0.00702344	2.27E-06	0.65422317		1.35E-07		197	1395113
Breast_tumors	0.731436695	0.169925033	0.232526791		0.237625256		208	1396392
Panc-AdenoCA	0.907123914	0.498003971	0.561850614		0.498740443		232	1410855
Hematopoietic_tumors	0.059114053	0.0002316	0.656551932		1.20E-07		235	1437483
Lung-SCC	0.921068822	0.654442534	0.739761779		0.634789122		47	1960685
Eso-AdenoCa	0.949624324	0.759730327	0.829427682		0.726554527		97	2557599
Squamous_tumors	0.94270079	0.701117651	0.747912941		0.677539257		121	2902110
Female_reproductive_system_tumors	0.885428477	0.410332778	0.470997613		0.463893301		378	3012549
Lung_tumors	0.938624452	0.724653039	0.795634455		0.68159458		84	3192235
Liver-HCC	0.952694706	0.690347567	0.751170561		0.708881822		314	3600337
ColoRect-AdenoCA	0.966494387	0.770510396	0.804909184		0.718308479		50	8270011
Skin-Melanoma	0.955168667	0.825755625	0.899761551		0.727146043		107	11585437
Digestive_tract_tumors	0.978536019	0.881160974	0.895504079		0.837177826		792	17188295
Adenocarcinoma_tumors	0.978324345	0.882774356	0.918856564		0.844775567		1622	22927110
Carcinoma_tumors	0.979283788	0.899521417	0.912160523		0.852608566		1847	26554707
Pancan	0.97963231	0.898492051	0.923424002		0.853363275		2279	28682089

Table D.3: Variance explained of SNV counts in tiled regions by method.

PCAWG cancer code	Dig: Genes 1-1.5kb	dNdScv: Genes 1-1.5kb	MutSigCV: Genes 1-1.5kb	N_SAMPLES	N_SNVS
Adenocarcinoma_tumors	0.36984283	0.324439295	0.132546817	1622	22927110
Biliary-AdenoCA	0.009527537	0.00499941	0.001019776	34	421403
Bladder-TCC	0.011985841	0.001984919	0.00404049	23	486025
Bone-Leiomyo	0.018263114	0.006164119	0.012463617	34	172618
Bone-Osteosarc	0.004832731	0.000103288	0.002505564	41	147296
Breast-AdenoCa	0.025021686	0.008613777	0.000335188	195	1305975
Breast_tumors	0.027037909	0.008935645	0.00023034	208	1396392
CNS-GBM	0.019712364	0.006369507	8.11E-07	39	478194
CNS-Medullo	0.008021052	0.006501051	0.002295228	141	184168
CNS_tumors	0.02636428	0.017306363	0.000539927	287	726664
Carcinoma_tumors	0.407337355	0.356241003	0.168235206	1847	26554707
ColoRect-AdenoCA	0.252008236	0.236454417	0.041074992	50	8270011
Digestive_tract_tumors	0.372783347	0.319727908	0.119288222	792	17188295
Eso-AdenoCa	0.200348903	0.131525874	0.033390872	97	2557599
Female_reproductive_system_tumors	0.063067647	0.035675414	0.004900271	378	3012549
Glioma_tumors	0.021897343	0.008053415	1.02E-08	146	542496
Head-SCC	0.033617321	0.005892584	0.000261147	56	835990
Hematopoietic_tumors	0.037876133	0.071388491	0.0685116	235	1437483
Kidney-RCC	0.008902512	0.005027209	0.000275905	143	857733
Kidney_tumors	0.007802823	0.005789137	0.000136655	186	932040
Liver-HCC	0.108228558	0.077156601	0.009880537	314	3600337
Lung-AdenoCA	0.128046322	0.08122856	0.004437913	37	1231550
Lung-SCC	0.17547039	0.099875624	0.043329839	47	1960685
Lung_tumors	0.249289702	0.154580288	0.056641124	84	3192235
Lymph-BNHL	0.042179309	0.071468146	0.041003086	105	1164199
Lymph-CLL	0.020667031	0.012079374	0.007162756	90	205996
Lymph_tumors	0.051449924	0.074423614	0.070511112	197	1395113
Ovary-AdenoCA	0.05285004	0.025742013	0.001478249	110	911513
Panc-AdenoCA	0.108227537	0.06937479	0.030511114	232	1410855
Panc-Endocrine	0.009100995	0.000825028	0.00029645	81	239462
Pancan	0.39455346	0.357141607	0.177603941	2279	28682089
Prost-AdenoCA	0.018111828	0.002106043	5.10E-05	199	607393
Sarcoma_tumors	0.021513503	0.011916866	0.009533251	95	349465
Skin-Melanoma	0.500551255	0.43681911	0.226255725	107	11585437
Squamous_tumors	0.153275439	0.102728507	0.029573042	121	2902110
Stomach-AdenoCA	0.065932971	0.036848403	0.011671477	65	928090
Uterus-AdenoCA	0.009994638	0.005796513	0.009084403	40	591830

Table D.4: Variance explained of SNV counts in genes by method. Genes were restricted to 1-1.5kb in size to avoid inflated correlation due to long genes harboring more mutations.

PCAWG cancer code	Dig: Noncoding 0.5-1kb	DriverPower: Noncoding 0.5-1kb	Larva: Noncoding 0.5-1kb	N_SAMPLES	N_SNVS
Adenocarcinoma_tumors	0.462461624	0.444856467	0.245211891	1622	22927110
Biliary-AdenoCA	0.010615543	0.007413831	0.004666811	34	421403
Bladder-TCC	0.007294131	0.009252924	0.004000205	23	486025
Bone-Leiomyo	0.007334541	0.004253133	0.001046878	34	172618
Bone-Osteosarc	0.008328242	0.002897138	0.002576206	41	147296
Breast-AdenoCa	0.031599694	0.02779063	0.020006713	195	1305975
Breast_tumors	0.035613099	0.028178649	0.020625374	208	1396392
CNS-GBM	0.011916703	0.008137553	0.004966878	39	478194
CNS-Medullo	0.002570318	0.000853499	9.32E-05	141	184168
CNS_tumors	0.017990944	0.014501308	0.006042504	287	726664
Carcinoma_tumors	0.482895789	0.472310708	0.260955005	1847	26554707
ColoRect-AdenoCA	0.254869028	0.269728386	0.119435972	50	8270011
Digestive_tract_tumors	0.420043013	0.411958887	0.216591575	792	17188295
Eso-AdenoCa	0.257735967	0.210169807	0.096802307	97	2557599
Female_reproductive_system_tumors	0.073915485	0.06686976	0.043663629	378	3012549
Glioma_tumors	0.015297448	0.011705383	0.006110022	146	542496
Head-SCC	0.029824354	0.02456643	0.016150845	56	835990
Hematopoietic_tumors	0.075731138	0.17249458	0.02316826	235	1437483
Kidney-RCC	0.014439495	0.012149736	0.005348982	143	857733
Kidney_tumors	0.017508012	0.015149273	0.006051284	186	932040
Liver-HCC	0.13358578	0.113670384	0.055806986	314	3600337
Lung-AdenoCA	0.096926586	0.076168012	0.035972826	37	1231550
Lung-SCC	0.119020435	0.112093102	0.044193796	47	1960685
Lung_tumors	0.186841845	0.166541019	0.073213185	84	3192235
Lymph-BNHL	0.063599692	0.223539672	0.02017316	105	1164199
Lymph-CLL	0.01477102	0.019964615	0.002869134	90	205996
Lymph_tumors	0.076303003	0.187018554	0.022518406	197	1395113
Ovary-AdenoCA	0.025488055	0.022336736	0.013814855	110	911513
Panc-AdenoCA	0.066334436	0.05330806	0.025622923	232	1410855
Panc-Endocrine	0.008090195	0.004234909	0.004911819	81	239462
Pancan	0.489666221	0.475063443	0.263919607	2279	28682089
Prost-AdenoCA	0.017733388	0.014463711	0.005592124	199	607393
Sarcoma_tumors	0.012440446	0.008517678	0.003491204	95	349465
Skin-Melanoma	0.55459782	0.471832625	0.129979884	107	11585437
Squamous_tumors	0.133050942	0.129394307	0.0548656	121	2902110
Stomach-AdenoCA	0.101500776	0.085926229	0.042556644	65	928090
Uterus-AdenoCA	0.025900057	0.019636014	0.013893136	40	591830

Table D.5: Variance explained of SNV counts in enhancers and noncoding RNAs restricted in length to 0.5-1kb to prevent inflation of correlation due to larger genomic regions harboring more mutation counts.

	Dig						dNdScv						MutSigCV					
PCAWG cancer code	TP	FP	TN	FN	RECALL	PRECISION	F1	TP	FP	TN	FN	RECALL	PRECISION	F1	TP	FP	TN	FN	RECALL	PRECISION	F1
meta_Adenocarcinoma	48	21	16724	1	0.97959184	0.69565217	0.81356	43	20	16725	6	0.87755102	0.68253968	0.76786	33	15	16730	16	0.67346939	0.6875	0.68041
Biliary-AdenoCA	4	1	16789	0	1	0.8	0.88889	4	1	16789	0	1	0.8	0.88889	2	0	16790	2	0.5	1	0.66667
Bladder-TCC	6	1	16787	0	1	0.85714286	0.92308	4	2	16786	2	0.66666667	0.66667	0.66667	0	0	16788	6	0	#N/A	#N/A
Bone-Leiomyo	2	0	16792	0	1	1	1	2	0	16792	0	1	1	1	1	0	16792	1	0.5	1	0.66667
Bone-Osteosarc	2	1	16791	0	1	0.66666667	0.8	2	1	16791	0	1	0.66666667	0.8	1	0	16792	1	0.5	1	0.66667
Breast-AdenoCa	12	0	16782	0	1	1	1	10	0	16782	2	0.83333333	1	0.90909	6	0	16782	6	0.5	1	0.66667
meta_Breast	12	0	16781	1	0.92307692	1	0.96	13	0	16781	0	1	1	1	7	1	16780	6	0.53846154	0.875	0.66667
CNS-GBM	3	0	16790	1	0.75	1	0.85714	4	0	16790	0	1	1	1	3	0	16790	1	0.75	1	0.85714
CNS-Medullo	4	4	16785	1	0.8	0.5	0.61538	5	4	16785	0	1	0.55555556	0.71429	1	2	16787	4	0.2	0.33333333	0.25
meta_CNS	15	4	16774	1	0.9375	0.78947368	0.85714	14	4	16774	2	0.875	0.77777778	0.82353	5	1	16777	11	0.3125	0.83333333	0.45455
meta_Carcinoma	56	25	16709	4	0.93333333	0.69135802	0.79433	53	22	16712	7	0.88333333	0.70666667	0.78519	41	19	16715	19	0.68333333	0.68333333	0.68333
ColoRect-AdenoCA	7	4	16780	3	0.7	0.63636364	0.66667	9	2	16782	1	0.9	0.81818182	0.85714	6	1	16783	4	0.6	0.85714286	0.70588
meta_Digestive_tract	26	13	16752	3	0.89655172	0.66666667	0.76471	26	15	16750	3	0.89655172	0.63414634	0.74286	19	12	16753	10	0.65517241	0.61290323	0.63333
Eso-AdenoCa	6	1	16787	0	1	0.85714286	0.92308	5	3	16785	1	0.83333333	0.625	0.71429	5	1	16787	1	0.83333333	0.83333333	0.83333
meta_Female_reproductive_tract	19	0	16774	1	0.95	1	0.97436	17	2	16772	3	0.85	0.89473684	0.87179	11	1	16773	9	0.55	0.91666667	0.6875
meta_Glioma	10	0	16782	2	0.83333333	1	0.90909	11	0	16782	1	0.91666667	1	0.95652	4	0	16782	8	0.33333333	1	0.5
Head-SCC	7	1	16784	2	0.77777778	0.875	0.82353	9	2	16783	0	1	0.81818182	0.9	4	0	16785	5	0.44444444	1	0.61538
Kidney-RCC	5	0	16789	0	1	1	1	5	0	16789	0	1	1	1	5	0	16789	0	1	1	1
meta_Kidney	6	0	16788	0	1	1	1	6	0	16788	0	1	1	1	6	0	16788	0	1	1	1
Liver-HCC	13	10	16771	0	1	0.56521739	0.72222	12	10	16771	1	0.92307692	0.54545455	0.68571	10	5	16776	3	0.76923077	0.66666667	0.71429
Lung-AdenoCA	4	0	16790	1	1	1	1	4	0	16790	0	1	1	1	4	0	16790	0	1	1	1
Lung-SCC	4	0	16789	1	0.8	1	0.88889	5	1	16788	0	1	0.83333333	0.90909	4	0	16789	1	0.8	1	0.88889
meta_Lung	9	0	16785	0	1	1	1	8	1	16784	1	0.88888889	0.88888889	0.88889	6	0	16785	3	0.66666667	1	0.8
Ovary-AdenoCA	4	0	16790	0	1	1	1	4	0	16790	0	1	1	1	1	0	16790	3	0.25	1	0.4
Panc-AdenoCA	12	3	16779	0	1	0.8	0.88889	10	3	16779	2	0.83333333	0.76923077	0.8	8	2	16780	4	0.66666667	0.8	0.72727
Panc-Endocrine	4	0	16790	0	1	1	1	4	0	16790	0	1	1	1	3	0	16790	1	0.75	1	0.85714
PANCANCER	61	20	16707	6	0.91044776	0.75308642	0.82432	54	27	16700	13	0.80597015	0.66666667	0.72973	43	25	16702	24	0.64179104	0.63235294	0.63704
Prost-AdenoCA	5	0	16789	0	1	1	1	5	0	16789	0	1	1	1	4	0	16789	1	0.8	1	0.88889
meta_Sarcoma	3	1	16790	0	1	0.75	0.85714	3	1	16790	0	1	0.75	0.85714	2	0	16791	1	0.66666667	1	0.8
meta_Squamous	11	1	16780	2	0.84615385	0.91666667	0.88	12	1	16780	1	0.92307692	0.92307692	0.92308	5	0	16781	8	0.38461538	1	0.55556
Stomach-AdenoCA	5	2	16787	0	1	0.71428571	0.83333	5	4	16785	0	1	0.55555556	0.71429	1	0	16789	4	0.2	1	0.33333
Uterus-AdenoCA	7	1	16785	1	0.875	0.875	0.875	7	1	16785	1	0.875	0.875	0.875	3	0	16786	5	0.375	1	0.54545
COMBINED	392	114	536872	30	0.92890995	0.77470356	0.84483	375	127	536859	47	0.88862559	0.74701195	0.81169	254	85	536901	168	0.60189573	0.74926254	0.66754

Table D.6: Accuracy measures of driver gene detection methods at FDR<0.1.

	Cohort	Dig	dNdScv	MutSigCV
PCAWG	Pan-cancer (all samples)	58295	49652.5	40561.5
	Pan-cancer (no hypermutated samples)	57901	52334	40561.5
Dietlein et al.	Bladder	37267	33340.5	29320.5
	Brain	24532.5	23397.5	19871
	Breast	37888.5	32580.5	29319
	Colorectal	23872.5	20874	20042.5
	Endometrial	27592.5	25399.5	20346
	Gastroesophageal	35422.5	23426.5	22685.5
	HeadNeck	28046.5	25705.5	20171
	KindeyClear	19718	18498	14988
	Liver	20413	17948	16798.5
	LungAD	26213.5	26173.5	20509
	LungSC	17703	18769.5	11413
	Ovarian	19460	15299	9856.5
	Pancreas	32653.5	29874	#N/A
	Pancancer	74422.5	63422	50455.5
	Prostate	28090	27177.5	22944
	Sarcoma	13788	14309	9373

Table D.7: Areas under the approximated ROC curves of Fig fig. 6-1e and Supplementary Fig fig. D-4.

