A computational search for mutational drivers of cancer

by

Collin Tokheim

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2018

Abstract

The notion that DNA changes could drive the growth of cancer was first speculated more than a century ago, and has acquired overwhelming evidence in the past several decades. The recent decrease in cost of next-generation sequencing has spurred the growth of cancer sequencing studies that catalog mutations observed in cancer. However, the vast majority of mutations in cancer do not increase the fitness of cancer cells. As a consequence, computational methods have become essential to distinguish the specific driver mutations implicated in cancer by leveraging statistical patterns of genetic variation observed across many cancer samples.

Here, I introduce several new computational methods to analyze cancer drivers at different levels of resolution – including at the gene (20/20+), protein region (HotMAPS), and mutation (CHASMplus) level. I use these methods to interrogate fundamental questions regarding cancer driver mutations, such as their cancer type specificity, commonness or rarity, and the characteristics of oncogenes and tumor suppressor genes. Different types of cancer varied

substantially on the precise cancer driver genes and the balance of oncogenes versus tumor suppressor genes, but shared clusters of cancer driver genes were seen in cancer types with a common cell of origin. Results also indicate a prominent emerging role for rare driver mutations, suggesting interpretation of a cancer genome will need to be increasingly personalized, as a patient's driver mutation may have not been previously observed.

I also probe the efficacy of computational methods, which is difficult because there is no accepted gold-standard. I first analyze consequences expected analytically, and then compare existing methods on newly developed benchmarks. I found many prior computational methods do not appropriately model the heterogeneity of mutations expected by chance.

The recent completion of The Cancer Genome Atlas has provided a unique capability to understand cancer at an unprecedented scale. I comprehensively discover both cancer driver genes and mutations across nearly 10,000 cancers from 33 cancer types. This revealed 299 cancer driver genes and >3,000 driver mutations. Although this expansive analysis found 59 novel genes not previously associated as cancer drivers, some evidence points to diminishing returns for future driver discovery.

Primary Reader and Advisor: Dr. Rachel Karchin

Secondary Reader: Dr. Joel Bader

Acknowledgments

This work could not have been accomplished isolation. I am greatly thankful for the countless hours of mentorship, encouragement, and inspiration provided by Rachel Karchin. I also appreciate the valuable advice provided by Noushin Niknafs, Chris Douville, and Violeta Beleva-Guthrie, especially when I just started in the lab. My lively experiences in the lab also would not have been the same without Rohit Bhattacharya, Ashok Sivakumar, Lily Zheng, and Melody Shao.

I would also like to thank collaborators who were essential in my projects. Bert Vogelstein's insights into cancer drivers were critical for formulating 20/20+. The CHASMplus analysis would not be the same without Nick Roberts' and Neha Nanda's experimental perspective on ATM. I especially thank Matthew Bailey and Eduard Porta-Pardo for their creativity, insights, enthusiasm, and camaraderie when analyzing cancer drivers as part of the driver's group of The Cancer Genome Atlas PancanAtlas.

Dedication

To *my parents*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cancer is a disease defined by aberrant proliferation of cells that have acquired invasiveness into surrounding tissues of the human body [1]. As a whole, cancer is estimated to have caused 600,000 deaths in the United States in 2017 [2]. The biological process of cancer development has been associated with numerous hallmarks known to circumvent the otherwise restricted growth of a normal cell [3]. The ongoing effort to reduce cancer mortality, whether by prevention or new treatments, may require a deeper understanding of the processes that lead to the development and progression of cancer. Given the limited throughput to study human cancers experimentally [4], my dissertation is focused on developing new computational methods to identify mutational drivers of cancer from the big data arising through large-scale DNA sequencing. Specifically, I will analyze protein-coding mutations that happen

somatically, i.e., starting from embryogenesis, mutations that occur in the cells of the body (excluding germ cells), and therefore are not inherited.

1.1 Cancer as a genetic disease

Cancer's foundation as a genetic disease was first proposed more than a century ago by observations of cells with chromosomal aberrations [5]. There was only sparse support for this hypothesis until the observation that chicken cells contained a homologous sequence to a gene in a known cancer-related virus, avian sarcoma virus, [6] and, further, that a single nucleotide change at codon 12 of the human gene *HRAS* could oncogenically transform bladder cells [1, 7]. Endogenous human genes, when mutated, could therefore contribute to the growth of cancer. As a technical note, I will refer to such genes that contain mutations which increase the net growth of cells toward cancer as "cancer driver genes". However, it was not clear at the time whether all such cancer driver genes would fit the mold of *HRAS*. Now it is understood that cancer driver genes fall into two broad categories, oncogenes and tumor suppressor genes. Oncogenes, like *HRAS*, acquire mutations that generate gain-of-function, while tumor suppressor genes acquire mutations that cause loss of function. Originally the view of tumor suppressor genes was as biallelic loss-of-function of both gene copies (the "two-hit hypothesis" [8]), such as by the combined effect

of a deletion (or loss-of-heterozygosity) and a mutation, like *RB1* in retinoblastoma [9] and TP53 in colorectal cancer [10]. However, for some tumor suppressor genes that are either haploinsufficient or dominant-negative, the mutation of only one copy may be sufficient [11–13].

A particular driver mutation may be neither necessary nor sufficient for the development of cancer. Rather, carcinogenesis, the development of cancer, often is a multi-step process (estimated 2-8 [14]) involving several driver mutations, where the combined effect of multiple mutations is sufficient. In the case of colorectal cancer, it is estimated 3 mutational drivers are required [15]. The driver mutation at each step causes a clonal expansion of cells because of their selective growth advantage; thus, leading to progression from a small adenoma to a large adenoma and eventually to a carcinoma in colorectal cancers [16]. As an example, a particular cancer's sequence of driver mutations could initiate with an *APC* gene mutation followed by a *KRAS* mutation and subsequent *TP53* mutations. But particular driver mutations are not necessarily exclusive to each stage that leads to colorectal cancer [16]. Moreover, in another patient's cancer, driver mutations in different genes could also lead to colorectal cancer [17]. Lastly, even within a single tumor, there may be multiple competing subclones with different compositions of driver mutations (termed "intra-tumor heterogeneity"). Carcinogenesis, therefore, is not a simple fixed linear path of driver mutations, but instead a remarkably heterogeneous mix of multiple

possible paths.

Only in the past decade have improvements in DNA sequencing technology made cancer sequencing studies feasible for cataloging large numbers of mutations in human cancers. The first wave of cancer sequencing studies [18–20] analyzed common cancers, such as breast and colorectal cancers, and sequenced only targeted portions of the exome (the regions encoding genes). Due to technical limitations and prohibitive cost, they employed a Discovery-Validation study design where mutations were first detected more comprehensively in a smaller number of samples but then validated against a larger set of samples. Although soon after, milestone studies would sequence the whole-exome of pancreatic cancers [21] and glioblastoma multiforme [22]. Also, in the same year, the first pilot project of the The Cancer Genome Atlas (TCGA) analyzed glioblastoma multiforme [23], the beginning of a consortium that would analyze thousands of human cancers in the coming years. The genomic breadth of sequencing was expanded by several whole-genome sequencing studies analyzing a few samples in leukemia, lung cancer, and melanoma [24–26]. By 2011 the TCGA analyzed 316 ovarian carcinoma samples by whole-exome sequencing [27], which started to reach the large sample sizes necessary for statistically implicating cancer drivers.

1.2 Positive selection and the statistical identification of cancer drivers

Cancer sequencing studies quickly made it evident that driver mutations only constitute a small percentage of the potential 100's or 1,000's of mutations observed with a single exome [14]. The major question is therefore not what mutations are detected in cancer but, rather, which mutations are drivers as opposed to "passengers" that do not contribute to tumorigenesis? Because the vast majority of mutations are passengers, it is difficult to distinguish a driver mutation from many passenger mutations; also, like a driver mutation, many passenger mutations may be completely clonal in a particular cancer sample, because they happened before the founding clone's driver mutation [28, 29]. Passenger mutations effectively hitchhike off of the selective growth advantage provided by driver mutations. The key distinction is, when considered across many cancer samples, that driver mutations are positively selected for in cancer and therefore should be disproportionately represented. Computational methods have therefore evaluated signals of positive selection in cancer at various scales, including at the protein, domain/region, and mutation level [30]. Although the precise way positive selection is statistically measured varies [31], the essence is to analyze patterns of mutations across many samples and rule out the possibility that the mutations are explainable by a ran-

dom background accumulation of mutations alone.

As a consequence, identifying positive selection of driver mutations also requires understanding the converse, how passenger mutations accumulate in cancer. Passenger mutations have only <1% of non-silent mutations eliminated by negative selection, indicating that most of the variability in the number of passenger mutations is due to mutation rates rather than selection [32]. A simple model is that passenger mutations accumulate at a universal background rate per base across the genome [20], after adjustments for the nucleotide sequence context. This model fails to capture key mutational processes in cancer, such as the background mutation rate varying by over two orders of magnitudes between cancer types [33] and also varying patient-by-patient, especially when a tumor has defective DNA damage repair or mutagen exposure [34]. Moreover, regional variation within the genome of replication timing, gene expression, and chromatin structure leads to ~3-fold differences in the background mutation rate [33]. In certain cancers, kataegis causes genomically localized hypermutation [35]. Accurate models therefore need to account for the greater dispersion caused by this mutational heterogeneity.

A focus for many large cancer sequencing studies has been to identify cancer driver genes, usually done in one of three ways. The most common approach, which I term as a significantly mutated gene method, has been to compare the number of mutations within a gene to that expected by a background

mutation rate [18, 20, 36, 37]. The accuracy of the background mutation rate, therefore, becomes the critical parameter to estimate. Recent improvements in estimating regional variation in mutation rate across the genome lead to reduced false positives in the method MutSigCV [38, 39]. The second approach, functional impact bias, evaluates whether a gene contains mutations that are skewed towards higher predicted impact [40]. The score of a mutation usually reflects either evolutionary conservation of the protein sequence or machine learning methods that predict the deleteriousness of a variant. Lastly, evaluating the positional clustering of mutations has also been used at the sequence- and structure-level of a protein [38, 41–55]. However, not all tumor suppressor genes may exhibit mutational clustering, and therefore these methods may be better at identifying oncogenes due to their highly localized activating mutations.

Although a cancer driver gene, by definition, contains a driver mutation, not all mutations within a cancer driver gene are necessarily cancer drivers. Especially considering the large size of genes, passenger mutations will be observed in large numbers when analyzing many cancer samples [32]. To address this issue, recent methods, known as hotspot mutation detection, have therefore focused on smaller regions, such as protein domains [55], protein-protein interfaces [40], and individual codons [9].

However, if driver mutations are in separate locations of the protein se-

quence, hotspot detection based on protein sequence may lose statistical power. Often this results from residues being far apart in protein sequence but actually proximal in the folded protein structure. Since there is a relationship between protein structure and function [56, 57], I and others have developed computational methods for hotspot detection in 3D space of protein structure [41, 43, 45, 46, 48, 52–54, 58]. Hotspot detection in protein structure has the advantage of generating plausible hypotheses about the function of the mutation given the spatial proximity to known functional sites in the protein [43, 59].

Cancer driver prediction at the level of individual mutations has largely focused on missense mutations, the most common type of protein-coding mutation in cancer [14]. Typically, machine learning approaches have been used to leverage features characterizing mutations. Although features vary substantially by method [39, 60–65], they usually include inter-species evolutionary conservation of the protein sequence, features of the local protein environment, molecular function annotations, and biophysical characterizations of the amino acid substitution. Cancer-focused machine learning methods have previously tried to enhance performance by training cancer type-specific models [26, 66] or boosting data with synthetic passenger missense mutations [26]. Despite the capability of utilizing many features, with the exception of a few gene-level features in ParsSNP [25], machine learning methods typically have not used mutational patterns characterizing the genetic variation observed in

human cancers. Furthermore, a systematic comparative study of 15 methods concluded that none of them were sufficiently reliable for experimental or clinical follow-through [49]. I, and others, have hypothesized that determining the impact of missense mutations requires proper context [59, 67], which have not been sufficiently leveraged in a comprehensive manner from the current generation of methods. Context includes both prior knowledge about the functional importance of genes or gene subregions in which a mutation occurs, and mutational patterns that are now evident from cancer sequencing studies of many thousands of patients.