	TOTAL_HITS	Dig (hypermutated samples removed)							Dig (no samples removed)						
	TP	FP	TN	FN	RECALL	PRECISION	F1	TP	FP	TN	FN	RECALL	PRECISION	F1
meta_Adenocarcinoma	15	12	6	95740	3	0.8	0.666666667	0.727273	13	10	95736	2	0.866667	0.565217391	0.684211
Biliary-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A	0	1	95760	0	#N/A	0	#N/A
Bladder-TCC	6	3	1	95754	3	0.5	0.75	0.6	5	5	95750	1	0.833333	0.5	0.625
Bone-Leiomyo	1	0	0	95760	1	0	#N/A	#N/A	1	5	95755	0	1	0.166666667	0.285714
Bone-Osteosarc	0	0	0	95761	0	#N/A	#N/A	#N/A	0	5	95756	0	#N/A	0	#N/A
Breast-AdenoCa	4	0	0	95757	4	0	#N/A	#N/A	4	8	95749	0	1	0.333333333	0.5
meta_Breast	5	1	0	95756	4	0.2	1	0.333333	5	7	95749	0	1	0.416666667	0.588235
CNS-GBM	1	1	0	95760	0	1	1	1	1	0	95760	0	1	1	1
CNS-Medullo	1	1	0	95760	0	1	1	1	1	0	95760	0	1	1	1
meta_CNS	1	1	0	95760	0	1	1	1	1	0	95760	0	1	1	1
meta_Carcinoma	27	23	4	95730	4	0.851852	0.851851852	0.851852	24	26	95708	3	0.888889	0.48	0.623377
ColoRect-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A	0	5	95756	0	#N/A	0	#N/A
meta_Digestive_tract	9	9	1	95751	0	0	0.9	0.947368	8	8	95744	1	0.888889	0.5	0.64
Eso-AdenoCa	2	1	0	95759	1	0.5	1	0.666667	2	3	95756	0	1	0.4	0.571429
meta_Female_reproductive_tract	6	1	0	95755	5	0.166667	1	0.285714	4	4	95751	2	0.666667	0.5	0.571429
meta_Glioma	1	1	0	95760	0	0	1	1	1	0	95760	0	1	1	1
Head-SCC	2	1	0	95759	1	0.5	1	0.666667	2	2	95757	0	1	0.5	0.666667
Kidney-RCC	1	1	0	95760	0	1	1	1	1	0	95760	0	1	1	1
meta_Kidney	1	1	0	95760	0	1	1	1	1	1	95759	0	1	0.5	0.666667
Liver-HCC	11	11	1	95749	0	0	1	0.916666667	0.956522	9	1	95749	2	0.818182	0.9	0.857143
Lung-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A	0	2	95759	0	#N/A	0	#N/A
Lung-SCC	0	0	0	95761	0	#N/A	#N/A	#N/A	0	1	95760	0	#N/A	0	#N/A
meta_Lung	0	0	0	95761	0	#N/A	#N/A	#N/A	0	1	95760	0	#N/A	0	#N/A
Ovary-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A	0	0	95761	0	#N/A	#N/A	#N/A
Panc-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A	0	3	95758	0	#N/A	0	#N/A
Panc-Endocrine	0	0	0	95761	0	#N/A	#N/A	#N/A	0	0	95761	0	#N/A	#N/A	#N/A
PANCANCER	23	18	4	95734	5	0.782609	0.818181818	0.8	19	8	95730	4	0.826087	0.703703704	0.76
Prost-AdenoCA	1	0	0	95760	1	0	#N/A	#N/A	0	4	95756	1	0	0	#N/A
meta_Sarcoma	0	0	0	95761	0		#N/A	#N/A	0	6	95755	0	#N/A	0	#N/A
meta_Squamous	1	1	0	95760	0	1	1	1	1	2	95758	0	1	0.333333333	0.5
Stomach-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A	0	1	95760	0	#N/A	0	#N/A
Uterus-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A	0	0	95761	0	#N/A	#N/A	#N/A
COMBINED	119	87	17	3064216	32	0.731092	0.836538462	0.780269	103	119	3064114	16	0.865546	0.463963964	0.604106

Table D.8: Accuracy measures of noncoding driver detection methods at FDR<0.1 for Dig.

	TOTAL_HITS	DriverPower						
		TP	FP	TN	FN	RECALL	PRECISION	F1
meta_Adenocarcinoma	15	8	0	95746	7	0.533333	1	0.695652174
Biliary-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Bladder-TCC	6	3	0	95755	3	0.5	1	0.666666667
Bone-Leiomyo	1	0	0	95760	1	0	#N/A	#N/A
Bone-Osteosarc	0	0	0	95761	0	#N/A	#N/A	#N/A
Breast-AdenoCa	4	2	0	95757	2	0.5	1	0.666666667
meta_Breast	5	2	0	95756	3	0.4	1	0.571428571
CNS-GBM	1	1	0	95760	0	1	1	1
CNS-Medullo	1	1	0	95760	0	1	1	1
meta_CNS	1	1	0	95760	0	1	1	1
meta_Carcinoma	27	13	0	95734	14	0.481481	1	0.65
ColoRect-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
meta_Digestive_tract	9	5	0	95752	4	0.555556	1	0.714285714
Eso-AdenoCa	2	1	0	95759	1	0.5	1	0.666666667
meta_Female_reproductive_tract	6	1	0	95755	5	0.166667	1	0.285714286
meta_Glioma	1	1	0	95760	0	1	1	1
Head-SCC	2	1	0	95759	1	0.5	1	0.666666667
Kidney-RCC	1	0	0	95760	1	0	#N/A	#N/A
meta_Kidney	1	0	0	95760	1	0	#N/A	#N/A
Liver-HCC	11	5	0	95750	6	0.454545	1	0.625
Lung-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Lung-SCC	0	0	0	95761	0	#N/A	#N/A	#N/A
meta_Lung	0	0	0	95761	0	#N/A	#N/A	#N/A
Ovary-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Panc-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Panc-Endocrine	0	0	0	95761	0	#N/A	#N/A	#N/A
PANCANCER	23	14	1	95737	9	0.608696	0.933333333	0.736842105
Prost-AdenoCA	1	0	0	95760	1	0		
meta_Sarcoma	0	0	0	95761	0	#N/A	#N/A	#N/A
meta_Squamous	1	1	0	95760	0	1	1	1
Stomach-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Uterus-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
COMBINED	119	60	1	3064232	59	0.504202	0.983606557	0.666666667

Table D.9: Accuracy measures of noncoding driver detection methods at FDR<0.1 for DriverPower.

	TOTAL_HITS	ActiveDriverWGS						
		TP	FP	TN	FN	RECALL	PRECISION	F1
meta_Adenocarcinoma	15	6	2	95744	9	0.4	0.75	0.521739
Biliary-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Bladder-TCC	6	4	0	95755	2	0.666667	1	0.8
Bone-Leiomyo	1	0	0	95760	1	0	#N/A	#N/A
Bone-Osteosarc	0	0	0	95761	0	#N/A	#N/A	#N/A
Breast-AdenoCa	4	3	0	95757	1	0.75	1	0.857143
meta_Breast	5	4	0	95756	1	0.8	1	0.888889
CNS-GBM	1	1	0	95760	0	1	1	1
CNS-Medullo	1	1	0	95760	0	1	1	1
meta_CNS	1	1	0	95760	0	1	1	1
meta_Carcinoma	27	14	3	95731	13	0.518519	0.823529412	0.636364
ColoRect-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
meta_Digestive_tract	9	5	0	95752	4	0.555556	1	0.714286
Eso-AdenoCa	2	2	0	95759	0	1	1	1
meta_Female_reproductive_tract	6	4	0	95755	2	0.666667	1	0.8
meta_Glioma	1	1	0	95760	0	1	1	1
Head-SCC	2	1	0	95759	1	0.5	1	0.666667
Kidney-RCC	1	1	0	95760	0	1	1	1
meta_Kidney	1	1	0	95760	0	1	1	1
Liver-HCC	11	5	0	95750	6	0.454545	1	0.625
Lung-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Lung-SCC	0	0	0	95761	0	#N/A	#N/A	#N/A
meta_Lung	0	0	0	95761	0	#N/A	#N/A	#N/A
Ovary-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Panc-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Panc-Endocrine	0	0	0	95761	0	#N/A	#N/A	#N/A
PANCANCER	23	11	3	95735	12	0.478261	0.785714286	0.594595
Prost-AdenoCA	1	0	0	95760	1	0	#N/A	#N/A
meta_Sarcoma	0	0	0	95761	0	#N/A	#N/A	#N/A
meta_Squamous	1	1	0	95760	0	1	1	1
Stomach-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Uterus-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
COMBINED	119	66	8	3064225	53	0.554622	0.891891892	0.683938

Table D.10: Accuracy measures of noncoding driver detection methods at FDR<0.1 for ActiveDriverWGS.

	TOTAL_HITS	Larva						
		TP	FP	TN	FN	RECALL	PRECISION	F1
meta_Adenocarcinoma	15	2	0	95746	13	0.133333	1	0.235294
Biliary-AdenoCA	0	0	0	95761	0	#N/A	#N/A	#N/A
Bladder-TCC	6	6	1	95754	0	1	0.857142857	0.923077
Bone-Leiomyo	1	0	2	95758	1	0	0	#N/A
Bone-Osteosarc	0	0	7	95754	0	#N/A	#N/A	
Breast-AdenoCa	4	3	6	95751	1	0.75	0.333333333	0.461538
meta_Breast	5	4	7	95749	1	0.8	0.363636364	0.5
CNS-GBM	1	1	0	95760	0	1	1	1
CNS-Medullo	1	1	0	95760	0	1	1	1
meta_CNS	1	1	0	95760	0	1	1	1
meta_Carcinoma	27	7	0	95734	20	0.259259	1	0.411765
ColoRect-AdenoCA	0	0	8	95753	0	#N/A	0	#N/A
meta_Digestive_tract	9	2	0	95752	7	0.222222	1	0.363636
Eso-AdenoCa	2	0	5	95754	2	0	0	
meta_Female_reproductive_tract	6	6	8	95747	0	1	0.428571429	0.6
meta_Glioma	1	1	0	95760	0	1	1	1
Head-SCC	2	2	2	95757	0	1	0.5	0.666667
Kidney-RCC	1	0	0	95760	1	0	#N/A	#N/A
meta_Kidney	1	0	0	95760	1	0	#N/A	#N/A
Liver-HCC	11	1	0	95750	10	0.090909	1	0.166667
Lung-AdenoCA	0	0	3	95758	0	#N/A	0	#N/A
Lung-SCC	0	0	0	95761	0	#N/A	#N/A	#N/A
meta_Lung	0	0	2	95759	0	#N/A	0	#N/A
Ovary-AdenoCA	0	0	3	95758	0	#N/A	0	#N/A
Panc-AdenoCA	0	0	4	95757	0	#N/A	0	#N/A
Panc-Endocrine	0	0	0	95761	0	#N/A	#N/A	#N/A
PANCANCER	23	7	0	95738	16	0.304348	1	0.466667
Prost-AdenoCA	1	1	4	95756	0	1	0.2	0.333333
meta_Sarcoma	0	0	6	95755	0	#N/A	0	#N/A
meta_Squamous	1	1	0	95760	0	1	1	1
Stomach-AdenoCA	0	0	1	95760	0	#N/A	0	#N/A
Uterus-AdenoCA	0	0	1	95760	0	#N/A	0	#N/A
COMBINED	119	46	70	3064163	73	0.386555	0.396551724	0.391489

Table D.11: Accuracy measures of noncoding driver detection methods at FDR<0.1 for Larva.

	CATEGORY	SUBCATEGORY	OBSERVED	EXPECTED	ENRICHMENT	CI_LOWER	CI_UPPER	P_VALUE
Tumor suppressor genes in CGC	Gene	Synonymous	1502	1502.955749	0.999364087	0.952123841	1.050596467	0.513126806
		Missense	5574	4324.921641	1.288809477	1.249242287	1.329977853	2.93E-71
		Nonsense	941	326.5870386	2.881314592	2.682209936	3.08968171	2.26E-167
		Canonical splice	265	107.3748176	2.467990224	2.169968763	2.775324855	1.38E-37
		indels	1323	354.9731821	3.727042116	3.524209892	3.95811591	0
	Noncanonical splice (all)	0.2 < Delta < 0.5	450	415.5678698	1.082855612	0.976976397	1.176703098	0.049525369
		0.5 < Delta < 0.8	128	92.55979298	1.382889869	1.145205673	1.609770238	0.000281627
		0.8 < Delta < 1.0	70	41.69617301	1.678811146	1.295082884	2.062539408	3.93E-05
	Noncanonical intronic splice	0.2 < Delta < 0.5	427	390.056267	1.094713856	0.999855747	1.199826896	0.034022184
		0.5 < Delta < 0.8	115	85.29369163	1.348282596	1.113505562	1.606214919	0.001262974
		0.8 < Delta < 1.0	64	36.56239029	1.750432603	1.312824452	2.215391263	2.52E-05
	Noncanonical coding splice	0.2 < Delta < 0.5	18	23.36427861	0.770406838	0.428003799	1.155610257	0.891324075
		0.5 < Delta < 0.8	13	6.690311119	1.94310844	1.04628916	3.13886748	0.019712989
		0.8 < Delta < 1.0	5	4.858393713	1.029146729	0.205829346	2.058293459	0.534310275
Oncogenes in CGC	Gene	Synonymous	1159	1106.145653	1.047782448	0.989019843	1.109257173	0.060648465
		Missense	4048	3043.214193	1.330172556	1.285803349	1.376833747	2.17E-63
		Nonsense	215	216.7673072	0.991846985	0.857947642	1.125746328	0.55666725
		Canonical splice	64	72.52468519	0.882458157	0.648055209	1.116861104	0.855968363
		indel	327	251.5069378	1.300162941	1.164977804	1.431372046	3.15E-06
	Noncanonical splice (all)	0.2 < Delta < 0.5	279	260.9752659	1.069066829	0.950281626	1.195515594	0.139546584
		0.5 < Delta < 0.8	79	60.6762861	1.301991356	1.021816001	1.582166711	0.013623882
		0.8 < Delta < 1.0	28	25.82239446	1.084330117	0.697069361	1.510316948	0.359717383
	Noncanonical intronic splice	0.2 < Delta < 0.5	257	243.2484832	1.056532796	0.929091097	1.183974495	0.197075089
		0.5 < Delta < 0.8	72	55.33012907	1.301280174	0.975960131	1.608526882	0.017859353
		0.8 < Delta < 1.0	27	22.33268733	1.208990194	0.761216048	1.701541755	0.186556579
	Noncanonical coding splice	0.2 < Delta < 0.5	22	15.85375059	1.38768425	0.883071795	1.955373261	0.083173728
		0.5 < Delta < 0.8	7	4.957884417	1.411892535	0.403397867	2.420387203	0.231694962
		0.8 < Delta < 1.0	1	3.225359871	0.310042922	0	0.930128767	0.960250871
500 randomly sampled genes not in CGC	Gene	Synonymous	1460	1468.836443	0.993984052	0.942872167	1.043717295	0.593685717
		Missense	3961	4045.0097	0.979231274	0.949317872	1.010380766	0.90254688
		Nonsense	273	291.4519586	0.936689536	0.83375662	1.049915744	0.866571618
		Canonical splice	88	98.19266123	0.896197322	0.702700173	1.089694471	0.860325645
		indel	320	332.6485447	0.961976251	0.850747747	1.061180052	0.762661792
	Noncanonical splice (all)	0.2 < Delta < 0.5	369	338.2765684	1.09082341	0.981445453	1.19436295	0.051905393
		0.5 < Delta < 0.8	87	80.78904491	1.07687868	0.854076194	1.312059081	0.259056092
		0.8 < Delta < 1.0	37	34.35692798	1.076929812	0.75676149	1.426204346	0.348232631
	Noncanonical intronic splice	0.2 < Delta < 0.5	332	309.0511675	1.07425577	0.957689312	1.197212756	0.10204519
		0.5 < Delta < 0.8	75	72.39284104	1.036014044	0.814652377	1.270843894	0.395144359
		0.8 < Delta < 1.0	35	29.64978819	1.180446881	0.80944929	1.551444473	0.184653962
	Noncanonical coding splice	0.2 < Delta < 0.5	32	26.78910786	1.194515329	0.821229289	1.605129974	0.179700476
		0.5 < Delta < 0.8	11	7.703338144	1.427952375	0.649069261	2.336649341	0.155815964
		0.8 < Delta < 1.0	2	4.431819399	0.451281927	0	1.128204818	0.935385357

Table D.12: SNV enrichment across coding and cryptic splice sites in the PCAWG pan-cancer cohort.