1.3 Large-scale cancer driver discovery

The application of computational methods to identify mutational drivers of cancer has expanded with the growth in sample size of cancer sequencing studies [33,36,51,55,58,66,68–73]. A comprehensive analysis of 3,281 cancers comprising 12 cancer types from the TCGA revealed 125 associated cancer driver genes [66]. A subsequent study identified 224 cancer driver genes in 21 cancer types, and further suggested by sub-sampling analysis that the discovery of new cancer driver genes does not show evidence of saturation at current sample sizes [69]. Combined with the low mutation frequency of many identified cancer driver genes, it has been hypothesized there is a "long tail" of drivers of

increasing rarity, with more cancer drivers on the horizon [74,75]. Prior statistical power calculations suggest this is reasonable given driver genes estimated at a 2% frequency of cancers may require sample size as high as 5,000 for certain cancer types with high mutation rate and, in total across all cancer types, 100,000 sequenced cancer samples may be needed [69]. This is well above the number of samples per cancer type available in The Cancer Genome Atlas.

Only a relatively few studies have started to focus on identifying cancer drivers at sub-gene resolution. Initially, studies focused on identifying protein domains [49,50]. More recently, studies have progressed to varying sized hotspots and towards codon-level resolution [38,42,76]. However, I and others have noted such approaches currently are biased towards finding hotspot regions in oncogenes as opposed to tumor suppressor genes. This is a result of tumor suppressor genes having loss-of-function driver mutations that are more diverse and spread over a larger region of the protein [43].

1.4 Relevance to current study

Ultimately, computational methods have progressively sought to understand whether a particular mutation in a patient's cancer is a cancer driver mutation. However, given current approaches may require 100,000 cancer samples to just identify cancer driver genes, identifying particular driver mutations within

those genes may require an even greater number of sequenced cancers. Due to the millions of mutations currently being identified in human cancers, laborious experimental validation of all mutations is not feasible because of lack of throughput [4,59]. A more statistically powerful computational method would be greatly beneficial to the cancer genomics community; enabling improved utilization of cancer sequencing studies. Insight into cancer driver mutations can be of substantial clinical relevance, such as indicators of prognosis [22,77], therapeutic response [78], drug targets [79], and as biomarkers for early detection of cancer [80].

My dissertation covers four primary aims: **(1)** the modeling of somatic mutations, **(2)** development of new computational methods, **(3)** benchmarking of computational predictions, and **(4)** systematic driver discovery across thousands of human cancer samples. Where pertinent, results of methodological benchmarks (3) and systematic driver discovery (4) are combined in the same chapter as the corresponding developed computational method (2).

Although there are many existing computational methods, I find that prior methods have not yet adequately combined multiple signals of positive selection of driver mutations in cancer. I hypothesize better characterization of cancer drivers, particularly those that occur at relatively low prevalence, can be obtained by a carefully designed machine learning approach, which leverages mutational patterns in cancer sequencing studies that are characteristic

of oncogenes and tumor suppressor genes. Moreover, I show that combining multiple scales of information - including at the gene, region, and mutation level - has substantial benefits.

It has been difficult to evaluate progress in this area because many published methods do not rigorously compare their relative merits to those developed by others. I establish extensive benchmarks for cancer driver prediction to address this shortcoming. Importantly, I develop novel metrics to assess the current landscape of predictions, which addresses the need for rigorous evaluation criteria given the lack of a true gold standard for predicting cancer drivers. My analysis points to the strengths and weaknesses of each of the currently available methods and offers guidance for improving them in the future.

The recent completion of The Cancer Genome Atlas has provided a unique capability to understand cancer at an unprecedented scale. Here, I predicted protein-coding driver mutations in nearly 10,000 cancers and characterize the landscape of drivers across human cancers. This involves interrogating fundamental questions regarding cancer and driver mutations, such as their cancer type specificity, commonness or rarity, the balance and characteristics of oncogenes and tumor suppressor genes, and the likely future trajectory of cancer driver discovery.

Chapter 2

Statistically modeling the accumulation of somatic mutations in cancer

Somatic mutations accumulate randomly in all cells of the body, starting from the beginning of embryogenesis through the entire lifetime of an individual [81]. Somatic mutations arise because of both endogenous and exogenous sources, or from a mutator phenotype acquired in a cancer cell. Endogenous sources include intrinsic DNA replication mistakes, damage caused by free radicals from metabolism, and spontaneous deamination of nucleotides [82]. In contrast, exogenous sources originate from the environment and include ultraviolet radiation [82] and various mutagenic chemicals like aristolochic

acid [83]. Also, defective DNA damage repair in a cancer cell may lead to a mutator phenotype where a substantial number of unrepaired mutations become fixed, such as defective mismatch repair genes leading to microsatellite instability [84]. Each mutational source leaves a mark on the cancer genome that reflects the mutational signatures during the lifetime of the cell's progeny leading to cancer [82].

An essential first step towards implicating cancer driver genes from a cancer sequencing study requires understanding how mutations accumulate in cancer.

2.1 Variability in the background accumulation of mutations

To investigate the potential for using elevated mutation rate per base as a means to detect cancer drivers, I sought to examine the background variability of mutation rate in human cancers. The median background mutation rate per base for each cancer type in my pan-cancer data set (see Appendix B) [85] varied by over two orders of magnitude (Figure 2.1), with individual samples varying over an even larger range, which is consistent with prior observations [33, 34]. Because only a small fraction of the total somatic mutations in any common solid tumor affects driver genes, the remaining mu-

tations can be considered passengers. The total number of mutations (drivers plus passengers) per base is therefore only slightly larger than the number of passenger mutations per base, and, for simplicity, I refer to this number as the background mutation rate. Mutation rates are also known to vary across the genome [33] and are influenced by nucleotide sequence context, gene expression, chromatin state, transcription factor occupancy, replication timing, DNA strand, and perhaps by a variety of factors that have yet to be discovered [34, 86–88]. For example, melanoma mutations in The Cancer Genome Atlas (TCGA) are predominately C to T transition mutations, which is not seen in other cancer types (Figure 2.2).

However, mutation rate is not the only statistic capable of statistically implicating cancer drivers. One alternative is to use ratiometric features that normalize for the total number of mutations within a gene. For example, the ratio of non-silent to silent mutations within a gene is relative to silent mutations. Figure 2.1 shows the variability of the median ratio of non-silent to silent mutations for cancer types in our pancancer set. Ratiometric features had significantly less variability among cancer types than background mutation rates. The considerably lower variability suggests less covariates would need to be modeled when developing a statistical model of somatic mutations.

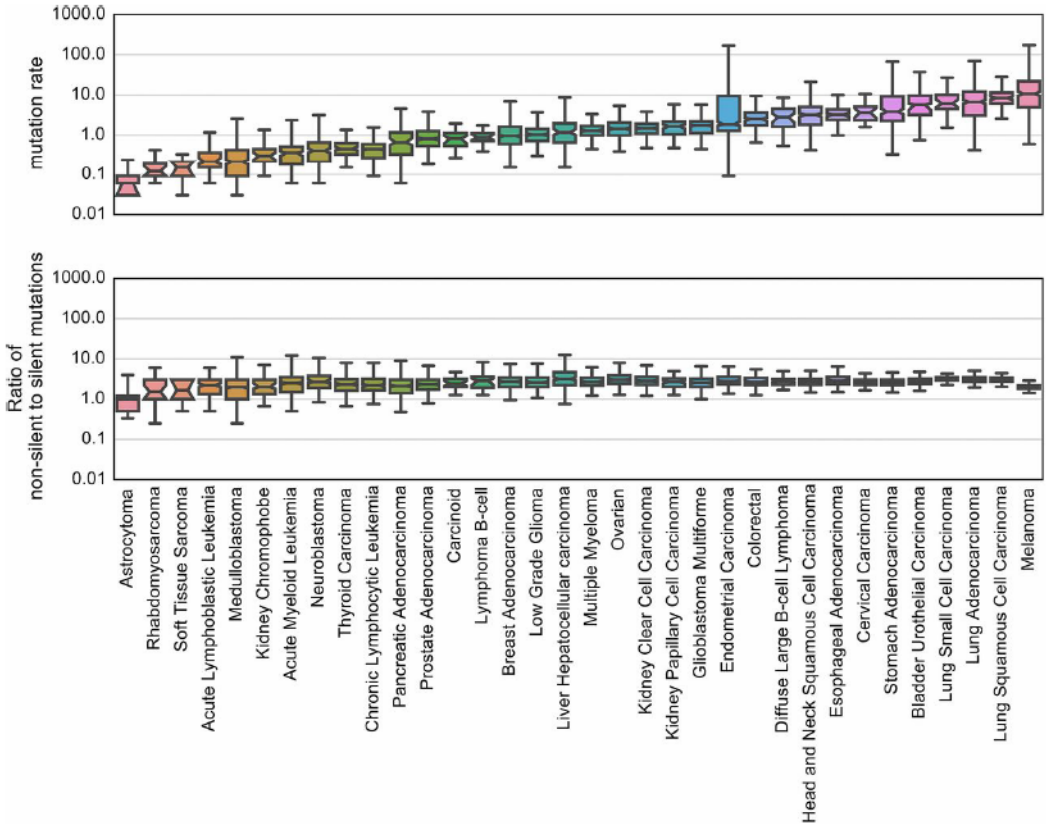

Figure 2.1: Background mutation rate is more variable than the ratio of nonsilent to silent mutations across 34 cancer types in data from [69, 89]. Boxplots are plotted on a log10 scale. The top boxplot shows the mutation rate in coding sequence for the samples in our pancancer dataset. The bottom boxplot shows the ratio of nonsilent to silent mutations in coding sequence for the same samples. A pseudocount for a silent mutation was added for each sample to avoid dividing by zero. Notches indicate bootstrap 95% confidence interval (1,000 iterations) for the median. Outliers, defined as 1.5*IQR away from the first and third quartile, are not shown.

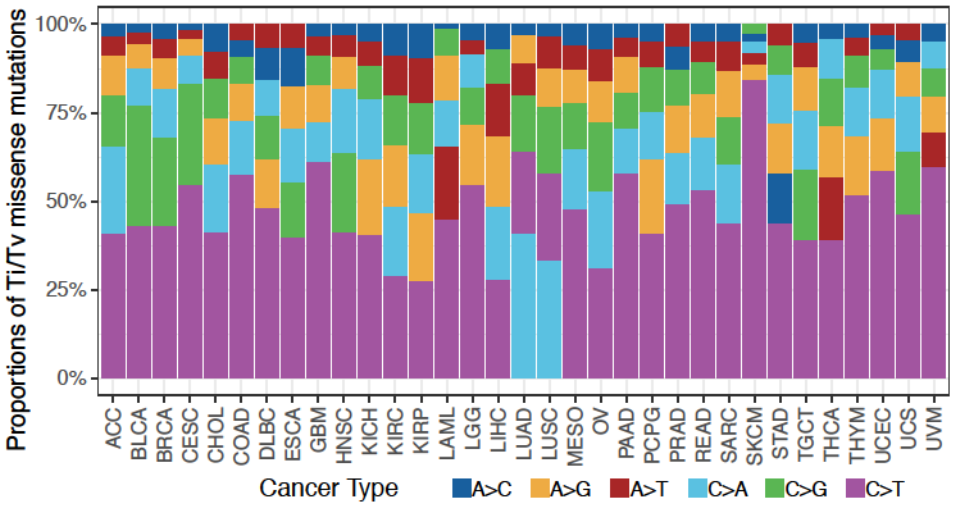

Figure 2.2: Transition and transversion proportions are shown for 6 nucleotide changes from 33 cancer types available from the The Cancer Genome Atlas (https://synapse.org/MC3). The stacked proportion bar chart is sorted by increasing transition/transversion fraction.

2.2 Expected consequence of variable background mutation rate

2.2.1 Increased mutational heterogeneity results in reduced statistical power or increased false positives

I analyzed the possible impact of unexplained variability in background mutation rate on expected false-positive driver gene predictions. First, I applied a

binomial model previously used for driver gene detection power analysis [69]. The model assumes a gene-specific background mutation rate μ, which is set to a relatively high value, corresponding to genes in the 90th percentile of genes for mutation rate. I used the binomial distribution to set the critical value for driver gene prediction, that is, the number of mutations required for a gene to be considered significantly different from the background. Next, I modeled the situation where the genes actually had mutation rates that varied around μ using a beta-binomial model. I estimated the false positives expected under the binomial, after a highly conservative multiple-testing correction (Bonferroni), for levels of variability [beta-binomial coefficients of variation (CVs)], and for sample size ranging up to 8,000 (Figure 2.3). Levels of variability defined by CVs (CV = 0.05, 0.1, and 0.2) were chosen to approximate low, medium, and high unexplained variation around the mean. As the number of samples increased, so did the number of expected false positives. At the low end of background mutation rates (0.5 mutations per megabase (MB)), the expected false positives remained low, even when 8,000 samples were evaluated, regardless of the level of variability. At an intermediate background mutation rate of 3.0 mutations per MB and with high unexplained variability, 1,000 false positives were expected from 8,000 samples. At a high background mutation rate (10.0 mutations per MB), both medium and high unexplained variability produced many thousand expected false positives.

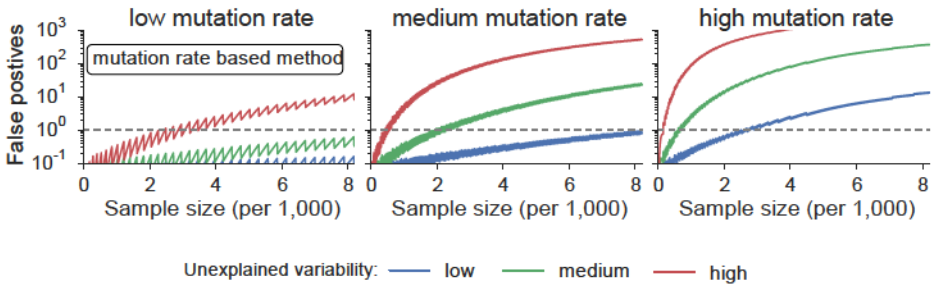

Figure 2.3: Expected false positives for a mutation rate-based predictor that identifies genes with increased mutation rate over background.

I reasoned that unexplained variability might also have an impact on power calculations to estimate how many samples must be sequenced to find the majority of cancer driver genes. To this end, I repeated previous calculations performed with a binomial power model, in which the required sample size was estimated to be 600-5,000 per cancer type [69]. The previous analysis was parameterized to detect intermediate frequency driver genes, having 2-20% mutation rates above background per sample, with background defined by genes in the 90th percentile of background mutation rates. First, I calculated the sample size required to detect 90% of these drivers, given exome-wide backgrounds of 0.1-10 mutations per MB, and a conservative estimate of 2% effect size (see subsection 2.2.2 for details). Next, I calculated the sample size required if the gene mutation rate varied around the original estimate, using a beta-binomial model with different CVs (CV = 0.05, 0.1, 0.2). The binomial power model was in accord with previous estimates. However, when unexplained variability was

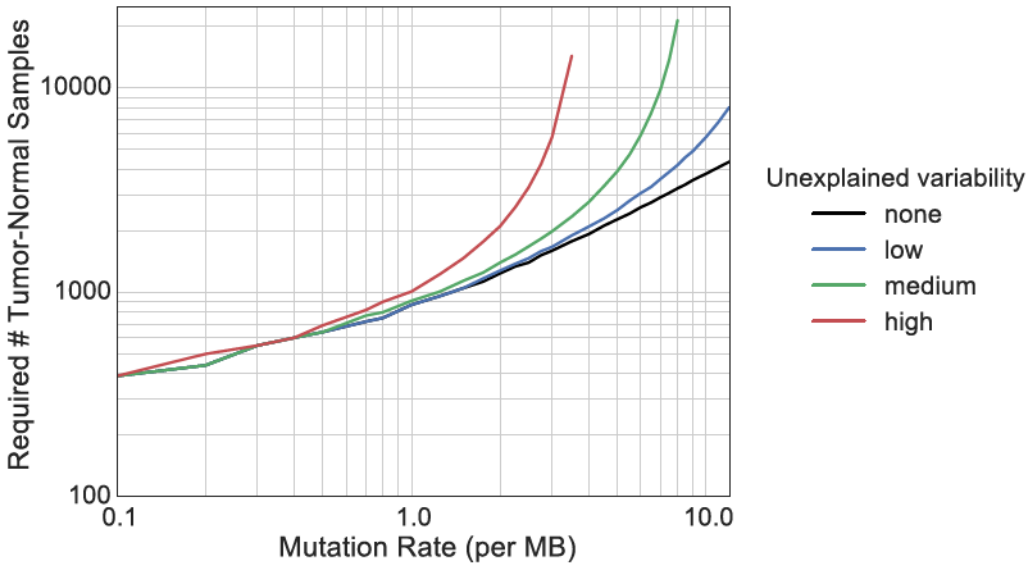

Figure 2.4: Sample size required for near-comprehensive detection of intermediate-effect driver genes (90% detection and 2% effect size/increase with respect to background). Results are shown for scenarios with no unexplained variability (black), low (blue), medium (green), and high (red) unexplained variability (CVs of 0.0, 0.05, 0.1, and 0.2, respectively). The number of required samples for the mutation rate-based method becomes very large for moderate-to-high mutation rates and levels of unexplained variability.

taken into account, the number of required samples increased sharply, particularly for higher background mutation rates (Figure 2.4).

2.2.2 Statistical models of mutation rate

Now I will describe the implementation of the statistical models used to evaluate the effects of unexplained variability in the mutation rate on false positives and statistical power. The first model assumes a correctly estimated background mutation rate μ for a particular gene (binomial model) and the second model assumes that gene background mutation rate varies around μ

(beta-binomial model). I used a binomial model similar to previously developed for driver gene power analysis [69]. The gene-specific mutation rate factor F_g calculated by MutsigCV [69] was set to represent a gene at the 90th percentile, given an exome-wide background mutation rate of π, so that $\mu = F_g\pi$ (F_g = 3.9). Average gene length (L) was set to 1,500 bases and 3/4 mutations were assumed to be non-silent. Effective gene length for non-silent mutations was therefore adjusted as $L_{eff} = 3/4L$. Gene background mutation rate was calculated using the total number of potentially mutated bases that could yield a non-silent mutation (N_{eff}), which is the effective gene length multiplied by number of samples (S). A predicted driver was defined as a gene with significantly higher non-silent mutation rate per base than that gene's background mutation rate, where non-silent mutation rate per base is the following:

$$\mu^{es} = 1 - ((1-\mu)^{L_{eff}} - r)^{1/L_{eff}} \tag{2.1}$$

and r is the fraction of samples with non-silent mutations in the gene above background. Exome-wide background mutation rates of (π = 0.5e-6, 3e-6, or 10e-6) were considered.

The beta-binomial was designed to model several levels of unexplained variability around μ. To parameterize the beta-binomial with low, medium, and high variability levels, I used different coefficients of variation (CVs) for the mutation rate (0.05, 0.1, 0.2). Beta-binomial α and β parameters were com-

puted as follows:

$$\alpha = \mu \left(\frac{\mu(1-\mu)}{(CV * \mu)^2} - 1 \right) \tag{2.2}$$

$$\beta = (1-\mu) \left(\frac{\mu(1-\mu)}{(CV * \mu)^2} - 1 \right) \tag{2.3}$$

To compute the number of false positives expected from a binomial model when unexplained variability is present, I examined the probability that the number of mutations in a gene from a beta-binomial model (K_{bb}) would meet or exceed the critical value (for a genome-wide significant driver gene at α = 5e-6) by the binomial, k'_b:

$$E[FP] = g * P_{\mu, N_{eff}}[K_{bb} \geq k'_b] \tag{2.4}$$

where g is the total number of human genes (assumed 18,500) and both models use the same mean mutation rate μ and total number of potentially mutated bases N_{eff}.

A similar model is applicable to the effect of various levels of unexplained variability in mutation rate on the power to detect driver genes. I reproduced the binomial model power analysis of [69] to estimate the number of samples required for 90% power to detect genes in the 90th percentile of gene-specific background rate, with 2% mutation rate above background (r = 0.02). Using Equations 2.2 and 2.3 to parameterize the beta-binomial model, I calcu-

lated the number of samples required for 90% power at a Bonferroni genome-wide significance level of 5e-6. Samples were iteratively added until there was greater than or equal to 90% probability that a driver gene with mutation rate μ^{es} would be found significant.

As a technical detail, discrete distributions often do not obtain exactly the stated significance level, but rather achieve at least the target significance level. Depending on the precise critical count threshold, the actual significance level varies on how overly conservative it is. This results in a jagged power curve for discrete data [90], and consequently I found the minimum number of samples required to achieve 90% power.

2.3 A Monte Carlo simulation approach

2.3.1 Implementation

Given the previously highlighted limitations of using the mutation rate, I decided to instead model ratiometric features and statistically condition on the total number of mutations within a gene. This strategy tries to limit the effect of nuisance covariates influencing mutation rate that are not always measured or known. Briefly, for each gene, single nucleotide somatic mutations were moved with uniform probability to any matching position in the gene se-

quence, holding the total number of single nucleotide somatic mutations fixed (Fig. 5). A matching position was required to have the same nucleotide base context (C*pG, CpG*, TpC*, G*pA, A, C, G, T) as the observed position. This method of generating a null distribution controls for the particular gene sequence, gene length, and mutation's nucleotide sequence context. The number of somatic mutations remains the same, but the mutation consequence of a somatic mutation may change. For example, a somatic mutation that generates a missense mutation may generate a nonsense mutation in its new position. Since mutations that result in insertions and deletions will not change their mutation consequence type by being randomly moved to another position in the same gene, they were moved to a random position in a different gene. The gene was selected based on a multinomial model, with probability proportional to the coding DNA sequence length of the originating gene. This results in a multinomial model with a large number of categories (equal to the total number of protein-coding genes) and number of trials being the total number of indel mutations.

The simulated mutations allow calculating the statistical significance of an arbitrary test statistic computed from the mutation data. Let's say there is a function T of some set of mutation(s) M. I can then compute an estimated p-value based on the simulated mutations M^0 as follows,

$$\widehat{P(M)} = \frac{\#\{T(M_i^0) : T(M_i^0) \geq T(M), i \in 1..S\}}{\#\{T(M_i^0) : i \in 1..S\}} \tag{2.5}$$

where $\widehat{P(M)}$ is the estimated p-value and S is the total number of simulations.

2.3.2 Comparison to simulations in CHASM

A related simulation method was first established in the method Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM) [39]. However, there were several limitations that needed to be addressed. First, mutations are now simulated in a cohort-level manner, rather than considering each unique mutation in isolation. This allows computation of test statistics that may be a function of many mutations found in a single region, a single gene, or in multiple genes. Second, the original CHASM simulations assumed a background rate for mutations at certain nucleotide contexts (termed 'passenger tables'). Instead, I condition on the observed nucleotide sequence context and randomly select another position with the same context. Third, I do not assume a homogenous mutation rate for single nucleotide mutations in genes across the genome (highlighted as problematic above). Fourth, my Monte Carlo simulations also apply to all coding mutations, rather than just missense mutations. Lastly, an issue with the original simulations by CHASM was that it

assumed every mutation in a cancer driver gene was a cancer driver, which we know not to be the case. As a consequence, the simulations from CHASM blacklisted all simulated mutations in likely cancer driver genes. Here, I do not blacklist simulated mutations in driver genes.

2.3.3 Application to salivary gland adenoid cystic carcinoma (ACC)

2.3.3.1 ACC overview

As part of the statistical analysis, I analyzed coding mutations from 25 whole-genome sequenced Adenoid Cystic Carcinomas (ACC) [91]. Specifically, it was noticed that several chromatin regulator genes had more than one non-silent mutation. The question was whether these chromatin regulator genes had a high proportion of truncating mutations, suggesting that they could be tumor suppressor genes.

2.3.3.2 Model of truncating point mutations

I performed a randomization-based statistical test of increased proportion of truncating mutations (K) out of total non-silent mutations (N) for genes involved in chromatin regulation, controlling for the effect of gene sequence

and the nucleotide sequence context of the mutation. For each gene i, our test statistic was

$$T_i = \frac{\#\{t : t \in K\}}{|N|}, where K \subset N \tag{2.6}$$

Truncating mutations were defined as any nonsense, consensus splice-site mutations, or out-of-frame insertions/deletions (frameshift). Monte Carlo simulations were performed to approximate the null probability distribution of the test statistic T_i. Because frameshift mutations do not change consequence when moved to a different position, in the Monte Carlo sample, they were retained with probability equal to the observed proportion of frameshift mutations out of all mutations (maximum likelihood estimate), otherwise they were changed to a non-truncating mutation. After each iteration of this sampling procedure, the number of mutations in a gene is always the same, but the mutation consequence of each mutation may change. Thus, the test statistic T_i for the gene will change values at each iteration, and repeated iterations yield a null distribution of test statistics to estimate the P value of the gene's observed test statistic. For the gene group analysis, my test statistic was

$$T_c = \frac{\sum_{i \in c} \#\{t : t \in K_i\}}{\sum_{i \in c} |N_i|} \tag{2.7}$$

and it was computed both for the observed and simulated mutations. A one-

tailed empirical P value was calculated as the fraction of Monte Carlo samples in which the observed value of the test statistic was equal to or higher than the simulated value. Increasing the number of iterations of Monte Carlo sampling increases the precision of the P value; 10,000,000 iterations were chosen to achieve adequate precision.