PCWAG cancer code	Gene	OBS_SAMPLES	OBS_SNV	EXP_SNV	PVAL_SNV_BURDEN	Q_VALUE
meta_Digestive	CBFA2T3	6	6	1.085823261	0.000530317	0.043662784
meta_Hematopoietic	CIITA	4	4	0.162078147	1.53E-05	0.003787693
Lymph-BNHL	CIITA	4	4	0.187775328	2.81E-05	0.006929697
meta_Lymphatic	CIITA	4	4	0.206112555	3.71E-05	0.009161746
Lymph-CLL	NOTCH1	3	3	0.012776719	4.93E-07	0.000121709
meta_Hematopoietic	NOTCH1	3	3	0.108935917	0.000109741	0.013552959
meta_Lymphatic	NOTCH1	3	3	0.106952137	0.000113328	0.013995985
Kidney-RCC	PBRM1	3	3	0.100314372	8.20E-05	0.020244022
meta_Kidney	PBRM1	3	3	0.104355281	9.16E-05	0.022630276
Prost-AdenoCA	PTEN	2	2	0.037486646	0.000363654	0.08982256
Panc-AdenoCA	SMAD4	4	4	0.069937003	5.58E-07	0.00013777
meta_Digestive	SMAD4	5	5	0.437965655	5.13E-05	0.009190921
meta_Adenocarcinoma	SMAD4	5	5	0.645948075	0.000308198	0.03806249
meta_Carcinoma	SMAD4	5	5	0.771440753	0.000683982	0.059238233
PANCANCER	SMAD4	5	5	0.835384734	0.000976274	0.080379929
meta_Carcinoma	TP53	4	4	0.392226885	0.000396972	0.059238233
PANCANCER	TP53	4	4	0.415603648	0.000495085	0.080379929

Table D.13: Tumor suppressor genes with a FDR$<$0.1 significant burden of intronic cryptic splice SNVs.

Chromosome	Pos	Ref	Alt	Gene	Strand	spliceAI annot.	Delta score	Distance to exon boundary	Donor	Cohort	Notes
16	10972365	C	G	CIITA	+	DS_DG	0.6068	-1126	DO52681	MALY-DE	
16	10972968	C	G	CIITA	+	DS_DG	0.4698	-1729	DO52675	MALY-DE	* Evidence of alternative splicing in raw junc files
16	10972968	C	G	CIITA	+	DS_DG	0.4698	-1729	DO222305	DLBC-US	* Significant alternative splicing with one control but likely a low coverage artifact
18	48577503	A	G	SMAD4	+	DS_DG	0.5089	-1809	DO33392	PACA-AU	* No evidence of alternative splicing
18	48579077	C	G	SMAD4	+	DS_DG	0.341	2074	DO34736	PACA-AU	
17	7589749	C	T	TP53	-	DS_AG	0.6489	946	DO26540	LUSC-US	* No evidence of alternative splicing in raw junc files, but very low read coverage may suggest Nonsense mediated decay
17	7572139	A	T	TP53	-	DS_AG	0.5867	-419	DO9074	COAD-US	
17	7576525	A	C	TP53	-	DS_DG	0.4857	328	DO46591	OV-AU	

Significant alternative splicing cluster in carrier compared to majority of controls
Significant alternative splicing cluster in carrer compared to minority of controls or evidence of alternatie splicing in raw junction files
Low coverage in gene of interest
No significant evidence of alternative splicing

Table D.14: Evidence of cryptic splice events in RNA-seq data.

PCAWG cohort	GENE	OBS_SAMPLES	OBS_SNV	EXP_SNV	PVAL_SNV_BURDEN	QVAL_SNV_BURDEN
Lymph-BNHL	IGLL5	4	4	0.009279	4.39199E-10	0.000002
meta_Lymphatic	BTG2	4	4	0.016709	5.59702E-09	0.00002
Head-SCC	SLC7A14	3	3	0.031849	2.78859E-06	0.024839
meta_Female_repr.	RHO	3	3	0.031811	2.85351E-06	0.050835
CNS-Medullo	LHX1	2	2	0.00299	3.44882E-06	0.061441
meta_Hematopoietic	RPS25	2	2	0.008001	1.71631E-05	0.03822
meta_Hematopoietic	HIST1H2AC	2	2	0.013214	4.60815E-05	0.066642
Lymph-BNHL	ZNF48	2	2	0.015914	7.08543E-05	0.074251
meta_Hematopoietic	ADAM19	3	3	0.100254	8.5116E-05	0.089197

Bonferroni<0.05 significant corrected for all cancers
Bonferroni<0.05 significant corrected for a single cancer
FDR<0.1 significant corrected for a single cancer

Table D.15: Genes not in the CGC with a FDR<0.1 significant burden of intronic cryptic splice SNVs.

GENE	CHROM	POS	REF	ALT	PREDICTED EFFECT	SPLICEAI DELTA SCORE	N_CARRIERS
IGLL5	22	23230442	A	C	Donor loss	0.2352	2
			A	T	Donor loss	0.2348	2
LHX1	17	35299487	T	G	Acceptor gain	0.83	2
ZNF48	16	30407295	C	G	Donor gain	0.563	1
			C	T	Donor gain	0.6333	1
HIST1H2AC	6	26124982	G	A	Donor loss	0.3722	2

Table D.16: Recurrent predicted cryptic intronic splice mutations in genes in Supplementary Table table D.15.

GENE	ELEMENT	MUTATION TYPE	Observed	Expected	ENRICHMENT	CI_LOW	CI_HIGH
TP53	Coding	Synonymous	6	1.18288405	5.072348384	1.690782795	9.29930537
	Coding	Missense	543	3.532566677	153.7125976	122.2906854	188.248393
	Coding	Nonsense	104	0.242682613	428.5432675	317.2868423	552.1615177
	Coding	INDEL	141	0.321829878	438.1196701	332.4737922	556.1944749
	Splicing	Canonical	57	0.146940917	387.9110136	272.2182552	510.4092284
	Splicing	Cryptic exonic	4	0.044319716	90.2532867	22.56332168	180.5065734
	Splicing	Cryptic intronic	4	0.330592969	12.09947089	3.024867723	24.19894179
	5'UTR	INDEL	7	0.082792839	84.54837518	24.15667862	157.018411
	5'UTR	SNV	3	1.083210069	2.769545894	1.00E-16	6.462273753
ELF3	Coding	Synonymous	5	1.365059992	3.662842681	0.732568536	7.325685361
	Coding	Missense	10	3.597424017	2.779766842	1.111906737	4.725603632
	Coding	INDEL	7	0.33103758	21.14563549	6.041610139	39.2704659
	5'UTR	SNV	6	0.956931849	6.270038987	2.090012996	11.49507148
	Intron1	SNV	4	1.729358234	2.312996764	0.578249191	4.625993528
	Upstream	SNV	1	4.211543215	0.237442654	1.00E-16	0.712327963

Table D.17: Enrichment of mutations in TP53 and ELF3 5'UTRs in the PCAWG pan-cancer dataset.

ELEMENT	MUTATION TYPE	OBSERVED	POSSIBLE SNVs	ENRICHMENT Rel. to SYN SNVs	CI_LOW	CI_HIGH	P-VALUE
Synonymous	SNV	6	741	#N/A	#N/A	#N/A	#N/A
5'UTR	SNV	10	193	6.398963731	2.29721188	17.82453642	0.000387
Intron1	SNV	2	382	0.646596859	0.129886953	3.218856768	0.723539
Upstream	SNV	5	1000	0.6175	0.187741415	2.0310183	0.54391

Table D.18: Enrichment of mutations in the ELF3 5'UTR in the Hartwig Medical Foundation pan-cancer dataset.

CANCER	N_SAMPLE	N_MUTATION	Reference model	Data source reference	Data download URL	Database comparison cancers
Bladder	701	66475	Bladder-TCC_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017	Bladder Adenocarcinoma (BLCA)
Breast	3110	112916	Breast-AdenoCa_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/ https://pubmed.ncbi.nlm.nih.gov/30205045/	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017 https://www.cbioportal.org/study/summary?id=breast_msk_2018	Breast Adenocarcinoma (BRCA)
CNS	1666	39285	CNS_tumors_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/ https://www.ncbi.nlm.nih.gov/pubmed/31263031	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017 https://www.cbioportal.org/study/summary?id=glioma_mskcc_2019	High- grade glioma (HGG) Lower grade glioma (LGG) Medulloblastoma (MBL) Glioblastoma (GBM) Brain tumors
Gastroesophageal	1214	225208	Eso-AdenoCa_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/ https://pubmed.ncbi.nlm.nih.gov/29122777/ https://pubmed.ncbi.nlm.nih.gov/32437664/	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017 https://www.cbioportal.org/study/summary?id=egc_msk_2017 https://www.cbioportal.org/study/summary?id=egc_trap_msk_2020	Esogphageal adenocarcinoma (ESCA) Stomach adenocarcinoma (STAD) Gastroesophageal tumors
Head-Neck	644	59881	Head-SCC_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/ https://pubmed.ncbi.nlm.nih.gov/27442865/	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017 https://www.cbioportal.org/study/summary?id=hnc_mskcc_2016	Head-Neck cancer (HNSC)
Liver	748	46007	Liver-HCC_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/ https://pubmed.ncbi.nlm.nih.gov/30373752/	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017 https://www.cbioportal.org/study/summary?id=hcc_mskimpact_2018	Liver hepatocellular carcinoma (HC / LIHC)
Lung-NSC	2131	195178	Lung_tumors_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/ https://www.ncbi.nlm.nih.gov/pubmed/28336552 https://www.ncbi.nlm.nih.gov/pubmed/29337640	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017 https://www.cbioportal.org/study/summary?id=lung_msk_2017 https://www.cbioportal.org/study/summary?id=nsclc_pd1_msk_2018	Lung adenocarcinoma (LUAD) Lung squamous cell carcinoma (LUSC) Lung non-small cell carcinoma (NSCLC)
Ovarian	515	18347	Ovary-AdenoCA_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017	Ovarian adenocarcinoma (OV)
Pancreatic	1177	62217	Panc-AdenoCA_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017	Pancreas adenocarcinoma (PAAD)
Prostate	2112	59761	Prost-AdenoCA_SNV.Pretrained.h5	https://pubmed.ncbi.nlm.nih.gov/32015527/ https://pubmed.ncbi.nlm.nih.gov/28481359/ https://pubmed.ncbi.nlm.nih.gov/28250054/ https://www.ncbi.nlm.nih.gov/pubmed/32220891	http://www.cancer-genes.org/ https://www.cbioportal.org/study/summary?id=msk_impact_2017 https://www.cbioportal.org/study/summary?id=prad_mskcc_2017 https://www.cbioportal.org/study/summary?id=prad_mcspc_mskcc_2020	Prostate adenocarcinoma (PRAD)

Table D.19: Metadata on mega-cohorts of targeted and whole-exom sequenced samples.

CANCER	OBS_MUT	EXP_MUT	MEDIAN_BCK_MUT	CI_LOWER_BCK_MUT	CI_UPPER_BCK_MUT	N_SAMPLE	MEAN_RATE_EXCESS_MUTATIONS	CI_LOWER_RATE_EXCESS_MUTATION	CI_UPPER_RATE_EXCESS_MUTATION	P_VALUE
Bladder cancer	37	2.166005706	2	0	5	699	0.049257511	0.030042918	0.068705293	5.40E-32
CNS cancer	29	1.862493193	2	0	5	1660	0.016454819	0.009638554	0.024096386	6.00E-24
Breast cancer	33	2.324349441	2	0	6	3108	0.009908623	0.006113256	0.014157014	2.81E-26
Gastroesophageal cancer	39	3.339741681	3	0	7.025	1212	0.029688119	0.018151815	0.041254125	2.17E-27
Head-Neck cancer	20	1.74331829	2	0	5	644	0.028229814	0.01242236	0.045031056	8.08E-15
Liver cancer	34	0.778196133	1	0	3	748	0.044299465	0.028074866	0.061497326	5.64E-43
Lung-NSC cancer	48	2.906429227	3	0	7	2131	0.021106992	0.013139371	0.029563585	8.55E-40
Ovarian cancer	9	0.500226188	0	0	2	515	0.016551456	0.003883495	0.031067961	3.78E-09
Pancreatic cancer	28	1.219265847	1	0	4	1176	0.022712585	0.013584184	0.034013605	6.33E-28
Prostate cancer	24	2.610906295	2	0	6	2112	0.010114583	0.004723011	0.016098485	4.38E-15

Column name	Description
CANCER	Cancer name
OBS_MUT	Number of observed activating SNVs
EXP_MUT	Number of expected activating SNVs
MEDIAN_BCK_MUT	Median number of predicted background mutations
CI_LOWER_BCK_MUT	Lower bound of 95% confidence interval of predicted background mutations
CI_UPPER_BCK_MUT	Upper bound of 95% confidence interval of predicted background mutations
N_SAMPLE	Number of samples in the cohort
MEAN_RATE_EXCESS_MUTATIONS	Mean rate of excess mutations in simulations
CI_LOWER_RATE_EXCESS_MUTATION	Lower bound of 95% confidence interval of mutational enrichment
CI_UPPER_RATE_EXCESS_MUTATION	Upper bound of 95% confidence interval of mutational enrichment
P_VALUE	P-value of mutational burden

Table D.20: Burden of activating mutations in long-tail oncogenes in mega-cohorts.

CANCER	OBS_TRUNC	EXP_TRUNC	MEDIAN_BCK_MUT	CI_LOWER_BCK_MUT	CI_UPPER_BCK_MUT	N_SAMPLE	MEAN_RATE_EXCESS_MUTATIONS	CI_LOWER_RATE_EXCESS_MUTATION	CI_UPPER_RATE_EXCESS_MUTATION	P_VALUE
Bladder cancer	74	38.75410049	39	26	51	576	0.062277778	0.026041667	0.098958333	4.37E-07
CNS cancer	60	24.2098306	24	15	34	926	0.038485961	0.018358531	0.059395248	1.32E-09
Breast cancer	102	48.71643688	49	35	63	1760	0.030255682	0.016477273	0.044318182	3.40E-11
Gastroesoph. cancer	62	31.45281094	32	21	43	999	0.03071972	0.011986987	0.051051051	2.14E-06
Head-Neck cancer	58	26.38721463	26	17	38	492	0.06376626	0.026371951	0.101626016	1.35E-07
Liver cancer	29	12.62810469	13	6.975	20	545	0.030249541	0.00733945	0.056926606	5.86E-05
Lung-NSC cancer	170	95.97696494	96	77	117.025	1660	0.045120482	0.025903614	0.065060241	6.60E-11
Ovarian cancer	20	8.087469168	8	3	14	311	0.038138264	0.006430868	0.070819936	0.000309328
Pancreatic cancer	52	9.719740941	9	4	16	663	0.064120664	0.036199095	0.09653092	1.69E-16
Prostate cancer	42	32.03021079	32	21	45	1263	0.007807601	-0.007125891	0.021377672	0.055937754

Column name	Description
CANCER	Cancer name
OBS_TRUNC	Number of observed truncating SNVs
EXP_TRUNC	Number of expected truncating SNVs
MEDIAN_BCK_MUT	Median number of predicted background mutations
CI_LOWER_BCK_MUT	Lower bound of 95% confidence interval of predicted background mutations
CI_UPPER_BCK_MUT	Upper bound of 95% confidence interval of predicted background mutations
N_SAMPLE	Number of samples in the cohort
MEAN_RATE_EXCESS_MUTATIONS	Mean rate of excess mutations in simulations
CI_LOWER_RATE_EXCESS_MUTATION	Lower bound of 95% confidence interval of mutational enrichment
CI_UPPER_RATE_EXCESS_MUTATION	Upper bound of 95% confidence interval of mutational enrichment
P_VALUE	P-value of mutational burden

Table D.21: Burden of pLoF mutations in long-tail tumor suppressor genes.

COHORT	GENE	OBS_SYN	OBS_TRUNC	OBS_MIS	N_SAMP_SYN	N_SAMP_TRUNC	N_SAMP_MIS	EXP_SYN	EXP_TRUNC	EXP_MIS	PVAL_SYN_BURDEN	PVAL_TRUNC_BURDEN	PVAL_MIS_BURDEN	CARRIER_FREQ_Q
Bladder cancer	HORMAD1	1	4	4	1	4	4	0.453645796	0.26042868	1.63017202	0.21980935	0.000106563	0.060923634	0.011796755 0.089002873
Bladder cancer	WAC	2	5	6	2	5	4	0.609685017	0.310936325	2.334423412	0.074644848	9.95E-06	0.021150481	0.014791999 0.011946781
Bladder cancer	ZNF146	1	3	2	1	3	2	0.284013307	0.097459361	0.808007262	0.140342462	7.37E-05	0.121385555	0.00915628 0.064329724
Brain cancer	ATG5	0	3	1	0	3	1	0.086928441	0.047275899	0.279443968	0.541595219	9.09E-06	0.138057288	0.004230264 0.017464774
Brain cancer	NME8	1	4	2	1	3	2	0.420497472	0.230067836	1.071954753	0.204800655	6.04E-05	0.193659057	0.005401049 0.077767752
Brain cancer	PRSS35	0	3	1	0	3	1	0.366060905	0.051815041	0.778697481	0.646812689	2.27E-05	0.346811781	0.004223761 0.033592353
Brain cancer	SH3GL3	1	3	2	1	3	2	0.249593108	0.088178345	0.552870798	0.123534643	6.07E-05	0.064061109	0.004171664 0.077767752
Breast cancer	GGNBP2	2	5	2	2	5	2	0.799509923	0.377779913	3.075953458	0.119656268	2.70E-05	0.698907151	0.003209875 0.020779003
Breast cancer	RIC8A	2	4	9	2	2	7	0.911275916	0.151113012	2.097583564	0.150285187	1.26E-05	0.000472781	0.002672838 0.011021393
Breast cancer	TMED6	0	3	0	0	3	0	0.354912868	0.088459599	0.966874904	0.648413714	6.24E-05	0.805999995	0.002021903 0.044364771
Esogastric cancer	ARRB1	0	4	4	0	3	4	1.45939122	0.289617754	4.315957194	0.879502684	0.000147287	0.518611861	0.004486557 0.074804198
Esogastric cancer	FERMT1	2	5	6	2	3	6	2.168739405	0.528628375	6.926804907	0.497561189	0.000155761	0.594875804	0.005406737 0.074804198
Esogastric cancer	GIGYF2	3	9	7	3	8	7	2.972616685	1.114590365	9.86247694	0.455988551	2.36E-06	0.793392034	0.009534957 0.002386066
Esogastric cancer	MXD1	0	3	6	0	3	6	0.422744235	0.105471167	1.454639877	0.668302446	0.000133518	0.006127792	0.003500035 0.071246548
Esogastric cancer	PARD3	7	9	12	7	9	12	4.414198928	1.075150837	13.86207304	0.124216659	1.57E-06	0.660823253	0.009582647 0.001778056
Esogastric cancer	PDZD8	6	5	9	6	4	9	3.948402937	0.527544085	11.79204716	0.171106032	0.000171147	0.726793644	0.005408048 0.078533524
Esogastric cancer	PGM5	0	6	29	0	6	29	2.248883379	0.456658262	6.295446619	0.943012591	6.72E-06	7.25E-09	0.006702953 0.005381234
Head-neck cancer	CSNK2A1	0	4	2	0	4	2	0.355419531	0.202837883	1.163654364	0.647986076	4.50E-05	0.220623043	0.008934499 0.046025112
Head-neck cancer	RSPH9	0	3	1	0	3	1	0.384669657	0.079228534	0.91778419	0.657668178	4.95E-05	0.411190368	0.006872403 0.04754281
Head-neck cancer	STC1	1	3	1	1	3	1	0.486633789	0.103011732	1.103866268	0.233567813	0.000118144	0.473421803	0.006816443 0.089243116
Head-neck cancer	SUFI1	1	5	5	1	5	5	1.456643895	0.50529092	4.198421685	0.592581373	0.000120787	0.330427124	0.010575786 0.089243116
Head-neck cancer	TANC1	1	6	8	1	6	8	2.005924154	0.47964967	4.821250592	0.724982275	7.76E-06	0.0956153	0.01298906 0.012427538
Head-neck cancer	ZNF572	1	4	1	1	3	1	0.444117057	0.17951164	1.695056502	0.21394823	4.03E-05	0.628969624	0.008989384 0.045534567
NSC Lung cancer	ATAD5	2	8	12	2	7	12	2.68065192	1.311197668	10.14720907	0.618374537	5.75E-05	0.283580505	0.010876101 0.038056536
NSC Lung cancer	DSN1	1	5	3	1	5	3	0.580019588	0.362003595	1.975008758	0.27551323	3.82E-05	0.233840635	0.007541458 0.028251991
NSC Lung cancer	PEAK1	6	8	7	6	7	7	3.67840896	1.03112092	12.02761137	0.136041357	1.52E-05	0.892334724	0.011331511 0.012693087
NSC Lung cancer	RIC3	1	5	6	1	5	6	1.25454993	0.53824873	4.104898106	0.532595272	0.000158274	0.184585052	0.00725488 0.089425066
NSC Lung cancer	ZNF236	5	9	26	5	9	26	5.762840641	1.742908301	15.49227443	0.593054869	6.49E-05	0.011965336	0.011800149 0.041574123
Pancreatic cancer	MYO9B	2	4	12	1	4	12	1.750067459	0.238736174	3.586110186	0.387308317	6.70E-05	0.000484462	0.005328674 0.055939859

Table D.22: Genes not in driver gene databases with a FDR<0.1 significant burden of pLoF mutations in exome-sequenced samples.

CHROM	START	END	ELT	ELT_SIZE	OBS_SAMPLES	OBS_SNV	OBS_INDEL	EXP_SNV	EXP_INDEL	PVAL_SNV_BURDEN	PVAL_INDEL_BURDEN	PVAL_MUT_BURDEN
8	29952426	29953028	LEPROTL1_3758_promoter_23	602	14	14	0	3.35496809	4 0.172348891	1.35E-05	0.579079316	9.95E-05
12	12542268	2 12542576	1 NCOR2_6045_intergenic_26	3079	35	32	3	13.5478951	1 0.97232299	7.27E-05	0.04740057	4.68E-05
16	50729275	50730237	CYLD_7082_intergenic_17	962	14	14	0	4.08123534	0.321619887	7.31E-05	0.637429865	0.00051134
19	45981012	45982719	BCL3,ERCC2_8247_intergenic_72	1707	25	25	0	10.3756913	2 0.731958049	0.000118388	0.759042311	0.000927124
3	8485478	8486428	SRGAP3_1330_genic_5	950	13	13	0	3.74091310	9 0.293013393	0.000126534	0.626862424	0.000828255
12	125001659	125004288	NCOR2_6016_genic_84	2629	24	16	8	9.89303786	2 0.679885439	0.046646304	4.85E-07	4.21E-07
8	128753771	128756221	MYC_4137_genic,intergenic_34	2450	32	24	8	11.5675811	2 0.958130973	0.001413155	5.31E-06	1.48E-07
14	38059382	38060274	FOXA1_6296_genic_5	892	10	5	5	3.79130261	1 0.351934715	0.258356738	1.90E-05	6.49E-05
12	69195077	69196107	MDM2_5816_intergenic_9	1030	9	4	5	5.02138832	0.379383264	0.645363656	2.75E-05	0.000212058
17	54992281	54993673	HLF,MSI2,RNF43_7542_intergenic_16	1392	12	7	5	6.20253606	1 0.513335778	0.356902711	0.000118733	0.000469059

Table D.23: ABC enhancer elements with a FDR<0.1 significant burden of mutations in the PCAWG pan-cancer cohort.

Cancer type	N_SAMPLES	N_MUTATIONS	Reference Model
Bladder	317	61478	Bladder-TCC_SNV.Pretrained.h5
Brain	698	31788	CNS_tumors_SNV.Pretrained.h5
Breast	1442	105773	Breast_tumors_SNV.Pretrained.h5
Colorectal	223	75041	ColoRect-AdenoCA_SNV.Pretrained.h5
Endometrium	327	191117	Uterus-AdenoCA_SNV_msi_low.Pretrained.h5
Gastroesophageal	831	217561	Eso-AdenoCa_SNV.Pretrained.h5
HeadNeck	425	58000	Head-SCC_SNV.Pretrained.h5
KidneyClear	412	20979	Kidney-RCC_SNV.Pretrained.h5
Liver	407	44758	Liver-HCC_SNV.Pretrained.h5
LungAD	445	128816	Lung-AdenoCA_SNV.Pretrained.h5
LungSC	172	55297	Lung-SCC_SNV.Pretrained.h5
Lymph	184	22261	Lymph_tumors_SNV.Pretrained.h5
Ovarian	316	17694	Ovary-AdenoCA_SNV.Pretrained.h5
Pancreas	714	60660	Panc-AdenoCA_SNV.Pretrained.h5
Prostate	878	56029	Prost-AdenoCA_SNV.Pretrained.h5
Sarcoma	247	17096	Sarcoma_tumors_SNV.Pretrained.h5
Skin	579	398335	Skin-Melanoma_SNV.Pretrained.h5
Pancan	7569	1132911	Pancan_SNV.Pretrained.h5

Table D.24: Metadata about whole-exome sequenced cohorts from Dietlein et al. 2019 Nat. Genet..

Bibliography

[1]

[2] Nezar Abdennur. Python bindings to UCSC BigWig and BigBed library. contribute to nvictus/pybbi development by creating an account on GitHub, 2018.

[3] George Adam, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, 4(1):1–10, 2020.

[4] Henry Hu Bennett Stankovits Sue Felshin Yevgeni Berzak Helena Aparicio Boris Katz Ignacio Cases Adam U. Yaari, Jan DeWitt and Andrei Barbu. The aligned multimodal movie treebank: An audio, video, dependency-parse treebank. *The Conference on Empirical Methods in Natural Language Processing*, 2022.

[5] Christoph Adami. What is complexity? *BioEssays*, 24(12):1085–1094, 2002.

[6] Christoph Adami. The use of information theory in evolutionary biology. *Annals of the New York Academy of Sciences*, 1256(1):49–65, 2012.

[7] Christoph Adami, Charles Ofria, and Travis C Collier. Evolution of biological complexity. *Proceedings of the National Academy of Sciences*, 97(9):4463–4468, 2000.

[8] Lars Ahrenberg. Lines: An English-Swedish Parallel Treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, 2007.

[9] Kadir C. Akdemir, Victoria T. Le, Justin M. Kim, Sarah Killcoyne, Devin A. King, Ya-Ping Lin, Yanyan Tian, Akira Inoue, Samirkumar B. Amin, Frederick S. Robinson, Manjunath Nimmakayalu, Rafael E. Herrera, Erica J. Lynn, Kin Chan, Sahil Seth, Leszek J. Klimczak, Moritz Gerstung, Dmitry A. Gordenin, John O'Brien, Lei Li, Yonathan Lissanu Deribe, Roel G. Verhaak, Peter J. Campbell, Rebecca Fitzgerald, Ashby J. Morrison, Jesse R. Dixon, and P. Andrew Futreal. Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nature Genetics*, 52(11):1178–1188, 2020.

[10] Ludmil B. Alexandrov, Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R. Covington, Dmitry A. Gordenin, Erik N. Bergstrom, S. M. Ashiqul Islam, Nuria Lopez-Bigas, Leszek J. Klimczak, John R. McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A. Wheeler, Ville Mustonen, Gad Getz, Steven G. Rozen, and Michael R. Stratton. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

[11] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, Icgc PedBrain, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[12] Fiorenzo Artoni, Piergiorgio d'Orio, Eleonora Catricalà, Francesca Conca, Franco Bottoni, Veronica Pelliccia, Ivana Sartori, Giorgio Lo Russo, Stefano F Cappa, Silvestro Micera, and Andrea Moro. High gamma response tracks different syntactic structures in homophonous phrases. *Scientific Reports*, 10(1):1–10, 2020.

[13] Farzaneh Atashrazm and Sarah Ellis. The polarity protein PARD3 and cancer. *Oncogene*, 40(25):4245–4262, 2021.

[14] 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. *Nature Methods*, 18(10):1196–1203, 2021.

[15] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021.

[16] Taejeong Bae, Livia Tomasini, Jessica Mariani, Bo Zhou, Tanmoy Roychowdhury, Daniel Franjic, Mihovil Pletikos, Reenal Pattni, Bo-Juen Chen, Elisa Venturini, Bridget Riley-Gillis, Nenad Sestan, Alexander E. Urban, Alexej Abyzov, and Flora M. Vaccarino. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science*, 359(6375):550–555, 2018.

[17] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[18] Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C. Wendl, Jaegil Kim, Brendan Reardon, Patrick Kwok-Shing Ng, Kang Jin Jeong, Song Cao, Zixing Wang, Jianjiong Gao, Qingsong Gao, Fang Wang, Eric Minwei Liu, Loris Mularoni, Carlota Rubio-Perez, Niranjan Nagarajan, Isidro Cortés-Ciriano, Daniel Cui Zhou, Wen-Wei Liang, Julian M. Hess, Venkata D. Yellapantula, David Tamborero, Abel Gonzalez-Perez, Chayaporn Suphavilai, Jia Yu Ko, Ekta Khurana, Peter J. Park, Eliezer M. Van Allen, Han Liang, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau,

D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J. M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C. S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella

Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R.

Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson,

Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, Michael S. Lawrence, Adam Godzik, Nuria Lopez-Bigas, Josh Stuart, David Wheeler, Gad Getz, Ken Chen, Alexander J. Lazar, Gordon B. Mills, Rachel Karchin, and Li Ding. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.e18, 2018.

[19] Doris-Eva Bamiou, Frank E Musiek, and Linda M Luxon. The insula (island of reil) and its role in auditory processing: literature review. *Brain research reviews*, 42(2):143–154, 2003.

[20] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[21] Sallie Baxendale. The wada test. *Current opinion in neurology*, 22(2):185–189, 2009.

[22] Douglas K Bemis and Liina Pylkkänen. Flexible composition: Meg evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PloS one*, 8(9):e73949, 2013.

[23] Douglas K. Bemis and Liina Pylkkänen. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, 31(8):2801–2814, 2011.