2.3.3.3 Results

Several genes with well-known roles in chromatin regulation were mutated in multiple tumors: *MLL2*, *MLL3*, *EP300*, *SMARCA2*, *SMARCC1*, and *KDM6A*. The proportion of truncating mutations (nonsense codons, splice-site alterations, or out of-frame insertions and deletions) out of the total number of non-silent mutations in these genes was high (6 of 11), significantly greater than expected by chance (P = 3.8e-6). Furthermore, *MLL2* and *EP300*, when considered individually, had a significantly higher proportion of truncating mutations than expected by chance (P = 0.008 for *MLL2* and P = 0.01 for *EP300*). This finding is consistent with the hypothesis that several of these genes played an important role in these cancer samples.

2.4 Conclusions

The goal in this chapter was to first understand "how" somatic mutations accumulate in the absence of selection so as to, later, correctly interpret "which" mutations are cancer drivers in the presence of positive selection. A recent study indicates there is limited purifying selection of point mutations in cancer [32], suggesting the lack of incorporating negative selection in statistical models is not a major concern. I have shown in this chapter that background mutation rate is highly variable at multiple scales and therefore is difficult to statistically model. This can either lead to increased false positives or reduced statistical power when attempting to identify cancer driver genes. However, many of the known covariates are not at the same resolution as genes, as they vary across the genome at the scale of megabases [88], while nearly all genes span <1MB. I therefore developed Monte Carlo simulations to test any arbitrary test statistic by conditioning on the total number of mutations within a gene while accounting for nucleotide sequence context. The flexibility of the Monte Carlo simulations will be critical in later chapters when evaluating the significance of results from machine learning methods.

Chapter 3

20/20+: a machine learning method to predict cancer driver genes

3.1 Introduction

The first exomic analyses attempted to identify candidate driver genes as those having more mutations than expected from some presumed background somatic mutation rate, corrected for base context, gene size, and other variables [19,92]. Subsequent work has considerably refined the variables involved in determining whether a gene is more mutated in cancers than expected by chance. This has led to a variety of "significantly mutated gene" methods that

adjust for covariates such as replication timing and gene expression as well as including more sophisticated metrics of mutational base contexts [33, 73]. Although methods have been extended to utilize gene expression to identify cancer drivers [93–98], I will focus solely on driver gene analysis based on somatic point mutations.

An alternative method to finding cancer drivers employs a ratiometric approach. Rather than attempting to determine whether the observed mutation rate of a gene in cancers is higher than expected by chance, these methods simply assess the composition of mutations normalized by the total mutations in a gene. The ratiometric 20/20 rule [14] evaluates the proportion of inactivating mutation and recurrent missense mutations in a gene of interest. Other ratiometric approaches use mutation functional impact bias [40, 99], mutational clustering patterns [51, 55, 89], or patterns of mutation composition [89].

Here, I describe a machine-learning-based, ratiometric method (20/20+) that formalizes and extends the original 20/20 rule and enables automated integration of multiple features of positive selection.

3.2 Original 20/20 rule

The original 20/20 rule was a manually designed set of decision rules to identify cancer driver genes as either an oncogene or tumor suppressor gene

(Figure 3.1) [14]. The name derives from the rule that oncogenes should have at least 20% of mutations being recurrent (observed more than once) and tumor suppressor genes should have at least 20% inactivating mutations (i.e., frame shift indels, nonsense mutations, etc.). In addition, there are count thresholds that are specifically tuned for analyzing a specific version of the COSMIC database [100]. When I applied this rule to a later version of COSMIC, it had inadequate specificity (data not shown). Moreover, scaling the count thresholds relative to database size also produced similar results, which suggests additional manual curation of results would be necessary. Consistent with this observation, the 20/20 rule had good performance at distinguishing oncogenes versus tumor suppressor genes within already pre-defined driver genes [101].

3.3 Machine learning prediction of cancer driver genes

20/20+ utilizes a machine learning algorithm called Random Forests to predict cancer driver genes. The advantage of machine learning is that multiple features can automatically be incorporated when making predictions. The random forest algorithm was chosen because they outperformed logistic regression, boosting, and support vector machines (data not shown). I will first go over the mathematical background of random forests and then describe the

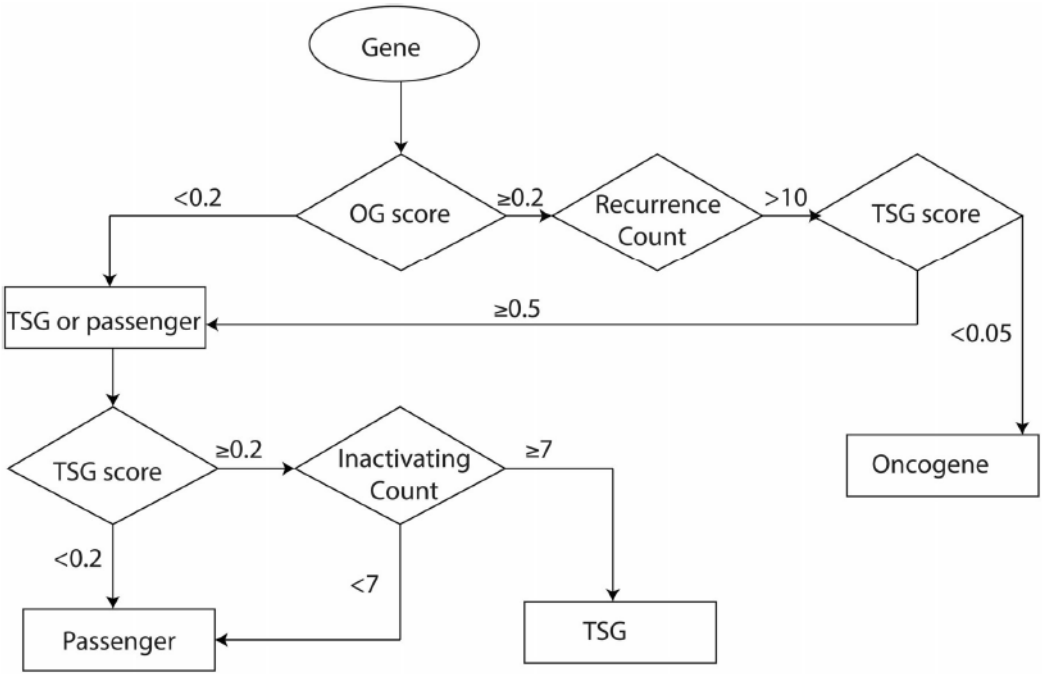

Figure 3.1: Decision tree underlying 20/20 rule. Each gene is input into the tree and oncogene (OG) and tumor suppressor gene (TSG) score computed. Thresholds of each score and the numerator of the OG score (recurrence count) and TSG score (inactivating count) are used to determine whether a gene is an OG, TSG, or passenger.

implementation in 20/20+.

3.3.1 Mathematical overview of random forests

3.3.1.1 Decision tree

The base component of a random forest is the decision tree, which, effectively, is a hierarchically organized set of questions. Since my focus will be on classification, a decision tree $\pi(X)$ will return 1 if the prediction is a cancer driver and 0 for a passenger. A decision tree is constructed from a set of possible questions $Q = \{Q_1, \ldots, Q_N\}$, with a question taking the following form,

$$Q_{\cdot}(X) = I_{X_i \leq c} \tag{3.1}$$

where I is the indicator function, X_i is the i'th feature value, and c is a constant. The question asked by a decision tree depends on the previous question asked. To keep track of order of questions, the index of the first question will be denoted ρ_1, where $\rho_1 \in \{1 \ldots N\}$. The second question will therefore be a function of the answer from the first question, $\rho_2 = \rho_2(a_1)$, where $a_1 = Q_{\rho_1}$. More generally there is a series of questions $\beta_n = \{Q_{\rho_1} = a_1, \ldots, Q_{\rho_{n-1}} = a_{n-1}\}$ prior to the n'th question.

Each question will split training examples depending on the answer to the question. The goal of the decision tree is to utilize questions that reduce the

uncertainty in the distribution of class labels. I will regard the distribution of class labels as the probability of observing a class label given the answer to a question n as follows,

$$p_k = Pr(Y = 1|Q_n = k, \beta_n) \tag{3.2}$$

where k is the answer to the question and p_k is the proportion of labeled drivers. For the decision tree algorithm, the best question defined by minimizing the expected gini impurity at each step:

$$\rho_n = \underset{1 \leq b \leq N}{\arg\min} \, 2 \sum_{k=0}^{1} Pr[Q_b = k] p_k (1 - p_k) \tag{3.3}$$

There are various practical criteria to decide when to stop asking further questions in a decision tree, but I will not cover this here. Assuming a decision tree is constructed, the predicted class is chosen by the most likely class following the terminal question.

$$\pi(X) = \underset{l}{\arg\max} \, Pr(Y = l|\beta_*) \tag{3.4}$$

where $l \in \{0, 1\}$ is the class label and β_* contains the history of answers for all questions.

3.3.1.2 Random Forest

A random forest is an ensemble of many randomized decision trees [102, 103], where each tree is trained on a random selection of training set examples and candidate features, via a recursive splitting process [104]. This involves constructing each tree π_j from a bootstrapped sample of the training data $D_j = \{(X(i), Y(i)), \ldots\}_{i=1..m}$. Then instead of allowing all questions for each split, a random subset of questions is used $Q^s \subset Q$, were the number of features s is usually taken to be proportional to the square root of the total features p, $|s| = \sqrt{p}$. Lastly, once J decision trees are constructed, random forest predictions result in a score between 0 and 1 that reflects the proportion of trees agreeing with the class label of cancer driver:

$$f(X) = \frac{1}{J} \sum_{j=1}^{J} \pi_j(X) \tag{3.5}$$

3.3.2 Random Forest implementation

Although the above mathematical description of a random forest was in terms of two classes (drivers and passengers) for simplicity, random forest classification also extends to multi-class classification. 20/20+ uses a three-class classifier which predicts a gene as either an oncogene, tumor suppressor genes, or passenger gene. I used the set of oncogenes and tumor suppressor genes

identified by the original 20/20 rule as a training set [14]. Considering cancer driver genes (oncogenes or tumor suppressor genes) constitute only a small proportion of all genes, I labeled all other genes as passenger for the purpose of training.

3.3.2.1 Features

I designed a set of 24 predictive features and whose feature importance is shown in Figure 3.2, as assessed by the mean decrease in gini impurity. A description of the features can also be found in Table 3.1. Many of the features are components of the 20/20 rule OG and TSG scores, and I included several ratiometric features not in the original 20/20 rule, for example, ratio of missense to silent mutations, as well as features that represented mutation functional impact and gene importance. Normalized missense entropy, a measure of positional clustering, was calculated as follows:

$$E_k = \frac{-\sum_i p_i \log_2 p_i}{\log_2 k} \tag{3.6}$$

where k is the total number of missense mutations in a gene and p_i = (count of missense mutations in the I'th codon)/k. Three of the 24 features represented p-values and were calculated using the monte carlo simulation described in section 2.3.