[24] Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[25] Manuela Berlingeri, Davide Crepaldi, Rossella Roberti, Giuseppe Scialfa, Claudio Luzzatti, and Eraldo Paulesu. Nouns and verbs in the brain: Grammatical class and task specific effects as revealed by fmri. *Cognitive neuropsychology*, 25(4):528–558, 2008.

[26] Johanna Bertl, Qianyun Guo, Malene Juul, Søren Besenbacher, Morten Muhlig Nielsen, Henrik Hornshøj, Jakob Skou Pedersen, and Asger Hobolth. A site specific model and analysis of the neutral somatic mutation rate in whole-genome cancer data. *BMC Bioinformatics*, 19(1):147, 2018.

[27] Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. Anchoring and agreement in syntactic annotations. *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016*, 2016.

[28] Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. The Alice Datasets: fMRI & EEG observations of natural language comprehension. In *Language Resources and Evaluation Conference (LREC)*, pages 120–125, 2020.

[29] Shohini Bhattasali and Philip Resnik. Using surprisal and fMRI to map the neural bases of broad and local contextual prediction during natural language comprehension. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3786–3798, Online, August 2021. Association for Computational Linguistics.

[30] Deniz Bilecen, Erich Seifritz, Klaus Scheffler, Jürgen Henning, and Anja-Carina Schulte. Amplitopicity of the human auditory cortex: An fmri study. *NeuroImage*, 17(2):710–718, 2002.

[31] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[32] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[33] Esti Blanco-Elorrieta, Itamar Kastner, Karen Emmorey, and Liina Pylkkänen. Shared neural correlates for building phrases in signed and spoken language. *Scientific reports*, 8(1):1–10, 2018.

[34] Esti Blanco-Elorrieta and Liina Pylkkänen. Bilingual language switching in the laboratory versus in the wild: The spatiotemporal dynamics of adaptive language control. *Journal of Neuroscience*, 37(37):9022–9036, 2017.

[35] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

[36] John Bradshaw, Alexander G. de G. Matthews, and Zoubin Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv:1707.02476 [stat]*, 2017.

[37] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[38] Jonathan Brennan. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313, 2016.

[39] Jonathan Brennan and Liina Pylkkänen. The time-course and spatial distribution of brain activity associated with sentence processing. *NeuroImage*, 60(2):1139–1148, 2012.

[40] Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157-158:81–94, 2016.

[41] Hale JT Brennan JR. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS ONE*, 2019.

[42] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8):1157–1166, 1997. Papers from the Sixth International World Wide Web Conference.

[43] Harm Brouwer, Francesca Delogu, Noortje J. Venhuizen, and Matthew W. Crocker. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12, 2021.

[44] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[45] Nithin Buduma, Nikhil Buduma, and Joe Papa. *Fundamentals of deep learning.* " O'Reilly Media, Inc.", 2022.

[46] Miriam Butt. The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective*, pages 48–78, 2010.

[47] Andrew Caines, Michael McCarthy, and Paula Buttery. Parsing transcripts of speech. *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2017*, 2017.

[48] Claudia Calabrese, Natalie R. Davidson, Deniz Demircioğlu, Nuno A. Fonseca, Yao He, André Kahles, Kjong-Van Lehmann, Fenglin Liu, Yuichi Shiraishi, Cameron M. Soulette, Lara Urban, Liliana Greger, Siliang Li, Dongbing Liu, Marc D. Perry, Qian Xiang, Fan Zhang, Junjun Zhang, Peter Bailey, Serap Erkek, Katherine A. Hoadley, Yong Hou, Matthew R. Huska, Helena Kilpinen, Jan O. Korbel, Maximillian G. Marin, Julia Markowski, Tannistha Nandi, Qiang Pan-Hammarström, Chandra Sekhar Pedamallu, Reiner Siebert, Stefan G. Stark, Hong Su, Patrick Tan, Sebastian M. Waszak, Christina Yung, Shida Zhu, Philip Awadalla, Chad J. Creighton, Matthew Meyerson, B. F. Francis Ouellette, Kui Wu, Huanming Yang, Alvis Brazma, Angela N. Brooks, Jonathan Göke, Gunnar Rätsch, Roland F. Schwarz, Oliver Stegle, and Zemin Zhang. Genomic basis for RNA alterations in cancer. *Nature*, 578(7793):129–136, 2020.

[49] Jason Calaiaro. Ai takes a dumpster dive: Computer-vision systems sort your recyclables at superhuman speed. *IEEE Spectrum*, 59(7):22–27, 2022.

[50] Ian M. Campbell, Chad A. Shaw, Pawel Stankiewicz, and James R. Lupski. Somatic mosaicism: implications for disease and transmission genetics. *Trends in Genetics*, 31(7):382–392, 2015.

[51] Peter J. Campbell, Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, Hardeep K. Nahal-Bose, B. F. Francis Ouellette, Constance H. Li, Esther Rheinbay, G. Petur Nielsen, Dennis C. Sgroi, Chin-Lee Wu, William C. Faquin, Vikram Deshpande, Paul C. Boutros, Alexander J. Lazar, Katherine A. Hoadley, David N. Louis, L. Jonathan Dursi, Christina K. Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Junjun Zhang, Wenyi Wang, David A. Wheeler, Li Ding, Jared T. Simpson, Brian D. O'Connor, Sergei Yakneen, Kyle Ellrott, Naoki Miyoshi, Adam P. Butler, Romina Royo, Solomon I. Shorser, Miguel Vazquez, Tobias Rausch, Grace Tiao, Sebastian M. Waszak, Bernardo Rodriguez-Martin, Suyash Shringarpure, Dai-Ying Wu, German M. Demidov, Olivier Delaneau, Shuto Hayashi, Seiya Imoto, Nina Habermann, Ayellet V. Segre, Erik Garrison, Andy Cafferkey, Eva G. Alvarez, José María Heredia-Genestar, Francesc Muyas, Oliver Drechsel, Alicia L. Bruzos, Javier Temes, Jorge Zamora, Adrian Baez-Ortega, Hyung-Lae Kim, R. Jay Mashl, Kai Ye, Anthony DiBiase, Kuan-lin Huang, Ivica Letunic, Michael D. McLellan, Steven J. Newhouse, Tal Shmaya, Sushant Kumar, David C. Wedge, Mark H. Wright, Venkata D. Yellapantula, Mark Gerstein, Ekta Khurana, Tomas Marques-Bonet, Arcadi Navarro, Carlos D. Bustamante, Reiner Siebert, Hidewaki Nakagawa, Douglas F. Easton, Stephan Ossowski, Jose M. C. Tubio, Francisco M. De La Vega, Xavier Estivill, Denis Yuen, George L. Mihaiescu, Larsson Omberg, Vincent Ferretti, Radhakrishnan Sabarinathan, Oriol Pich, Abel Gonzalez-Perez, Amaro Taylor-Weiner, Matthew W. Fittall, Jonas Demeulemeester, Maxime Tarabichi, Nicola D. Roberts, Peter Van Loo, Isidro Cortés-Ciriano, Lara Urban, Peter Park, Bin Zhu, Esa Pitkänen, Yilong Li, Natalie Saini, Leszek J. Klimczak, Joachim Weischenfeldt, Nikos Sidiropoulos, Ludmil B. Alexandrov, Raquel Rabionet, Georgia Escaramis, Mattia Bosio, Aliaksei Z. Holik, Hana Susak, Aparna Prasad, Serap Erkek, Claudia Calabrese, Benjamin Raeder, Eoghan Harrington, Simon Mayes, Daniel Turner, Sissel Juul, Steven A. Roberts, Lei Song, Roelof Koster, Lisa Mirabello, Xing Hua, Tomas J. Tanskanen, Marta Tojo, Jieming Chen, Lauri A. Aaltonen, Gunnar Rätsch, Roland F. Schwarz, Atul J. Butte, Alvis Brazma, Stephen J. Chanock, Nilanjan Chatterjee, Oliver Stegle, Olivier Harismendy, G. Steven Bova, Dmitry A. Gordenin, David Haan, Lina Sieverling, Lars Feuerbach, Don Chalmers, Yann Joly, Bartha Knoppers, Fruzsina Molnár-Gábor, Mark Phillips, Adrian Thorogood, David Townend, Mary Goldman, Nuno A. Fonseca, Qian Xiang, Brian Craft, Elena Piñeiro-Yáñez, Alfonso Muñoz, Robert Petryszak, Anja Füllgrabe, Fatima Al-Shahrour, Maria Keays, David Haussler, John Weinstein, Wolfgang Huber, Alfonso Valencia, Irene Papatheodorou, Jingchun Zhu, Yu Fan, David Torrents, Matthias Bieg, Ken Chen, Zechen Chong, Kristian Cibulskis, Roland Eils, Robert S. Fulton, Josep L. Gelpi, Santiago Gonzalez, Ivo G. Gut, Faraz Hach, Michael Heinold, Taobo Hu, Vincent Huang, Barbara Hutter, Natalie Jäger, Jongsun Jung, Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, Ana Milovanovic, Morten Muhlig Nielsen, Nagarajan Paramasivam, Jakob Skou

Pedersen, Montserrat Puiggròs, S. Cenk Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Jeremiah A. Wala, Jiayin Wang, Michael Wendl, Johannes Werner, Zhenggang Wu, Hong Xue, Takafumi N. Yamaguchi, Venkata Yellapantula, Brandi N. Davis-Dusenbery, Robert L. Grossman, Youngwook Kim, Michael C. Heinold, Jonathan Hinton, David R. Jones, Andrew Menzies, Lucy Stebbings, Julian M. Hess, Mara Rosenberg, Andrew J. Dunford, Manaswi Gupta, Marcin Imielinski, Matthew Meyerson, Rameen Beroukhim, Jüri Reimand, Priyanka Dhingra, Francesco Favero, Stefan Dentro, Jeff Wintersinger, Vasilisa Rudneva, Ji Wan Park, Eun Pyo Hong, Seong Gu Heo, André Kahles, Kjong-Van Lehmann, Cameron M. Soulette, Yuichi Shiraishi, Fenglin Liu, Yao He, Deniz Demircioğlu, Natalie R. Davidson, Liliana Greger, Siliang Li, Dongbing Liu, Stefan G. Stark, Fan Zhang, Samirkumar B. Amin, Peter Bailey, Aurélien Chateigner, Milana Frenkel-Morgenstern, Yong Hou, Matthew R. Huska, Helena Kilpinen, Fabien C. Lamaze, Chang Li, Xiaobo Li, Xinyue Li, Xingmin Liu, Maximillian G. Marin, Julia Markowski, Tannistha Nandi, Akinyemi I. Ojesina, Qiang Pan-Hammarström, Peter J. Park, Chandra Sekhar Pedamallu, Hong Su, Patrick Tan, Bin Tean Teh, Jian Wang, Heng Xiong, Chen Ye, Christina Yung, Xiuqing Zhang, Liangtao Zheng, Shida Zhu, Philip Awadalla, Chad J. Creighton, Kui Wu, Huanming Yang, Jonathan Göke, Zemin Zhang, Angela N. Brooks, Matthew W. Fittall, Iñigo Martincorena, Carlota Rubio-Perez, Malene Juul, Steven Schumacher, Ofer Shapira, David Tamborero, Loris Mularoni, Henrik Hornshøj, Jordi Deu-Pons, Ferran Muiños, Johanna Bertl, Qianyun Guo, Abel Gonzalez-Perez, Qian Xiang, and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pancancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020.

[52] Song Cao, Daniel Cui Zhou, Clara Oh, Reyka G. Jayasinghe, Yanyan Zhao, Christopher J. Yoon, Matthew A. Wyczalkowski, Matthew H. Bailey, Terrence Tsou, Qingsong Gao, Andrew Malone, Sheila Reynolds, Ilya Shmulevich, Michael C. Wendl, Feng Chen, and Li Ding. Discovery of driver non-coding splice-site-creating mutations in cancer. *Nature Communications*, 11(1):5573, 2020.

[53] Marinella Cappelletti, Felipe Fregni, Kevin Shapiro, Alvaro Pascual-Leone, and Alfonso Caramazza. Processing nouns and verbs in the left frontal cortex: A transcranial magnetic stimulation study. *Journal of Cognitive Neuroscience*, 20(4):707–720, 2008.

[54] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012.

[55] Michael Wy Chan, Yi-Wen Huang, Corinna Hartman-Frey, Chieh-Ti Kuo, Daniel Deatherage, Huaxia Qin, Alfred Sl Cheng, Pearlly S. Yan, Ramana V. Davuluri, Tim H.-M. Huang, Kenneth P. Nephew, and Huey-Jen L. Lin. Aberrant transforming growth factor beta1 signaling and SMAD4 nuclear translocation confer epigenetic repression of ADAM19 in ovarian cancer. *Neoplasia (New York, N.Y.)*, 10(9):908–919, 2008.

[56] Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. BLLIP 1987-89 WSJ Corpus Release 1. *Linguistic Data Consortium*, 2000.

[57] Shanyu Chen, Zhipeng He, Xinyin Han, Xiaoyu He, Ruilin Li, Haidong Zhu, Dan Zhao, Chuangchuang Dai, Yu Zhang, Zhonghua Lu, et al. How big data and high-performance computing drive brain science. *Genomics, proteomics & bioinformatics*, 17(4):381–392, 2019.

[58] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

[59] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface*, 15(141):20170387, 2018.

[60] Noam Chomsky. *Syntactic structures*. De Gruyter Mouton, 2009.

[61] Héctor Climente-González, Eduard Porta-Pardo, Adam Godzik, and Eduardo Eyras. The functional impact of alternative splicing in cancer. *Cell Reports*, 20(9):2215–2226, 2017.

[62] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[63] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, et al. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.

[64] The ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[65] Kelsy C. Cotto, Yang-Yang Feng, Avinash Ramu, Zachary L. Skidmore, Jason Kunisaki, Megan Richters, Sharon Freshour, Yiing Lin, William C. Chapman, Ravindra Uppaluri, Ramaswamy Govindan, Obi L. Griffith, and Malachi Griffith. RegTools: Integrated analysis of genomic and transcriptomic data for the discovery of splicing variants in cancer. *bioRxiv*, page 436634, 2021.

[66] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

[67] Anders Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179 – 194, 1999.

[68] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07 2021.

[69] Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968 – 980, 2006.

[70] Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, 2010.

[71] Felix Dietlein, Donate Weghorn, Amaro Taylor-Weiner, André Richters, Brendan Reardon, David Liu, Eric S. Lander, Eliezer M. Van Allen, and Shamil R. Sunyaev. Identification of cancer driver genes based on nucleotide context. *Nature Genetics*, pages 1–11, 2020.

[72] F. T. Durso and C. S. O'Sullivan. Naming and remembering proper and common nouns and pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(3):497–510, 1983.

[73] Kerryn Elliott and Erik Larsson. Non-coding driver mutations in human cancer. *Nature Reviews Cancer*, pages 1–10, 2021.

[74] Haitham A. Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, Taylor E. Arnoff, Jihye Park, William C. Hahn, and Eliezer M. Van Allen. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021.

[75] Katey S. S. Enfield, Erin A. Marshall, Christine Anderson, Kevin W. Ng, Sara Rahmati, Zhaolin Xu, Megan Fuller, Katy Milne, Daniel Lu, Rocky Shi, David A. Rowbotham, Daiana D. Becker-Santos, Fraser D. Johnson, John C. English, Calum E. MacAulay, Stephen Lam, William W. Lockwood, Raj Chari, Aly Karsan, Igor Jurisica, and Wan L. Lam. Epithelial tumor suppressor ELF3 is a lineage-specific amplified oncogene in lung adenocarcinoma. *Nature Communications*, 10(1):5438, 2019.

[76] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.