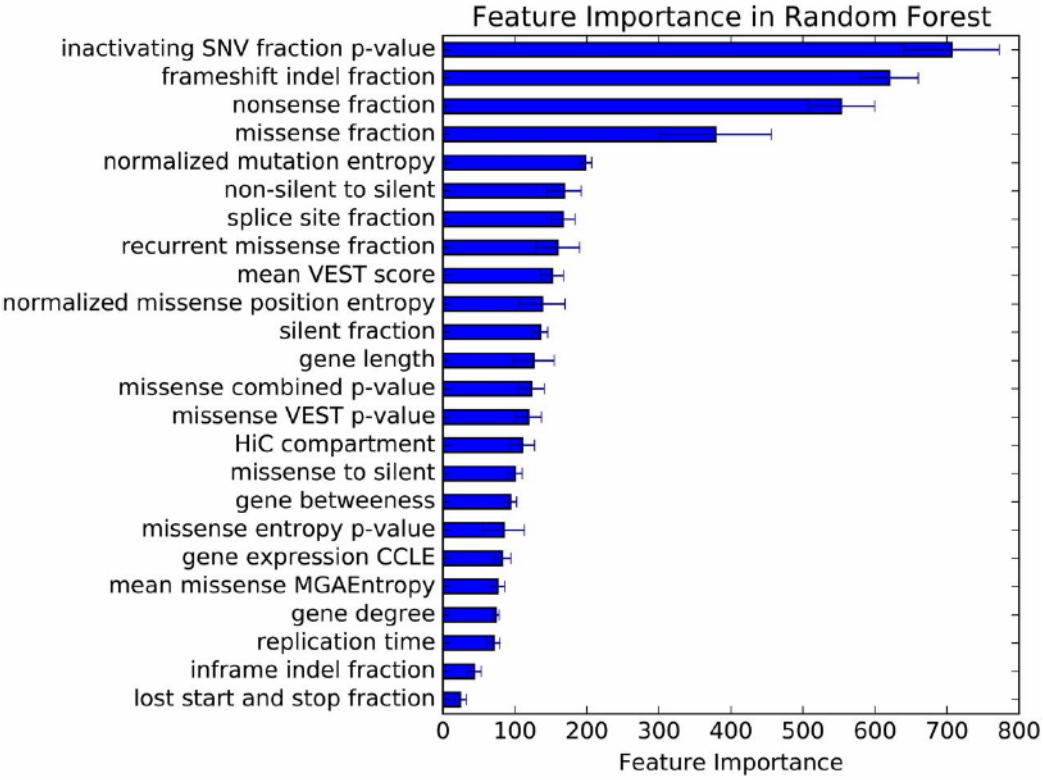

Figure 3.2: Random Forest feature importance ranking for the 24 predictive features. The mean decrease in Gini index is plotted for each feature. Error bars indicate SD when feature importance calculation was repeated on 10 different cross-validation partitions. CCLE, Cancer Cell Line Encyclopedia [33]; HiC, 3D chromatin interaction capture [33]; MGAEntropy, Shannon entropy in column of a vertebrate genome 46-way alignment corresponding to location of the mutation [105]; SNV, single-nucleotide variant; VEST, Variant Effect Scoring Tool [106].

Feature name	Source	Description
silent fraction	Calculated from mutations	Fraction of mutations that are silent mutations
nonsense fraction	Calculated from mutations	Fraction of mutations that are nonsense mutations
splice site fraction	Calculated from mutations	Fraction of mutations that are 2bp consensus splice site mutations
missense fraction	Calculated from mutations	Fraction of mutations that are missense mutations
recurrent missense fraction	Calculated from mutations	Fraction of mutations that are recurrent missense
frameshift indel fraction	Calculated from mutations	Fraction of mutations that are frameshift indel mutations
inframe indel fraction	Calculated from mutations	Fraction of mutations that are inframe indel mutations
lost start and stop fraction	Calculated from mutations	Fraction of mutations that are either lost start or lost stop mutations
normalized missense position entropy	Calculated from mutations	See above
missense to silent	Calculated from mutations	Ratio of missense to silent mutations. A pseudo count is added to silent mutations to avoid divide by zero.
non-silent to silent	Calculated from mutations	Ratio of non-silent to silent mutations. A pseudo count is added to silent mutations to avoid divide by zero.
normalized mutation entropy	Calculated from mutations	Normalized entropy score (see above). Missense mutations are binned together based on codon position. Each silent mutation is regarded in its own bin. Potentially inactivating mutations (nonsense, splice site, lost stop, and lost start) mutations are grouped into a single bin.
mean missense MGAEntropy	Calculated from mutations. MGAEntropy scores obtained from SNVBox (30).	Mean MGAEntropy score for missense mutations (30). MGAEntropy for a missense mutation is the entropy of the column for a protein-translated version of UCSC's 46-way vertebrate alignment
mean VEST score	Calculated from mutations	Mean score. Score for missense mutations are taken as the VEST score, silent mutations receive a score of 0, and other mutations receive a score of 1.
inactivating SNV p-value	Calculated from mutations	Statistical significance of proportion of inactivating mutations. SNV=single nucleotide variant.
missense entropy p-value	Calculated from mutations	Statistical significance of normalize missense position entropy
missense VEST p-value	Calculated from mutations	One-tailed tatistical significance of proportion of having higher mean VEST score for missense mutations
missense combined p-value	Calculated from mutations	Combined p-value composed of missense entropy and missense VEST p-value using Fisher's method
gene degree	BioGrid	Number of other genes that are connected in the BioGrid interaction network
gene betweenness centrality	BioGrid	Fraction of shortest paths that pass through a gene's node in the BioGrid interaction network
gene length	Longest SNVBox transcript	CDS length of reference transcript
expression CCLE	MutsigCV (4)	Average expression of a gene in the Cancer Cell Line Encyclopedia
replication time	MutsigCV (4)	DNA replication time during cell cycle
HiC compartment	MutsigCV (4)	HiC measures open vs closed chromatin

Table 3.1: Features used in 20/20+. Features use mutations that are small somatic variants, including single base substitutions and small insertions/deletions. CCLE = cancer cell line encyclopedia. SNV = single nucleotide variant. SNVBox = database of features of single nucleotide variants. Biogrid = database of gene networks.

3.3.2.2 Handling class imbalance

With only 54 OGs and 71 TSGs labeled by the original 20/20 rule, the number of passenger genes far exceeds the number of labeled driver genes, creating a problematic class imbalance [107]. I used a subsampling approach, previously recommended for random forests [108], in which, for my case, the passenger genes are sampled at a 1:1 ratio to OGs plus TSGs. To compensate for the remaining OG and TSG imbalance, the Random Forest is trained with class weights inversely proportional to the sampled frequency of the class. Predictions were made with a random forest of 200 trees. The number of trees only had minor impact on the overall performance.

Because of the limited number of driver genes, I decided to use 10-fold cross-validation instead of a train-test split of the data to evaluate performance. Out-of-bag estimates of performance by random forests is also an alternative, but it does not generalize to scenarios with biologically correlated predictions. The procedure of 10-fold cross-validation was repeated five times (1,000 trees in total), and the resulting scores from each gene were aggregated to limit minor fluctuations in scores due to randomization in the cross-validation folds.

3.3.2.3 Random forest prediction

Each gene was scored as the fraction of trees that voted for oncogene, tumor suppressor gene, or passenger gene. A driver score for a gene was calculated as

the sum of the oncogene and tumor suppressor gene scores. The purpose of a general driver score was included for cases where the gene was likely a cancer driver but the precise type of cancer driver gene was hard to determine (i.e. oncogene or tumor suppressor gene).

3.4 Statistical significance

The statistical significance of each gene score was computed with an extension of the Monte Carlo simulation algorithm described in section 2.3. For each gene, the Monte Carlo simulation was repeated 10 times, and for each simulation all 24 features were computed. In this process, protein interaction network features (degree and betweenness) were, additionally, permuted as a pair. The features of gene length, replication timing, HiC value, and average Cancer Cell Line Encyclopedia (CCLE) gene expression were not altered. Next, each "simulated" gene was scored with the Random Forest previously trained on the real data. The resulting OG, TSG, and driver scores for all simulated genes were used as an empirical null distribution. To compute a P value for a gene score, I used the fraction of simulated genes with a score equal to or greater than the score. P values were adjusted by the Benjamini-Hochberg method [109] for multiple hypotheses. I compute a Benjamini-Hochberg q-value as follows,

$$q(i) = \min\left(\min_{i..n} \frac{p(i)}{i/n}, 1\right) \tag{3.7}$$

where $p(i)$ is the i'th smallest p-value, $q(i)$ is the corresponding q-value, n is the total number of p-values, and $\min_{i..n}$ is the cumulative minimum from index n to i. Consistent with other driver gene prediction methods, I considered a gene to be significant ($q \leq 0.1$) if any of the OG, TSG, or driver scores were significant. The strategy of converting p-values to q-values is for convenience and does not change the significant p-values by the procedure originally outlined by Benjamini and Hochberg [109].

3.5 Conclusions

In this chapter, I developed a new method, 20/20+, which addresses two major issues with the original 20/20 rule: use of a limited number of features and a lack of a statistically principled threshold to limit false positives. 20/20+ uses the random forest algorithm to make predictions using a non-linear combination of features. Importantly, 20/20+ uses ratiometric features to normalize artifactual differences between cancer types and regions of the genome. Because Random Forests do not intrinsically include hypothesis testing techniques, I used simulated mutations to assess the statistical significance of scores. P-values were estimated from a simulated null distribution, controlling for se-

quence composition, and corrected for multiple testing with the Benjamini-Hochberg method [109]. The application and benchmarking of 20/20+ will be considered in Chapters 4 and 7.

Chapter 4

Benchmarking cancer driver gene predictions

Rigorous and unbiased evaluation is necessary to inform users about the comparative utility of prediction methods. In many investigative domains, there is a generally accepted gold standard against which predictions can be benchmarked [110, 111]. However, only a limited number of genes have been fully vetted as cancer drivers. In previous work, driver prediction has been benchmarked by significant overlap with the Cancer Gene Census (CGC) [112], which is a manually curated list of likely but not necessarily validated driver genes [40,55] or by agreement with a consensus gene list of drivers predicted by multiple methods [70]. To our knowledge, a systematic framework for the evaluation of somatic mutations that can be generally applied has not been previ-

ously developed. Eight methods were evaluated: MutsigCV [33], ActiveDriver [71], MuSiC [73], OncodriveClust [55], OncodriveFM [40], OncodriveFML [99], Tumor Suppressor and Oncogenes (TUSON) [89], and 20/20+ [85].

In this chapter, I present a framework for such evaluations. The framework has five components, some of which have been previously applied in isolation, but not as part of a unified system. I considered overlap with CGC, agreement between methods, comparison of observed vs. theoretical P values, number of significant genes predicted, and prediction consistency on independent partitions of the dataset. To implement this framework, I first collected 729,205 published somatic mutations from 34 cancer types (Figure 4.1) [69, 89]. These mutations were composed of single base substitutions and in-frame and out-of-frame insertions and deletions (indels) of less than 10 bp. I then compared various methods on the full pancancer set.

4.1 Overlap of the Driver Genes Predicted by Each Method

First, I assessed overlap of the predicted driver genes with the CGC. I considered only those CGC genes typed as somatic, missense, frameshift, nonsense or splice site, excluding translocations, large amplifications/deletions, and other mutation consequence types not addressed in our study, yielding

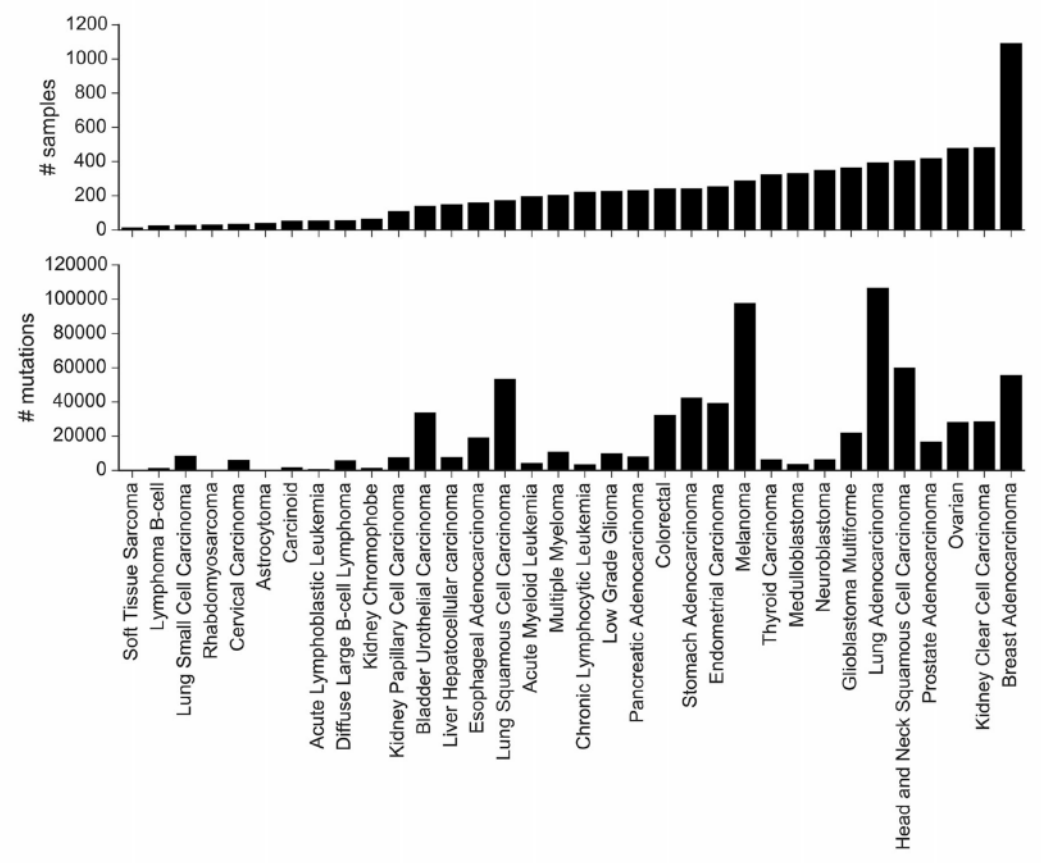

Figure 4.1: Summary of evaluation dataset. The evaluation dataset consisted of mutations spanning 34 cancer types. All included mutations were small somatic variants. Cancer types are ordered from Left to Right by number of samples, ranging from 15 for soft-tissue sarcoma to 1,093 for breast adenocarcinoma, with an average of 232 samples per cancer type. These cancer types span a wide range of solid and several liquid cancers, including multiple tissues and cell types of origin, different background mutation rates, and different numbers of available samples. For each cancer type, total mutations and number of available samples are shown.

a total of 188 CGC genes. Although the driver genes predicted by all methods were enriched for CGC genes, the predicted drivers by any individual method did not contain a majority of CGC genes (Figure 4.2A). Three methods (20/20+, MutsigCV, and TUSON) had substantially higher fractions of predicted drivers in the CGC than the other methods. When I considered a subset of 99 CGC genes supported by functional studies [68], the results were very similar. The ranking of methods by fraction predicted was essentially the same as with the full CGC, with the three methods listed above having substantially higher fractions than the rest.

Genes predicted by more than one method may be more likely to be true drivers [70]. For each method, I calculated the fraction of predicted drivers that were unique or predicted by at least one, two, or three other methods (Figure 4.3). As shown in Figure 4.3, there was little consensus in prediction of driver genes among the methods. The majority (59-80%) of genes identified by MuSiC, ActiveDriver, OncodriveClust, OncodriveFML, or OncodriveFM were not observed by any of the other seven methods. The fractions of genes identified by TUSON, 20/20+, and MutsigCV that were not identified as driver genes in at least one of the other seven methods was 14%, 19%, and 33%, respectively. Although it is likely that some of the uniquely predicted drivers are bona fide, I could not find convincing literature support for the top-ranked unique predictions of MuSiC, ActiveDriver, and the Oncodrive methods.

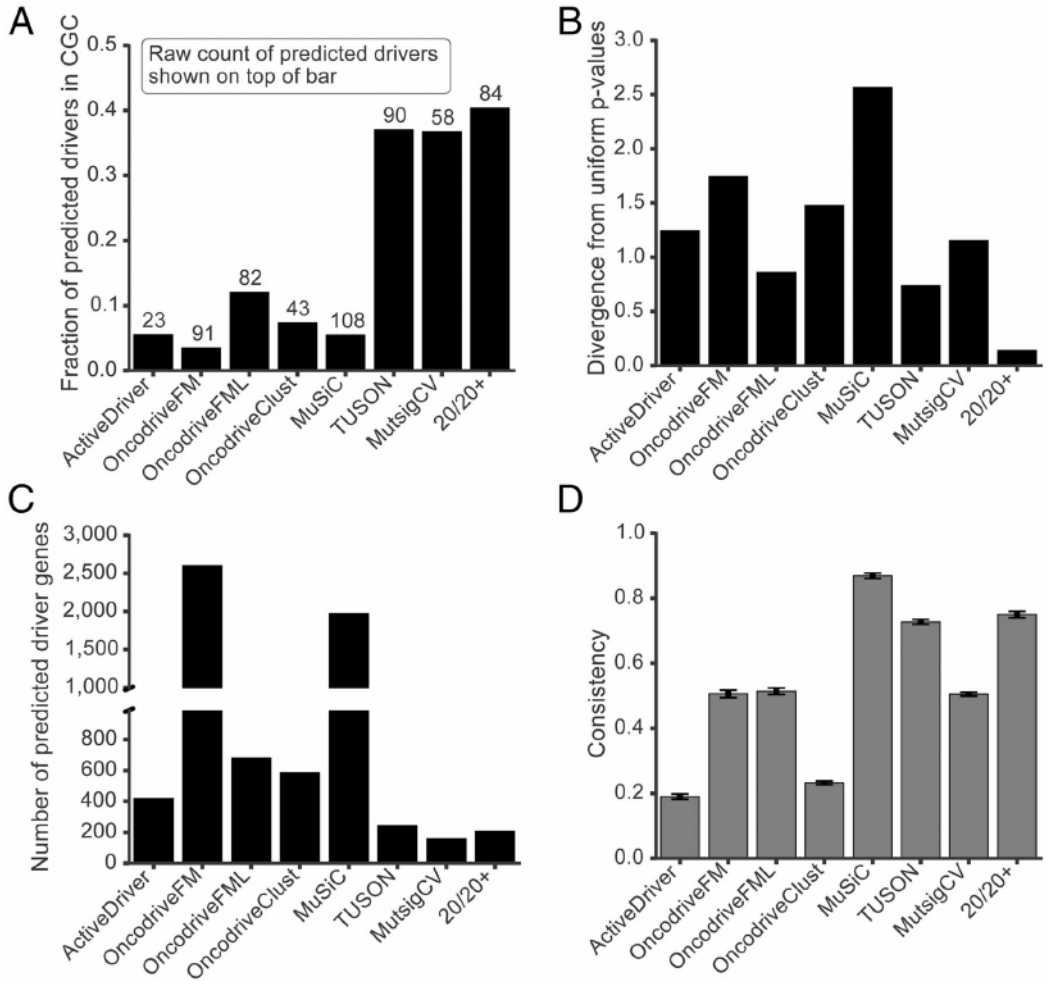

Figure 4.2: Outputs of eight driver prediction methods run through the evaluation protocol. (A) Fraction of predicted driver genes (q 0.1) that are found in the Cancer Gene Census (CGC) (downloaded April 1, 2016). Raw count of predicted driver genes indicated on Top of each bar. (B) Divergence from uniform P values, measured as mean log fold change (MLFC) between a method's observed and desired theoretical P values. (C) Number of predicted driver genes. Driver gene is defined as having Benjamini-Hochberg adjusted P value, $q \leq 0.1$. (D) Consistency of each method measured by TopDrop consistency (TDC) at depth of 100 in the method's ranked list of genes. Error bars indicate +/-1 SEM across 10 repeated splits of the data.

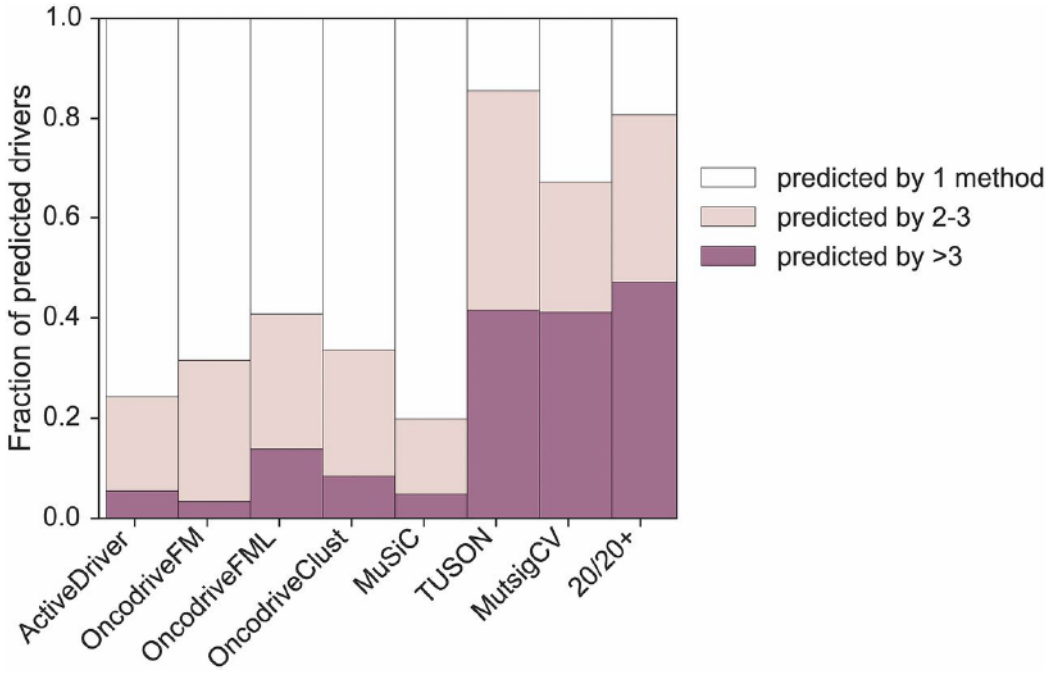

Figure 4.3: Fraction of predicted driver genes for each method by consensus among methods. Fraction of predicted drivers unique to each method, predicted by two to three methods or predicted by more than three methods are shown. A predicted driver gene is defined by Benjamini-Hochberg adjusted P value ($q \leq 0.1$)

4.2 Observed vs. Expected P Values

Given the lack of agreement among these various methods, I compared P values reported by each method to those expected theoretically. Such comparisons are often used in statistics and can indicate invalid assumptions or inappropriate heuristics. Theoretically, the P value distribution should be approximately uniform after likely driver genes are removed [113]. Therefore, I removed all genes predicted to be drivers by at least three methods after Benjamini-Hochberg multiple-testing correction ($q \leq 0.1$) and any remaining genes in the CGC. I assumed that the number of bona fide driver genes not removed by this procedure would be small enough to have minimal impact on the P value distribution. To quantify the differences between the observed P values and those expected from a uniform distribution, I developed a measure named mean absolute log2 fold change (MLFC) (see subsection 4.2.1). MLFC values near zero represent the smallest discrepancies and the closest agreement between observed and theoretical P values.

One method (20/20+) had an MLFC that was fivefold lower than the seven others (Figure 4.2B). I also compared observed and theoretical P values with quantile-quantile plots, which provide a detailed view of P value behavior (Figure 4.4A). 20/20+ P values had by far the best agreement with theoretical expectation across the entire range of supported values. In the critical range

typically used to assess statistical significance ($P \leq 0.05$), OncodriveClust, OncodriveFM, OncodriveFML, ActiveDriver, and MuSiC substantially underestimated P values, whereas MutsigCV substantially overestimated them (Figure 4.4B). For methods that combine multiple P values for each gene, failure to model correlation between P values may be responsible for this underestimation. The null P value distributions at the other end of the distribution (0.2-1.0) should also be uniform and in this case independent of the actual number of true driver genes. This is because regardless of whether all cancer driver genes are known and eliminated from the P value distribution, only a small proportion of all genes contain potential driver mutations.