[77] Yasmeen Faroqi-Shah, Rajani Sebastian, and Ashlyn Vander Woude. Neural representation of word categories is distinct in the temporal lobe: An activation likelihood analysis. *Human brain mapping*, 39(12):4925–4938, 2018.

[78] Kara D. Federmeier, Jessica B. Segal, Tania Lombrozo, and Marta Kutas. Brain responses to nouns, verbs and class-ambiguous words in context. *Brain*, 123(12):2552–2566, 2000.

[79] Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, page 201112937, 2011.

[80] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

[81] Evelina Fedorenko, Alfonso Nieto-Castanon, and Nancy Kanwisher. Lexical and syntactic representations in the brain: an fmri investigation with multi-voxel pattern analyses. *Neuropsychologia*, 50(4):499–513, 2012.

[82] Evelina Fedorenko, Terri L Scott, Peter Brunner, William G Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262, 2016.

[83] B. Fischl, A. Liu, and A. M. Dale. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Medical Imaging*, 20(1):70–80, Jan 2001.

[84] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355, 2002.

[85] Bruce Fischl and Anders M. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):11050–11055, 2000.

[86] Bruce Fischl, David H. Salat, André J.W. van der Kouwe, Nikos Makris, Florent Ségonne, Brian T. Quinn, and Anders M. Dale. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23(Supplement 1):S69 – S84, 2004. Mathematics in Brain Imaging.

[87] Bruce Fischl, Martin I. Sereno, and Anders Dale. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195 – 207, 1999.

[88] Bruce Fischl, Martin I. Sereno, Roger B.H. Tootell, and Anders M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–284, 1999.

[89] Bruce Fischl, André van der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, David H. Salat, Evelina Busa, Larry J. Seidman, Jill Goldstein, David Kennedy, Verne Caviness, Nikos Makris, Bruce Rosen, and Anders M. Dale. Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, 14(1):11–22, 2004.

[90] ALEXANDER FLEMING. THE DISCOVERY OF PENICILLIN. *British Medical Bulletin*, 2(1):4–5, 01 1944.

[91] Lars A. Forsberg, David Gisselsson, and Jan P. Dumanski. Mosaicism in health and disease — clones picking up speed. *Nature Reviews Genetics*, 18(2):128–142, 2017.

[92] Thomas S Frank, Amie M Deffenbaugh, Julia E Reid, Mark Hulick, Brian E Ward, Beth Lingenfelter, Kathi L Gumpper, Thomas Scholl, Sean V Tavtigian, Dmitry R Pruss, et al. Clinical characteristics of individuals with germline mutations in brca1 and brca2: analysis of 10,000 individuals. *Journal of Clinical Oncology*, 20(6):1480–1490, 2002.

[93] Nils J. Fredriksson, Lars Ny, Jonas A. Nilsson, and Erik Larsson. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics*, 46(12):1258–1263, 2014.

[94] Yves Frégnac. Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain? *Science*, 358(6362):470–477, 2017.

[95] Florian Fuchs, Yunlong Song, Elia Kaufmann, Davide Scaramuzza, and Peter Dürr. Super-human performance in gran turismo sport using deep reinforcement learning. *IEEE Robotics and Automation Letters*, 6(3):4257–4264, 2021.

[96] Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. Predicting 3d genome folding from DNA sequence with akita. *Nature Methods*, 17(11):1111–1117, 2020.

[97] Robert G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.

[98] Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. *arXiv:1809.11165 [cs, stat]*, 2019.

[99] Levi A. Garraway. Genomics-driven oncology: framework for an emerging paradigm. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 31(15):1806–1814, 2013.

[100] Murray Gell-Mann. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Macmillan, 1995.

[101] Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, Ellen T. Gelfand, Craig M. Bielski, Haoxin Li, Kevin Hu, Alexander Y. Andreev-Drakhlin, Jaegil Kim, Julian M. Hess, Brian J. Haas, François Aguet, Barbara A. Weir, Michael V. Rothberg, Brenton R. Paolella, Michael S. Lawrence, Rehan Akbani, Yiling Lu, Hong L. Tiv, Prafulla C. Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M. Korn, Dale A. Porter, Michael D. Jones, Javad Golji, Giordano Caponigro, Jordan E. Taylor, Caitlin M. Dunning, Amanda L. Creech, Allison C. Warren, James M. McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E. Maruvka, Andrew D. Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D. Jaffe, Andrew A. Lane, David M. Weinstock, Cory M. Johannessen, Michael P. Morrissey, Frank Stegmeier, Robert Schlegel, William C. Hahn, Gad Getz, Gordon B. Mills, Jesse S. Boehm, Todd R. Golub, Levi A. Garraway, and William R. Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, 2019.

[102] Judy W Gichoya, Siddhartha Nuthakki, Pallavi G Maity, and Saptarshi Purkayastha. Phronesis of ai in radiology: Superhuman meets natural stupidity. *arXiv preprint arXiv:1803.11244*, 2018.

[103] Brian S. Gloss and Marcel E. Dinger. Realizing the significance of noncoding functionality in clinical genomics. *Experimental & Molecular Medicine*, 50(8), 2018.

[104] John J Godfrey, Edward C Holliman, and Jane McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 1992.

[105] Ariel Goldstein, Zaid Kokaja Zada, Eliav Buchnik, Mariano Schain, Amy Rose Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Fanda Lora, Adeen Flinker, Sasha Devore, Werner K. Doyle, Daniel Friedman, Patricia Dugan, Avinatan Hassidim, Michael P. Brenner, Y. Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv*, 2020.

[106] Abel Gonzalez-Perez, Radhakrishnan Sabarinathan, and Nuria Lopez-Bigas. Local determinants of the mutational landscape of the human genome. *Cell*, 177(1):101–114, 2019.

[107] Google. Speech-to-text: Automatic speech recognition — Google Cloud, 2020.

[108] Joe G Greener, Shaun M Kandathil, Lewis Moffat, and David T Jones. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55, 2022.

[109] David M. Groppe, Stephan Bickel, Andrew R. Dykstra, Xiuyuan Wang, Pierre Mégevand, Manuel R. Mercier, Fred A. Lado, Ashesh D. Mehta, and Christopher J. Honey. ielvis: An open source matlab toolbox for localizing and visualizing human intracranial electrode data. *Journal of Neuroscience Methods*, 281:40–48, 2017.

[110] Dandan Guo, Bo Chen, Hao Zhang, and Mingyuan Zhou. Deep poisson gamma dynamical systems. In *Advances in Neural Information Processing Systems 31*, pages 8442–8452. Curran Associates, Inc., 2018.

[111] Thilo Hagendorff and Katharina Wezel. 15 challenges for ai: or what ai (currently) can't do. *AI & SOCIETY*, 35(2):355–365, 2020.

[112] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[113] Liberty S. Hamilton and Alexander G. Huth. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, 2020.

[114] Liberty S. Hamilton, Yulia Oganian, and Edward F. Chang. Topography of speech-related acoustic and phonological feature encoding throughout the human core and parabelt auditory cortex. *bioRxiv*, 2020.

[115] Liberty S. Hamilton, Yulia Oganian, Jeffery Hall, and Edward F. Chang. Parallel and distributed encoding of speech across human auditory cortex. *Cell*, 184(18):4626–4639.e13, 2021.

[116] Xiao Han, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Paul Maguire, Diana Rosas, Nikos Makris, Anders Dale, Bradford Dickerson, and Bruce Fischl. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194, 2006.

[117] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[118] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[119] Viktória Havas, Andreu Gabarrós, Montserrat Juncadella, Xavi Rifa-Ros, Gerard Plans, Juan José Acebes, Ruth de Diego Balaguer, and Antoni Rodríguez-Fornells. Electrical stimulation mapping of nouns and verbs in broca's area. *Brain and Language*, 145-146:53–63, 2015.

[120] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, page 187–197, USA, 2011. Association for Computational Linguistics.

[121] Brian Hie, Bryan D. Bryson, and Bonnie Berger. Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Systems*, 11(5):461–477.e9, 2020.

[122] Brian Hie, Ellen D. Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, 2021.

[123] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, Alison M. Taylor, Andrew D. Cherniack, Vésteinn Thorsson, Rehan Akbani, Reanne Bowlby, Christopher K. Wong, Maciej Wiznerowicz, Francisco Sanchez-Vega, A. Gordon Robertson, Barbara G. Schneider, Michael S. Lawrence, Houtan Noushmehr, Tathiane M. Malta, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger

Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J. M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C. S. Silveira,

Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, olanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Es-

maeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, An-

dreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, Joshua M. Stuart, Christopher C. Benz, and Peter W. Laird. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018.

[124] Gerard Hoyne, Caroline Rudnicka, Qing-Xiang Sang, Mark Roycik, Sarah Howarth, Peter Leedman, Markus Schlaich, Patrick Candy, and Vance Matthews. Genetic and cellular studies highlight that a disintegrin and metalloproteinase 19 is a protective biomarker in human prostate cancer. *BMC cancer*, 16:151, 2016.

[125] Feng-hsiung Hsu. Ibm's deep blue chess grandmaster chips. *IEEE micro*, 19(2):70–81, 1999.

[126] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[127] Minho Hwang, Brijen Thananjeyan, Daniel Seita, Jeffrey Ichnowski, Samuel Paradis, Danyal Fer, Thomas Low, and Ken Goldberg. Superhuman surgical peg transfer using depth-sensing and deep recurrent neural networks. *arXiv preprint arXiv:2012.12844*, 2020.

[128] 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, Eric D. Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J. Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, 2019.

[129] Sharmistha Jat, Erika J C Laing, Partha Talukdar, and Tom Mitchell. Spatio-temporal characteristics of noun and verb processing during sentence comprehension in the brain. In *34th Conference on Neural Information Processing Systems*, 2020.

[130] Eshin Jolly. Pymer4: Connecting r and python for linear mixed modeling. *Journal of Open Source Software*, 3(31):862, 2018.

[131] Philip Jonsson, Andrew L. Lin, Robert J. Young, Natalie M. DiStefano, David M. Hyman, Bob T. Li, Michael F. Berger, Ahmet Zehir, Marc Ladanyi, David B. Solit, Angela G. Arnold, Zsofia K. Stadler, Diana Mandelker, Michael E. Goldberg, Juliann Chmielecki, Maryam Pourmaleki, Shahiba Q. Ogilvie, Shweta S. Chavan, Andrew T. McKeown, Malbora Manne, Allison Hyde, Kathryn Beal, T. Jonathan Yang, Craig P. Nolan, Elena Pentsova, Antonio Omuro, Igor T. Gavrilovic, Thomas J. Kaley, Eli L. Diamond, Jacqueline B. Stone, Christian Grommes, Adrienne Boire, Mariza Daras, Anna F. Piotrowski, Alexandra M. Miller, Philip H. Gutin, Timothy A. Chan, Viviane S. Tabar, Cameron W. Brennan, Marc Rosenblum, Lisa M. DeAngelis, Ingo K. Mellinghoff, and Barry S. Taylor. Genomic correlates of disease progression and treatment response in prospectively characterized gliomas. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 25(18):5537–5547, 2019.

[132] A. Joshi, D. Scheinost, H. Okuda, D. Belhachemi, I. Murphy, L. H. Staib, and X. Papademetris. Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics*, 9(1):69–84, Mar 2011.

[133] Jorge Jovicich, Silvester Czanner, Douglas Greve, Elizabeth Haley, Andre van der Kouwe, Randy Gollub, David Kennedy, Franz Schmitt, Gregory Brown, James MacFall, Bruce Fischl, and Anders Dale. Reliability in multi-site structural mri studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30(2):436 – 443, 2006.

[134] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[135] Malene Juul, Johanna Bertl, Qianyun Guo, Morten Muhlig Nielsen, Michał Świtnicki, Henrik Hornshøj, Tobias Madsen, Asger Hobolth, and Jakob Skou

Pedersen. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. *eLife*, 6, 2017.

[136] Malene Juul, Tobias Madsen, Qianyun Guo, Johanna Bertl, Asger Hobolth, Manolis Kellis, and Jakob Skou Pedersen. ncdDetect2: improved models of the site-specific mutation rate in cancer and driver detection with robust significance evaluation. *Bioinformatics*, 35(2):189–199, 2019.

[137] Erik Kaestner, Adam Milton Morgan, Joseph Snider, Meilin Zhan, Xi Jiang, Roger Levy, Victor S. Ferreira, Thomas Thesen, and Eric Halgren. Toward a database of intracranial electrophysiology during natural language presentation. *Language, Cognition and Neuroscience*, 35(6):729–738, 2018.

[138] André Kahles, Kjong-Van Lehmann, Nora C. Toussaint, Matthias Hüser, Stefan G. Stark, Timo Sachsenberg, Oliver Stegle, Oliver Kohlbacher, Chris Sander, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J. M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J.

Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C. S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang,

Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McK-

ercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, and Gunnar Rätsch. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell*, 34(2):211–224.e6, 2018.

[139] Tze Zhen Evangeline Kang, Lina Zhu, Du Yang, Dongbo Ding, Xiaoxuan Zhu, Yi Ching Esther Wan, Jiaxian Liu, Saravanan Ramakrishnan, Landon Long

Chan, Siu Yuen Chan, Xin Wang, Haiyun Gan, Junhong Han, Toyotaka Ishibashi, Qing Li, and Kui Ming Chan. The elevated transcription of ADAM19 by the oncohistone h2be76k contributes to oncogenic properties in breast cancer. *The Journal of Biological Chemistry*, 296:100374, 2021.

[140] 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, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, Benjamin M. Neale, Mark J. Daly, and Daniel G. MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[141] Peter Kerpedjiev, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobelt, Jacob M. Luber, Scott B. Ouellette, Alaleh Azhir, Nikhil Kumar, Jeewon Hwang, Soohyun Lee, Burak H. Alver, Hanspeter Pfister, Leonid A. Mirny, Peter J. Park, and Nils Gehlenborg. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, 19(1):125, 2018.

[142] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, pages 1–32, 2022.

[143] Ekta Khurana, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A. Rubin, and Mark Gerstein. Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(2):93–108, 2016.

[144] Eugene V Koonin. The meaning of biological information. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2063):20150065, 2016.

[145] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[146] Brent M. Kuenzi, Jisoo Park, Samson H. Fong, Kyle S. Sanchez, John Lee, Jason F. Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38(5):672–684.e6, 2020.

[147] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–28, 2022.

[148] Thomas A. Kunkel and Dorothy A. Erie. Dna mismatch repair. *Annual Review of Biochemistry*, 74(1):681–710, 2005.

[149] G. R. Kuperberg, M. Broome, P. K. McGuire, A. S. David, M. Eddy, F. Ozawa, D. Goff, W. C. West, S.C.R. Williams, Andre van der Kouwe, David Salat, Anders Dale, and Bruce Fischl. Regionally localized thinning of the cerebral cortex in Schizophrenia. *Archives of General Psychiatry*, 60:878–888, 2003.

[150] Esther Landhuis. Neuroscience: Big brain, big data. *Nature*, 541(7638):559–561, 2017.

[151] Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, Adam Kiezun, Peter S. Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H. Ramos, Trevor J. Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L. Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I. Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M. Dulak, Jens Lohr, Dan-Avi Landau, Catherine J. Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A. McCarroll, Jaume Mora, Ryan S. Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Charles W. M. Roberts, Jaclyn A. Biegel, Kimberly Stegmaier, Adam J. Bass, Levi A. Garraway, Matthew Meyerson, Todd R. Golub, Dmitry A. Gordenin, Shamil Sunyaev, Eric S. Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[152] Andrew R. J. Lawson, Federico Abascal, Tim H. H. Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, Mathijs A. Sanders, Anne Y. Warren, Krishnaa T. A. Mahbubani, Bethany Bareham, Timothy M. Butler, Luke M. R. Harvey, Alex Cagan, Andrew Menzies, Luiza Moore, Alexandra J. Colquhoun, William Turner, Benjamin Thomas, Vincent Gnanapragasam, Nicholas Williams, Doris M. Rassl, Harald Vöhringer, Sonia Zumalave, Jyoti Nangalia, José M. C. Tubío, Moritz Gerstung, Kourosh Saeb-Parsy,

Michael R. Stratton, Peter J. Campbell, Thomas J. Mitchell, and Iñigo Martincorena. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science*, 2020.