4.2.1 MLFC and mathematical justification

The MLFC is a metric of discrepancy between an observed P value distribution reported by a method and a theoretical uniform null distribution. I define MLFC as follows,

$$MLFC = 1/n \sum_{i=1}^{n} \left| \log_2 \frac{P(i)}{i/n} \right| \tag{4.1}$$

where P(i) = i'th smallest P value, n is the total number of genes after excluding likely driver genes, and i/n is the corresponding expected P value from a uniform distribution. Values of MLFC near zero indicate smaller discrep-

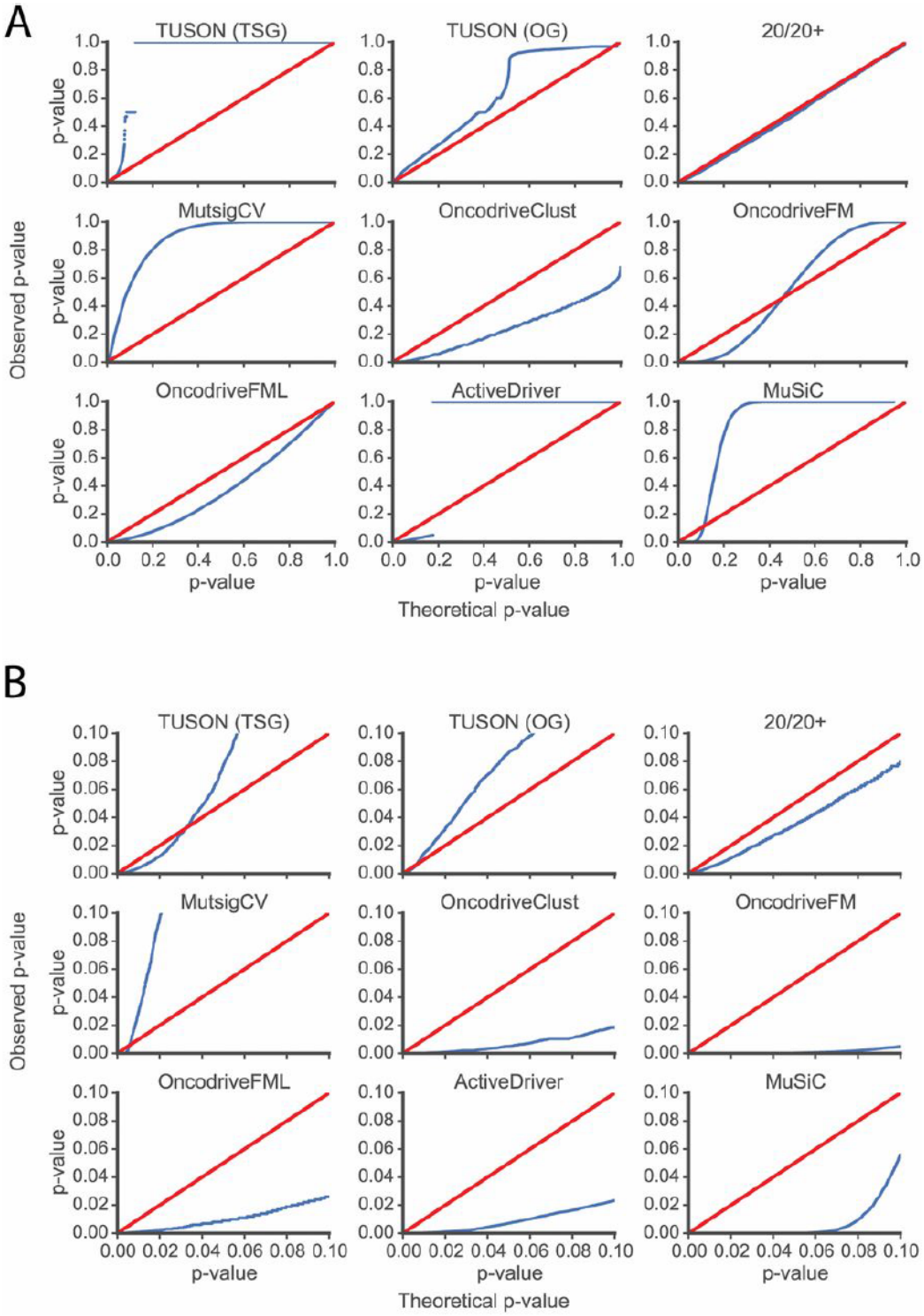

Figure 4.4: (Caption next page.)

Figure 4.4: (Figure: previous page) Quantile-quantile plots comparing observed and theoretical P values for the tested methods. (A) Full P value range from 0 to 1. (B) Blowup of P values from 0 to 0.1. Observed P values for the methods (blue) are compared with those expected from a uniform distribution (red). Genes predicted as drivers by at least three methods were removed along with genes in the CGC. TUSON OG and TSG P values are shown separately.

ancies, and therefore better statistical modeling of the passenger gene null distribution.

The absolute value was included in the MLFC to prevent p-value distributions which show evidence of over-dispersion from effectively canceling out over-estimation of p-values with under-estimation at other ranges of the p-value distribution (see OncodriveFM, Figure 4.4A). To get a better understanding of the effect of over-dispersion on reducing MLFC values without the absolute value term, I analyzed a simple case involving two normal distributions. Say the significance of a statistic is measured against the standard normal distribution centered at zero (standard deviation=1, mean=0), while in actuality the data is generated from a normal distribution with the same mean but greater variance (standard deviation $\geq$1, mean=0). If the number of generated samples is 18,500, roughly matching the total number of genes, than the MLFC metric without an absolute value systematically underestimates the divergence of the p-value distribution compared to the MLFC with the absolute value Figure 4.5. There is no difference between the MLFC values when there is no over-dispersion, and it grows as the standard deviation gets larger.

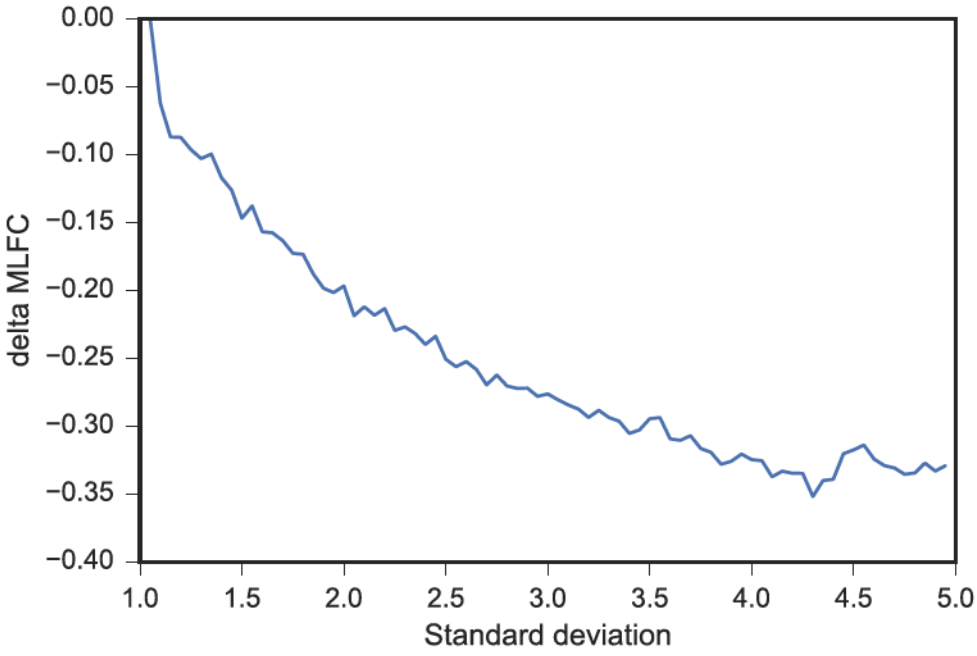

Figure 4.5: Underestimation of MLFC when absolute value is not taken into consideration. P-values were calculated from a standard normal distribution but data was generated from normal distributions with greater standard deviation.

There is a close connection between the MLFC and the Benjamini-Hochberg (BH) method [109] to control the false discovery rate, which is used by all methods evaluated in the benchmark. The BH procedure rejects hypotheses by selecting the largest index g such that,

$$\frac{P(i)}{i/n} \leq q^* \tag{4.2}$$

where q^* is the desired false discovery rate control, and then rejecting all hypotheses $H(i)$ from index $i = 1, .., g$. The critical part of the MLFC equation is the ratio of observed p-value to expected, which is intentionally the same statistic used in the Benjamini-Hochberg method. There are alternative approaches for testing whether a probability distribution differ from theoretical expectations, such as the Kolmogorov-Smirnov (KS) test [114]. The KS test evaluates the significance of the maximum absolute differences between cumulative distribution functions. However, given the large number of hypothesis tests, even small absolute differences in the cumulative distribution at small p-values may result in many false positives. Moreover, I reasoned that in contrast to MLFC, the KS statistic does not directly relate to the BH procedure and disproportionately focuses on discrepancies at large p-values, which are of less practical interest.

4.3 Number of Predicted Driver Genes

The number of predicted driver genes ($q \leq 0.1$) ranged from 158 (MutsigCV) to 2,600 (OncodriveFM) (Figure 4.2C). There were two obvious groups of methods with respect to predicted driver genes: MutSigCV, 20/20+, and TUSON predicted 158-243 genes, whereas the remaining had over 400 driver genes.

4.4 Driver Gene Prediction Consistency

Statistical methods suffer from both systematic and random prediction errors. When no gold standard is available, it is difficult to estimate systematic error, but possible to estimate random error by measuring the variability of predictions. I tested the eight methods on 10 repetitions of a random two-way split of the all samples in our dataset, while maintaining the proportion of samples in each cancer type. An ideal method would produce the same list of driver genes, ranked by P value, for each half of the split. For a fair comparison, I considered that methods predicting many drivers would be less likely to have consistent rankings than those predicting only a few. Thus, I developed a measure named TopDrop consistency (TDC) (see subsection 4.4.1) that examines the overlap between genes ranked at a defined depth (e.g., the top 100 genes) for each half of the random split. Examining TDC at a depth of 100 genes showed MuSiC, 20/20+, and TUSON to be the three with the high-

est consistency (Figure 4.2D). Most methods decreased in consistency when the gene depth was varied between 20 and 300, but the ordering of the TDC scores among the eight methods remained relatively stable.

4.4.1 TopDrop consistency and limitations

Consistency assesses stability in gene ranking. Each method was applied to 10 repeated random splits, consisting of two disjoint halves of the full data. For pancancer assessment, the proportion of samples from each cancer type was maintained in each half. Disjoint halves were scored separately by each method, and genes were ranked from low to high P values. For a fair comparison between methods, I considered a specific depth of top-ranked genes, rather than a fixed q value threshold. This is because consistency becomes harder to achieve as the number of top-ranked genes gets large. For example, a method that predicts 100 significant genes at $q \leq 0.1$ has an advantage in consistency over a method that predicts 1,000 significant genes at that threshold. I define $TopDropconsistency = |I_d|/d$, where d is the designated depth of interest for the ranked gene list and I_d is the TopDrop intersection, $I_d = A^{(1:d)} \cap B^{(1:2d)}$, defined as the intersection between predictions from the two random halves "A" and "B" such that the top d genes in "A" do not fall past twice the designated depth (2d) in "B."

I expect that all methods will lose statistical power and have greater ran-

dom sampling error when they are predicting on a dataset that has been split in half. Therefore, I chose to allow genes to fall twice as far down the list in the "B" half of the split, to better distinguish random effects and methods with intrinsically low consistency. One notable exception is the method MuSiC. I suspect that because MuSiC prioritizes genes largely based on the total number of mutations within a gene, that it is generally stable when split in half.

A substantial limitation is TopDrop only evaluates consistency with independent identically distributed samples from the same data set. A better alternative would be to evaluate cross-study consistency [115], as there may be multiple reasons why findings in one study would not generalize to another. However, from a practical perspective, I can only use data currently available to evaluate performance as this represents a substantial fraction of the data as it existed at the time.

4.5 Overall Performance

In Table 4.1, I summarize the performance of each method according to the criteria described above on the pancancer mutation data. The overall protocol is shown as a flowchart in Figure 4.6. I assume that a preferable method would predict a higher fraction of driver genes that overlap with the CGC, that overlap with at least one other method, that have the least deviation from expected

null P values, and that have the highest consistency. Each method is accordingly ranked by these four criteria and the average rank is shown. The top ranked methods in order are 20/20+, TUSON, OncodriveFML, and MutsigCV.

4.6 Conclusion

A major goal of the huge public investment in large-scale cancer sequencing has been to find driver genes. Robust computational prediction of drivers from small numbers of somatic variants is critical to this mission, and it is essential that the best methods for this purpose be identified. Although many such methods have been proposed (see review [31]), it has been difficult to evaluate them because there is no gold standard to use as a benchmark. Here, I developed an evaluation framework for driver gene prediction methods that does not require a gold standard. The framework includes a large set of small somatic mutations from a wide range of cancer types and five evaluation metrics. It can be used to systematically evaluate new prediction methods and compare them to existing methods. The results would be more informative to users of these methods than current ad hoc approaches.