[153] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.

[154] Guangye Li, Shize Jiang, Sivylla E. Paraskevopoulou, Meng Wang, Yang Xu, Zehan Wu, Liang Chen, Dingguo Zhang, and Gerwin Schalk. Optimal referencing for stereo-electroencephalographic (seeg) recordings. *NeuroImage*, 183:327–335, 2018.

[155] Yang I. Li, David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, 2018.

[156] Yang I. Li, David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, 2018.

[157] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, 2014.

[158] GB Lindsay, RM Merrill, and RJ Hedin. The contribution of public health and improved social conditions to increased life expectancy: An analysis of public awareness. *Journal of Community Medicine & Health Education*, 4(5):1–10, 2014.

[159] J. K. Lindsey. *Statistical Analysis of Stochastic Processes in Time*. Cambridge University Press, 2004.

[160] Jeremy W Linsley, Drew A Linsley, Josh Lamstein, Gennadi Ryan, Kevan Shah, Nicholas A Castello, Viral Oza, Jaslin Kalra, Shijie Wang, Zachary Tokuno, et al. Superhuman cell death detection with biomarker-optimized neural networks. *Science advances*, 7(50):eabf8142, 2021.

[161] Lucas Lochovsky, Jing Zhang, Yao Fu, Ekta Khurana, and Mark Gerstein. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Research*, 43(17):8123–8134, 2015.

[162] Cheng-Yu Ma, Yi-Ping Phoebe Chen, Bonnie Berger, and Chung-Shou Liao. Identification of protein complexes by integrating multiple alignment of protein interaction networks. *Bioinformatics*, 33(11):1681–1688, 2017.

[163] Jianzhu Ma, Samson H. Fong, Yunan Luo, Christopher J. Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk F. A. Wessels, Marc Hafner, Roded Sharan, Jian Peng, and Trey Ideker. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2(2):233–244, 2021.

[164] Salah Mahmoudi, Sofia Henriksson, Martin Corcoran, Cristina Méndez-Vidal, Klas G. Wiman, and Marianne Farnebo. Wrap53, a natural p53 antisense transcript required for p53 induction upon DNA damage. *Molecular Cell*, 33(4):462–471, 2009.

[165] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. 1993.

[166] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041, 2017.

[167] Iñigo Martincorena and Peter J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[168] Iñigo Martincorena, Joanna C. Fowler, Agnieszka Wabik, Andrew R. J. Lawson, Federico Abascal, Michael W. J. Hall, Alex Cagan, Kasumi Murai, Krishnaa Mahbubani, Michael R. Stratton, Rebecca C. Fitzgerald, Penny A. Handford, Peter J. Campbell, Kourosh Saeb-Parsy, and Philip H. Jones. Somatic mutant clones colonize the human esophagus with age. *Science*, 2018.

[169] Iñigo Martincorena, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041.e21, 2017.

[170] Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572, 2020.

[171] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, volume 8, 2015.

[172] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[173] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[174] Daniel W McShea and Robert N Brandon. *Biology's first law: the tendency for diversity and complexity to increase in evolutionary systems*. University of Chicago Press, 2010.

[175] Collin Melton, Jason A. Reuter, Damek V. Spacek, and Michael Snyder. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, 47(7):710–716, 2015.

[176] Nima Mesgarani, Connie Cheung, Keith Johnson, , and Edward F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

[177] Anna Mestres-Missé, Antoni Rodriguez-Fornells, and Thomas F Münte. Neural differences in the mapping of verb and noun concepts onto novel words. *NeuroImage*, 49(3):2826–2835, 2010.

[178] Anna Mestres-Missé, Antoni Rodriguez-Fornells, and Thomas F. Münte. Neural differences in the mapping of verb and noun concepts onto novel words. *NeuroImage*, 49(3):2826–2835, 2010.

[179] meuleman. meuleman/epilogos, 2021.

[180] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

[181] Charles Molnar and Jane Gair. *Concepts of Biology, 1st Canadian Edition*. BCCampus, 2015.

[182] Luiza Moore, Alex Cagan, Tim H. H. Coorens, Matthew D. C. Neville, Rashesh Sanghvi, Mathijs A. Sanders, Thomas R. W. Oliver, Daniel Leongamornlert, Peter Ellis, Ayesha Noorani, Thomas J. Mitchell, Timothy M. Butler, Yvette Hooks, Anne Y. Warren, Mette Jorgensen, Kevin J. Dawson, Andrew Menzies, Laura O'Neill, Calli Latimer, Mabel Teng, Ruben van Boxtel, Christine A. Iacobuzio-Donahue, Inigo Martincorena, Rakesh Heer, Peter J. Campbell, Rebecca C. Fitzgerald, Michael R. Stratton, and Raheleh Rahbari. The mutational landscape of human somatic and germline cells. *Nature*, 597(7876):381–386, 2021.

[183] Rachel L Moseley and Friedemann Pulvermüller. Nouns, verbs, objects, actions, and abstractions: Local fmri activity indexes semantics, not lexical categories. *Brain and language*, 132:28–42, 2014.

[184] Frederick Mosteller and R. A. Fisher. Questions and answers. *The American Statistician*, 2(5):30–31, 1948.

[185] Ferran Muiños, Francisco Martínez-Jiménez, Oriol Pich, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. In silico saturation mutagenesis of cancer genes. *Nature*, 596(7872):428–432, 2021.

[186] Roy Mukamel and Itzhak Fried. Human intracranial recordings and cognitive neuroscience. *Annual review of psychology*, 63:511–537, 2012.

[187] Loris Mularoni, Radhakrishnan Sabarinathan, Jordi Deu-Pons, Abel Gonzalez-Perez, and Núria López-Bigas. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biology*, 17(1):128, 2016.

[188] David J Musliner, James A Hendler, Ashok K Agrawala, Edmund H Durfee, Jay K Strosnider, and CJ Paul. The challenges of real-time ai. *Computer*, 28(1):58–66, 1995.

[189] Erica Baker Nancy Katz and John Macnamara. What's in a name? a study of how children learn common and proper names. *Child Development*, 45(2):469–473, 1974.

[190] Joseph Nasser, Drew T. Bergman, Charles P. Fulco, Philine Guckelberger, Benjamin R. Doughty, Tejal A. Patwardhan, Thouis R. Jones, Tung H. Nguyen, Jacob C. Ulirsch, Fritz Lekschas, Kristy Mualim, Heini M. Natri, Elle M. Weeks, Glen Munson, Michael Kane, Helen Y. Kang, Ang Cui, John P. Ray, Thomas M. Eisenhaure, Ryan L. Collins, Kushal Dey, Hanspeter Pfister, Alkes L. Price, Charles B. Epstein, Anshul Kundaje, Ramnik J. Xavier, Mark J. Daly, Hailiang Huang, Hilary K. Finucane, Nir Hacohen, Eric S. Lander, and Jesse M. Engreitz. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243, 2021.

[191] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B. Alexandrov, Sancha Martin, David C. Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B. Brinkman, Sandro Morganella, Miriam R. Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E. Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A. Foekens, Moritz Gerstung, Gerrit K. J. Hooijer, Se Jin Jang, David R. Jones, Hyung-Yong Kim, Tari A. King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A. Purdie, Keiran Raine, Kamna Ramakrishnan, F. Germán Rodríguez-González, Gilles Romieu, Anieta M. Sieuwerts, Peter T. Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A. Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G. Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos

Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T. Ueno, Christos Sotiriou, Alain Viari, P. Andrew Futreal, Peter J. Campbell, Paul N. Span, Steven Van Laere, Sunil R. Lakhani, Jorunn E. Eyfjord, Alastair M. Thompson, Ewan Birney, Hendrik G. Stunnenberg, Marc J. van de Vijver, John W. M. Martens, Anne-Lise Børresen-Dale, Andrea L. Richardson, Gu Kong, Gilles Thomas, and Michael R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

[192] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

[193] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May 2020. European Language Resources Association.

[194] Joakim Nivre, Jens Nilsson, and Johan Hall. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

[195] S. Oltean and D. O. Bates. Hallmarks of alternative splicing in cancer. *Oncogene*, 33(46):5311–5318, 2014.

[196] Sheli L. Ostrow, Ruth Barshir, James DeGregori, Esti Yeger-Lotem, and Ruth Hershberg. Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLOS Genetics*, 10(3):e1004239, 2014.

[197] Christopher J. Paciorek and Mark J. Schervish. Nonstationary covariance functions for gaussian process regression. In *Advances in Neural Information Processing Systems 16*, pages 273–280. MIT Press, 2004.

[198] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[199] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga,

Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *arXiv:1912.01703 [cs, stat]*, 2019.

[200] Paz Polak, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence, Alex Reynolds, Eric Rynes, Kristian Vlahoviček, John A. Stamatoyannopoulos, and Shamil R. Sunyaev. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364, 2015.

[201] Kivan Polimis, Ariel Rokem, and Bryna Hazelton. Confidence intervals for random forests in python. *Journal of Open Source Software*, 2(1), 2017.

[202] Robert Pollie. Machine learning produces superhuman chip designs, 2022.

[203] Gladys Y. P. Poon, Caroline J. Watson, Daniel S. Fisher, and Jamie R. Blundell. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nature Genetics*, 53(11):1597–1605, 2021.

[204] Peter Priestley, Jonathan Baber, Martijn P. Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, Susan Haidari, Arne van Hoeck, Wendy Onstenk, Paul Roepman, Mircea Voda, Haiko J. Bloemendal, Vivianne C. G. Tjan-Heijnen, Carla M. L. van Herpen, Mariette Labots, Petronella O. Witteveen, Egbert F. Smit, Stefan Sleijfer, Emile E. Voest, and Edwin Cuppen. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, 575(7781):210–216, 2019.

[205] Liina Pylkkänen. The neural basis of combinatory syntax and semantics. *Science*, 366(6461):62–66, 2019.

[206] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[207] Friedemann Pulvermüller Rachel L.Moseley. Nouns, verbs, objects, actions, and abstractions: Local fmri activity indexes semantics, not lexical categories. *Brain and Language*, 132:28–42, 2014.

[208] Partho Sarothi Ray, Richa Grover, and Saumitra Das. Two internal ribosome entry sites mediate the translation of p53 isoforms. *EMBO Reports*, 7(4):404–410, 2006.

[209] Pedram Razavi, Matthew T. Chang, Guotai Xu, Chaitanya Bandlamudi, Dara S. Ross, Neil Vasan, Yanyan Cai, Craig M. Bielski, Mark T. A. Donoghue, Philip Jonsson, Alexander Penson, Ronglai Shen, Fresia Pareja, Ritika Kundra, Sumit Middha, Michael L. Cheng, Ahmet Zehir, Cyriac Kandoth, Ruchi Patel,

Kety Huberman, Lillian M. Smyth, Komal Jhaveri, Shanu Modi, Tiffany A. Traina, Chau Dang, Wen Zhang, Britta Weigelt, Bob T. Li, Marc Ladanyi, David M. Hyman, Nikolaus Schultz, Mark E. Robson, Clifford Hudis, Edi Brogi, Agnes Viale, Larry Norton, Maura N. Dickler, Michael F. Berger, Christine A. Iacobuzio-Donahue, Sarat Chandarlapaty, Maurizio Scaltriti, Jorge S. Reis-Filho, David B. Solit, Barry S. Taylor, and José Baselga. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell*, 34(3):427–438.e6, 2018.

[210] Martin Reuter and Bruce Fischl. Avoiding asymmetry-induced bias in longitudinal image processing. *NeuroImage*, 57(1):19–21, 2011.

[211] Martin Reuter, Herminia Diana Rosas, and Bruce Fischl. Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, 53(4):1181–1196, 2010.

[212] Martin Reuter, Nicholas J. Schmansky, Herminia Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, 2012.

[213] Esther Rheinbay, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, Julian M Hess, Randi Istrup Juul, Ziao Lin, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, 2020.

[214] Esther Rheinbay, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A. Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, Julian M. Hess, Randi Istrup Juul, Ziao Lin, Lars Feuerbach, Radhakrishnan Sabarinathan, Tobias Madsen, Jaegil Kim, Loris Mularoni, Shimin Shuai, Andrés Lanzós, Carl Herrmann, Yosef E. Maruvka, Ciyue Shen, Samirkumar B. Amin, Pratiti Bandopadhayay, Johanna Bertl, Keith A. Boroevich, John Busanovich, Joana Carlevaro-Fita, Dimple Chakravarty, Calvin Wing Yiu Chan, David Craft, Priyanka Dhingra, Klev Diamanti, Nuno A. Fonseca, Abel Gonzalez-Perez, Qianyun Guo, Mark P. Hamilton, Nicholas J. Haradhvala, Chen Hong, Keren Isaev, Todd A. Johnson, Malene Juul, Andre Kahles, Abdullah Kahraman, Youngwook Kim, Jan Komorowski, Kiran Kumar, Sushant Kumar, Donghoon Lee, Kjong-Van Lehmann, Yilong Li, Eric Minwei Liu, Lucas Lochovsky, Keunchil Park, Oriol Pich, Nicola D. Roberts, Gordon Saksena, Steven E. Schumacher, Nikos Sidiropoulos, Lina Sieverling, Nasa Sinnott-Armstrong, Chip Stewart, David Tamborero, Jose M. C. Tubio, Husen M. Umer, Liis Uusküla-Reimand, Claes Wadelius, Lina Wadi, Xiaotong Yao, Cheng-Zhong Zhang, Jing Zhang, James E. Haber, Asger Hobolth, Marcin Imielinski, Manolis Kellis, Michael S. Lawrence, Christian von Mering, Hidewaki Nakagawa, Benjamin J. Raphael, Mark A. Rubin, Chris Sander, Lincoln D. Stein, Joshua M. Stuart, Tatsuhiko Tsunoda, David A. Wheeler, Rory Johnson, Jüri Reimand, Mark Gerstein, Ekta Khurana, Peter J. Campbell,

Núria López-Bigas, Joachim Weischenfeldt, Rameen Beroukhim, Iñigo Martincorena, Jakob Skou Pedersen, and Gad Getz. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, 2020.

[215] Esther Rheinbay, Prasanna Parasuraman, Jonna Grimsby, Grace Tiao, Jesse M. Engreitz, Jaegil Kim, Michael S. Lawrence, Amaro Taylor-Weiner, Sergio Rodriguez-Cuevas, Mara Rosenberg, Julian Hess, Chip Stewart, Yosef E. Maruvka, Petar Stojanov, Maria L. Cortes, Sara Seepo, Carrie Cibulskis, Adam Tracy, Trevor J. Pugh, Jesse Lee, Zongli Zheng, Leif W. Ellisen, A. John Iafrate, Jesse S. Boehm, Stacey B. Gabriel, Matthew Meyerson, Todd R. Golub, Jose Baselga, Alfredo Hidalgo-Miranda, Toshi Shioda, Andre Bernards, Eric S. Lander, and Gad Getz. Recurrent and functional regulatory mutations in breast cancer. *Nature*, 547(7661):55–60, 2017.

[216] Mark D. Risser. Review: Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. *arXiv:1610.02447 [stat]*, 2016.

[217] Hira Rizvi, Francisco Sanchez-Vega, Konnor La, Walid Chatila, Philip Jonsson, Darragh Halpenny, Andrew Plodkowski, Niamh Long, Jennifer L. Sauter, Natasha Rekhtman, Travis Hollmann, Kurt A. Schalper, Justin F. Gainor, Ronglai Shen, Ai Ni, Kathryn C. Arbour, Taha Merghoub, Jedd Wolchok, Alexandra Snyder, Jamie E. Chaft, Mark G. Kris, Charles M. Rudin, Nicholas D. Socci, Michael F. Berger, Barry S. Taylor, Ahmet Zehir, David B. Solit, Maria E. Arcila, Marc Ladanyi, Gregory J. Riely, Nikolaus Schultz, and Matthew D. Hellmann. Molecular determinants of response to anti–programmed cell death (PD)-1 and anti–programmed death-ligand 1 (PD-l1) blockade in patients with non–small-cell lung cancer profiled with targeted next-generation sequencing. *Journal of Clinical Oncology*, 36(7):633–641, 2018.

[218] 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. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, 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. Siebenthall, 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, 2015.

[219] H. D. Rosas, A. K. Liu, S. Hersch, M. Glessner, R. J. Ferrante, D. H. Salat, A. van der Kouwe, B. G. Jenkins, A. M. Dale, and B. Fischl. Regional and progressive thinning of the cortical ribbon in Huntington's disease. *Neurology*, 58(5):695–701, 2002.

[220] Ned T Sahin, Steven Pinker, Sydney S Cash, Donald Schomer, and Eric Halgren. Sequential processing of lexical, grammatical, and phonological information within broca's area. *Science*, 326(5951):445–449, 2009.

[221] Ned T Sahin, Steven Pinker, and Eric Halgren. Abstract grammatical processing of nouns and verbs in broca's area: evidence from fmri. *Cortex*, 42(4):540–562, 2006.

[222] David Salat, R.L. Buckner, A.Z. Snyder, Douglas N. Greve, R.S. Desikan, Evelina Busa, J.C. Morris, Anders Dale, and Bruce Fischl. Thinning of the cerebral cortex in aging. *Cerebral Cortex*, 14:721–730, 2004.

[223] Nicolae Sapoval, Amirali Aghazadeh, Michael G Nute, Dinler A Antunes, Advait Balaji, Richard Baraniuk, CJ Barberan, Ruth Dannenfelser, Chen Dun, Mohammadamin Edrisi, et al. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1–12, 2022.

[224] Aaron Schein, Hanna Wallach, and Mingyuan Zhou. Poisson-gamma dynamical systems. In *Advances in Neural Information Processing Systems 29*, pages 5005–5013. Curran Associates, Inc., 2016.

[225] Marianne Schell, Emiliano Zaccarella, and Angela D Friederici. Differential cortical contribution of syntax and semantics: An fmri study on two-word phrasal processing. *Cortex*, 96:105–120, 2017.

[226] Benjamin Schuster-Böckler and Ben Lehner. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–507, 2012.

[227] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[228] F. Segonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in mri. *NeuroImage*, 22(3):1060 – 1075, 2004.

[229] F. Segonne, J. Pacheco, and B. Fischl. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans Med Imaging*, 26:518–529, 2007.

[230] Terrence J Sejnowski. *The deep learning revolution*. MIT press, 2018.

[231] Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307, 2020.

[232] Kevin A. Shapiro, Lauren R. Moo, and Alfonso Caramazza. Cortical signatures of noun and verb production. *Proceedings of the National Academy of Sciences*, 103(5):1644–1649, 2006.

[233] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[234] Maxwell A. Sherman, Adam U. Yaari, Oliver Priebe, Felix Dietlein, Po-Ru Loh, and Bonnie Berger. Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nature Biotechnology*, 2022.

[235] Lynette M. Sholl, Khanh Do, Priyanka Shivdasani, Ethan Cerami, Adrian M. Dubuc, Frank C. Kuo, Elizabeth P. Garcia, Yonghui Jia, Phani Davineni, Ryan P. Abo, Trevor J. Pugh, Paul van Hummelen, Aaron R. Thorner, Matthew Ducar, Alice H. Berger, Mizuki Nishino, Katherine A. Janeway, Alanna Church, Marian Harris, Lauren L. Ritterhouse, Joshua D. Campbell, Vanesa Rojas-Rudilla, Azra H. Ligon, Shakti Ramkissoon, James M. Cleary, Ursula Matulonis, Geoffrey R. Oxnard, Richard Chao, Vanessa Tassell, James Christensen, William C. Hahn, Philip W. Kantoff, David J. Kwiatkowski, Bruce E. Johnson, Matthew Meyerson, Levi A. Garraway, Geoffrey I. Shapiro, Barrett J. Rollins, Neal I. Lindeman, and Laura E. MacConaill. Institutional implementation of clinical tumor profiling on an unselected cancer population. *JCI Insight*, 1(19), 2016.

[236] Shimin Shuai, PCAWG Drivers and Functional Interpretation Working Group, Steven Gallinger, Lincoln Stein, and PCAWG Consortium. Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nature Communications*, 11(1):734, 2020.

[237] Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[238] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[239] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034 [cs]*, 2014.

[240] Matthias J. Sjerps, Neal P. Fox, Keith Johnson, and Edward F. Chang. Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, 10(2465), 2019.

[241] J.G. Sled, A.P. Zijdenbos, and A.C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans Med Imaging*, 17:87–97, 1998.

[242] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

[243] Milan Straka. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2018.

[244] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[245] Fran Supek and Ben Lehner. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA repair*, 81:102647, 2019.

[246] Fran Supek, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–1335, 2014.

[247] David Tamborero, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P. Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, Carmen de Torres, Rodrigo Dienstmann, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(1):25, 2018.

[248] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2018.

[249] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E

Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47:D941–D947, 2019.

[250] Michalis K Titsias. Variational learning of inducing variables in sparse gaussian processes. *AISTATS*, page 8, 2009.

[251] Lorraine K. Tyler, Richard Russell, Jalal Fadili, and Helen E. Moss. The neural representation of nouns and verbs: PET studies. *Brain*, 124(8):1619–1634, 2001.

[252] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[253] Paul A. VanderLaan, Deepa Rangachari, and Daniel B. Costa. The rapidly evolving landscape of biomarker testing in non-small cell lung cancer. *Cancer cytopathology*, 129(3):179–181, 2021.

[254] Gabriella Vigliocco, David Vinson, Judit Druks, Horacio Barber, and Stefano Cappa. Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and biobehavioral reviews*, 35:407–26, 05 2010.

[255] Jan Vijg. Somatic mutations, genome mosaicism, cancer and aging. *Current opinion in genetics & development*, 26:141–149, 2014.

[256] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[257] Lina Wadi, Liis Uusküla-Reimand, Keren Isaev, Shimin Shuai, Vincent Huang, Minggao Liang, J. Drew Thompson, Yao Li, Luyao Ruan, Marta Paczkowska, Michal Krassowski, Irakli Dzneladze, Ken Kron, Alexander Murison, Parisa Mazrooei, Robert G. Bristow, Jared T. Simpson, Mathieu Lupien, Michael D. Wilson, Lincoln D. Stein, Paul C. Boutros, and Jüri Reimand. Candidate cancer driver mutations in superenhancers and long-range chromatin interaction networks. *bioRxiv*, page 236802, 2017.

[258] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, page 27, 2014.

[259] Jiuling Wang, Weimin Feng, Zhen Yuan, Jason D. Weber, and Yandong Zhang. Dhx33 interacts with ap-2β to regulate bcl-2 gene expression and promote cancer cell survival. *Molecular and Cellular Biology*, 39(17), 2019.

[260] Kirsten Weber, Morten H Christiansen, Karl Magnus Petersson, Peter Indefrey, and Peter Hagoort. fMRI syntactic and lexical repetition effects reveal the initial stages of learning a new language. *Journal of Neuroscience*, 36(26):6872–6880, 2016.

[261] Donate Weghorn and Shamil Sunyaev. Bayesian inference of negative and positive selection in human cancers. *Nature genetics*, 49(12):1785–1788, 2017.

[262] Donate Weghorn and Shamil Sunyaev. Bayesian inference of negative and positive selection in human cancers. *Nature Genetics*, 49(12):1785–1788, 2017.

[263] P. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

[264] Masha Westerlund, Itamar Kastner, Meera Al Kaabi, and Liina Pylkkänen. The latl as locus of composition: Meg evidence from english and arabic. *Brain and Language*, 141:124–134, 2015.

[265] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[266] Yuri I Wolf, Mikhail I Katsnelson, and Eugene V Koonin. Physical foundations of biological complexity. *Proceedings of the National Academy of Sciences*, 115(37):E8678–E8687, 2018.

[267] Caroline F. Wright, Nicholas M. Quaife, Laura Ramos-Hernández, Petr Danecek, Matteo P. Ferla, Kaitlin E. Samocha, Joanna Kaplanis, Eugene J. Gardner, Ruth Y. Eberhardt, Katherine R. Chao, Konrad J. Karczewski, Joannella Morales, Giuseppe Gallone, Meena Balasubramanian, Siddharth Banka, Lianne Gompertz, Bronwyn Kerr, Amelia Kirby, Sally A. Lynch, Jenny E. V. Morton, Hailey Pinz, Francis H. Sansbury, Helen Stewart, Britton D. Zuccarelli, Stuart A. Cook, Jenny C. Taylor, Jane Juusola, Kyle Retterer, Helen V. Firth,

Matthew E. Hurles, Enrique Lara-Pezzi, Paul J. R. Barton, and Nicola Whiffin. Non-coding region variants upstream of MEF2c cause severe developmental disorder through three distinct loss-of-function mechanisms. *The American Journal of Human Genetics*, 108(6):1083–1094, 2021.

[268] Jie Wu, Olga Anczuków, Adrian R. Krainer, Michael Q. Zhang, and Chaolin Zhang. OLego: fast and sensitive mapping of spliced mRNA-seq reads using small seeds. *Nucleic Acids Research*, 41(10):5149–5163, 2013.

[269] Chunming Xu and Scott A Jackson. Machine learning and complex biological data, 2019.

[270] Adam Uri Yaari, Maxwell Sherman, Oliver Clarke Priebe, Po-Ru Loh, Boris Katz, Andrei Barbu, and Bonnie Berger. Multi-resolution modeling of a discrete stochastic process identifies causes of cancer. *International Conference on Learning Representations*, 2021.

[271] A. I. Yang, X. Wang, W. K. Doyle, E. Halgren, C. Carlson, T. L. Belcher, S. S. Cash, O. Devinsky, and T. Thesen. Localization of dense intracranial electrode arrays using magnetic resonance imaging. *Neuroimage*, 63(1):157–165, Oct 2012.

[272] D.-Q. Yang, M.-J. Halaby, and Y. Zhang. The identification of an internal ribosomal entry site in the 5-untranslated region of p53 mRNA provides a novel mechanism for the regulation of its translation following DNA damage. *Oncogene*, 25(33):4613–4619, 2006.

[273] Ahmet Zehir, Ryma Benayed, Ronak H. Shah, Aijazuddin Syed, Sumit Middha, Hyunjae R. Kim, Preethi Srinivasan, Jianjiong Gao, Debyani Chakravarty, Sean M. Devlin, Matthew D. Hellmann, David A. Barron, Alison M. Schram, Meera Hameed, Snjezana Dogan, Dara S. Ross, Jaclyn F. Hechtman, Deborah F. DeLair, JinJuan Yao, Diana L. Mandelker, Donavan T. Cheng, Raghu Chandramohan, Abhinita S. Mohanty, Ryan N. Ptashkin, Gowtham Jayakumaran, Meera Prasad, Mustafa H. Syed, Anoop Balakrishnan Rema, Zhen Y. Liu, Khedoudja Nafa, Laetitia Borsu, Justyna Sadowska, Jacklyn Casanova, Ruben Bacares, Iwona J. Kiecka, Anna Razumova, Julie B. Son, Lisa Stewart, Tessara Baldi, Kerry A. Mullaney, Hikmat Al-Ahmadie, Efsevia Vakiani, Adam A. Abeshouse, Alexander V. Penson, Philip Jonsson, Niedzica Camacho, Matthew T. Chang, Helen H. Won, Benjamin E. Gross, Ritika Kundra, Zachary J. Heins, Hsiao-Wei Chen, Sarah Phillips, Hongxin Zhang, Jiaojiao Wang, Angelica Ochoa, Jonathan Wills, Michael Eubank, Stacy B. Thomas, Stuart M. Gardos, Dalicia N. Reales, Jesse Galle, Robert Durany, Roy Cambria, Wassim Abida, Andrea Cercek, Darren R. Feldman, Mrinal M. Gounder, A. Ari Hakimi, James J. Harding, Gopa Iyer, Yelena Y. Janjigian, Emmet J. Jordan, Ciara M. Kelly, Maeve A. Lowery, Luc G.T. Morris, Antonio M. Omuro, Nitya Raj, Pedram Razavi, Alexander N. Shoushtari, Neerav Shukla, Tara E. Soumerai, Anna M. Varghese, Rona Yaeger, Jonathan Coleman, Bernard

Bochner, Gregory J. Riely, Leonard B. Saltz, Howard I. Scher, Paul J. Sabbatini, Mark E. Robson, David S. Klimstra, Barry S. Taylor, Jose Baselga, Nikolaus Schultz, David M. Hyman, Maria E. Arcila, David B. Solit, Marc Ladanyi, and Michael F. Berger. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*, 23(6):703–713, 2017.

[274] Chaoxiong Zhang, Ning Wu, Fu Gao, Jiaqi Han, Yanyong Yang, Chuanfeng Zhou, Weimin Sun, Linfeng Xian, Ying Cheng, Bailong Li, Jianming Cai, and Cong Liu. Mterfd1 functions as an oncogene. *Oncotarget*, 5(0), 2014.

[275] Qian Zhang, Lei Yu, Dandan Qin, Rui Huang, Xiaochen Jiang, Chendan Zou, Qingchao Tang, Yinggang Chen, Guiyu Wang, Xishan Wang, and Xu Gao. Role of microRNA-30c targeting ADAM19 in colorectal cancer. *PloS One*, 10(3):e0120698, 2015.

[276] Xiaoyang Zhang and Matthew Meyerson. Illuminating the noncoding genome in cancer. *Nature Cancer*, 1(9):864–872, 2020.

[277] Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, 11(1):1–13, 2020.

[278] Siming Zhao, Jun Liu, Pranav Nanga, Yuwen Liu, A. Ercument Cicek, Nicholas Knoblauch, Chuan He, Matthew Stephens, and Xin He. Detailed modeling of positive selection improves detection of cancer driver genes. *Nature Communications*, 10(1):3399, 2019.

[279] Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, and Bonnie Berger. Reconstructing continuous distributions of 3d protein structure from cryo-EM images. *arXiv:1909.05215 [cs, eess, q-bio, stat]*, 2020.

[280] Helen Zhu, Liis Uusküla-Reimand, Keren Isaev, Lina Wadi, Azad Alizada, Shimin Shuai, Vincent Huang, Dike Aduluso-Nwaobasi, Marta Paczkowska, Diala Abd-Rabbo, Oliver Ocsenas, Minggao Liang, J. Drew Thompson, Yao Li, Luyao Ruan, Michal Krassowski, Irakli Dzneladze, Jared T. Simpson, Mathieu Lupien, Lincoln D. Stein, Paul C. Boutros, Michael D. Wilson, and Jüri Reimand. Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks. *Molecular Cell*, 77(6):1307–1321.e10, 2020.

[281] Tonći Šuštić, Sake van Wageningen, Evert Bosdriesz, Robert J. D. Reid, John Dittmar, Cor Lieftink, Roderick L. Beijersbergen, Lodewyk F. A. Wessels, Rodney Rothstein, and René Bernards. A role for the unfolded protein response stress sensor ERN1 in regulating the response to MEK inhibitors in KRAS mutant colon cancers. *Genome Medicine*, 10(1):90, 2018.