To apply the framework to a new method (Figure 4.6), a ranked list of predicted driver genes can be generated from the pancancer mutation dataset (see Appendix B), including a P value and a Benjamini-Hochberg corrected q value

Method	No. significant genes	CGC rank	Consensus rank	Pvalue rank	Consistency rank	Average rank
2020+	208	1	2	1	2	1.50
TUSON	243	2	1	2	3	2.00
OncodriveFML	679	4	4	3	4	3.75
MutsigCV	158	3	3	4	6	4.00
OncodriveClust	586	5	5	6	7	5.75
MuSiC	1,975	7	8	8	1	6.00
ActiveDriver	417	6	7	5	8	6.50
OncodriveFM	2,600	8	6	7	5	6.50

Table 4.1: Performance of eight evaluated cancer driver gene prediction methods on the pancancer dataset of small somatic mutations

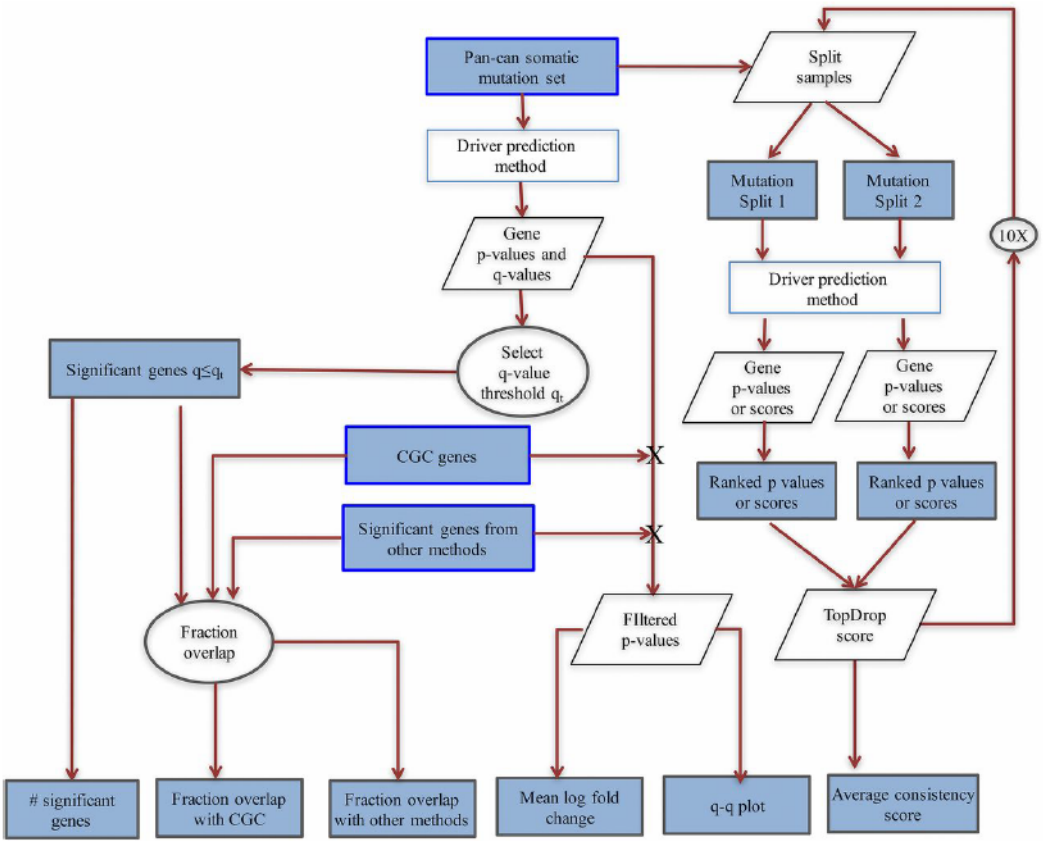

Figure 4.6: Flowchart of evaluation protocol. Overview of how a driver gene prediction method of interest can be evaluated. The input to the method is the pancan somatic mutation set provided in this work [85]. The initial output from the method to be evaluated is a list of predicted driver genes with associated P values and q values. A list of significant driver genes is produced by selecting a q value threshold. To compute fraction overlap of genes predicted as significant with Cancer Gene Census (CGC) and with the eight methods evaluated here, a freeze of CGC and predictions from the eight methods are provided. These gene lists are also used to subtract out putative driver genes and yield a list of filtered P values. Method consistency is estimated by 10 iterations of splitting the pancan somatic mutation set, outputting gene P values and scores for both halves, and applying the TopDrop metric. Jupyter notebooks for computing MLFC and qq plots from the filtered P value list, and the average TDC score are available at github.

for each gene. The choice of a threshold $q \leq 0.1$ to define driver genes worked well in our evaluations but can be adjusted if so desired. The same threshold should be used for fair comparison of different methods. If a driver prediction tool does not produce P values, a raw score threshold that represents the desired false-discovery rate could be selected.

The MLFC also has substantial implications for the accuracy of driver gene prediction methods. The relatively high MLFC of several methods brings into question the validity of the assumptions or analytic methods used in their construction. I believe that the most likely problem is with the assumptions rather than the analytic methods, which all appear to be well thought-out. In addition, the most likely problem with the assumptions is that there is unexplained variability in the background mutation rates (see chapter 2). This variability may be tumor type specific or even patient or tumor specific. If P values are underestimated in the range of low p-values, too many genes will be called as drivers. In fact, the methods that underestimate P values predict the largest number of drivers and have the highest fraction of uniquely predicted drivers.

Chapter 5

HotMAPS: Exome-scale discovery of mutation hotspots in 3D protein structure

5.1 Introduction

Missense mutations are perhaps the most difficult mutation type to interpret in human cancers. Truncating (loss-of-function) mutations and structural rearrangements generate major changes in the protein product of a gene, but a single missense mutation yields only a small change in protein chemistry. The impact of a missense mutation on protein function, cellular behavior, cancer etiology, and progression may be negligible or profound, for reasons that are

not yet well understood.

The recurrent observation across multiple cancer samples of missense mutations at the same amino acid residue position is well known to be a characteristic feature of both oncogenes (OG) and tumor suppressor genes (TSG; [116]). The idea that somatic mutations also frequently occur in positions proximal in protein sequence to the most highly recurrent positions has suggested that positional clustering of somatic missense mutations might be used to identify drivers [117]. These clusters, known as "hotspots," are regions where somatic missense mutations occur closer together in protein sequence than would be expected by chance. Hotspot regions can be rationalized as areas in a protein under positive selection in the cancer environment; missense mutations occurring in these regions are selected for because they alter protein function in a manner advantageous to the cancer cell. Numerous methods have been developed to identify hotspots based on the linear protein sequence [38,42,49–51,55].

However, only using the linear sequence of a protein may fail to capture hotspots that appear in the 3-dimensional (3D) structure of a protein [58]. Protein structure has long been known to relate to the function of a protein [56,57], and clustering of mutations within a structure may indicate mutations that are cancer drivers. A few algorithms leveraging this protein structure-function relationship have shown early signs of promise. An algorithm that leverages 3D protein structure information, but still performs clustering in 1D through a

dimensionality reduction step, has shown utility in detecting oncogenes [54]. A recent study of an aggregated collection of TCGA cancer mutations from 21 tumor types presented an algorithm to identify cancer genes based on 3D clustering of somatic missense mutations, yielding ten such genes [58].

Here, I present HotMAPS (Hotspot Missense mutation Areas in Protein Structure), a new, sensitive algorithm for high-throughput analysis of cancer 3D hotspot regions of missense mutation. HotMAPS finds clusters of amino acid residues with significantly increased local mutation density in 3D protein space, compared to an empirical null distribution. The statistical model is designed to handle higher-order protein complexes and can capture regions that span protein-protein interfaces. I apply HotMAPS to missense mutations from 23 tumor types sequenced by TCGA. By careful use of both experimentally derived protein biologic assemblies in the Protein Data Bank (PDB) and theoretical protein structure models, I substantially increase the number of amino acids that can be mapped into 3D protein space and the number of detectable hotspot regions compared to a prior approach [58].

5.2 HotMAPS algorithm

Standard clustering algorithms are not well suited for detecting rare clustering patterns in a large number of problems. I considered many standard

clustering algorithms for clustering mutations in protein structure [118–121], but each has substantial weaknesses. Methods like K-means and spectral clustering require the number of clusters to be specified as a parameter, but the number of clusters is not necessarily the same for every protein structure. In practice, the "elbow-method" is used to manually select the number of clusters by examining where there is a noticeable flattening in performance as the number of clusters is increased [122]. However, when applying clustering to ~65,000 protein structures, manual procedures are infeasible. Even if the clustering algorithm does not require the number of clusters as a parameter, such as the algorithms affinity propagation [119] or DBSCAN [118], they generally assume the minimum number of clusters is one. Since driver mutations are rare relative to all mutations, most protein structures should have no clusters due to the clustering of driver mutations.

There are also application specific concerns for clustering mutations in protein structure. A clustering algorithm would preferably adjust for the topology of a protein structure as this affects where mutations could possibly be located. Additionally, since protein structure may contain multiple identical protein subunits, a clustering algorithm needs to compensate for the non-independence of mutations. Lastly, the algorithm should ideally adjust statistically for potential false discovery of clusters due to the large number of clustering problems (~65,000 protein structures).

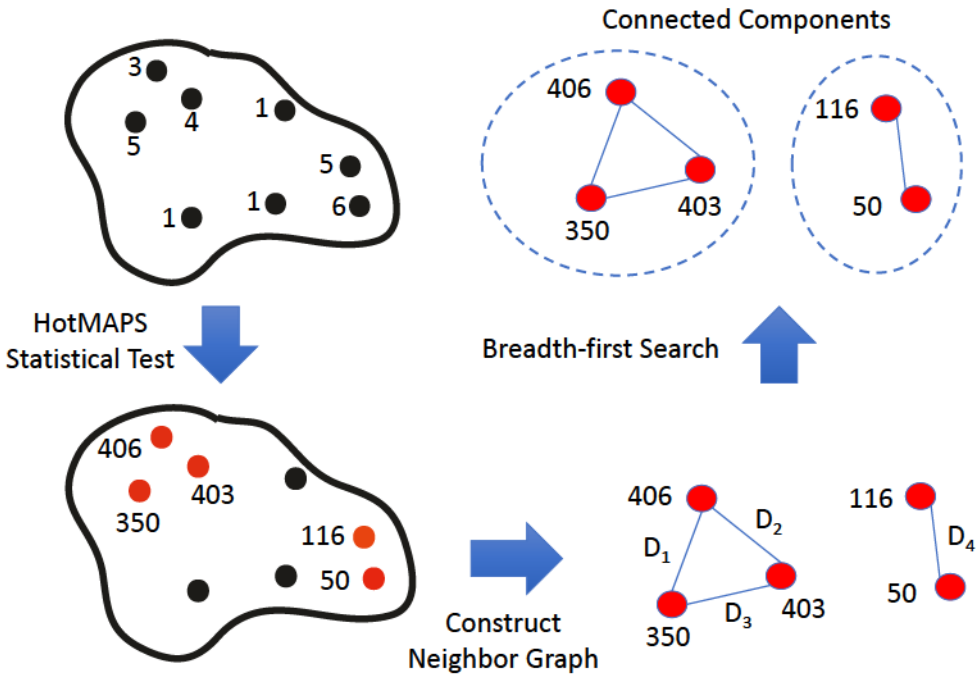

Figure 5.1: HotMAPS was run on 65,372 protein structures and models. For each structure or model, mutations were mapped from TCGA genomic coordinates to 3D protein space and for each mutated residue, its observed local mutation density was calculated. P-values were estimated based on simulations. If p-values for the same residue differed across multiple structures/models, the minimum was used and adjusted for multiple hypotheses testing with the Benjamini-Hochberg algorithm. Hotspot regions were identified as connected components in a graph of significantly mutated residues.

The HotMAPS algorithm was developed to address all of these noted limitations and is described below (Figure 5.1).

5.2.1 Mutation density

HotMAPS depends on calculating a mutation density for each amino acid residue. Let K be the set of all protein structures. Each protein structure or

model was an element $k \in K$. For each k, the center of geometry in Euclidean space (i.e., centroid) was calculated for each residue (r), considering all backbone and side-chain atoms,

$$C_r^k = \frac{1}{|r|} \sum_{a \in r} a \tag{5.1}$$

where C_r^k is the center of geometry for residue r in k, and a is a 3D position vector for each atom in residue r. The neighbors of residue r were identified using a 10 angstrom radius cutoff from the center of geometry,

$$N_r^k = \{r' : dist(C_r^k, C_{r'}^k) \leq 10, r' \in R^k\} \tag{5.2}$$

where R^k is the set of residues for k, N_r^k is the set of neighbor residues for residue r, and $dist$ is the Euclidean distance function. The density D of mutations at residue r was calculated as the sum of mutations in the residue's neighborhood,

$$D_r^k = \sum_{n \in N_r^k} M_n^k \tag{5.3}$$

$$D_{obs}^k = \{D_r^k : M_r^k > 0, r \in R^k\} \tag{5.4}$$

where M_n^k is the number of missense mutations for the n'th residue neighbor, D_r^k is the density of mutations for a specific residue and k, and D_{obs}^k is the

set of observed mutation densities for all mutated residues in a given k.

5.2.2 Statistical model

Next, I simulated the expected null hypothesis if mutations on the protein structure were under no selective pressure to occur in any particular region. The null distribution is reasonably modeled by a discrete uniform distribution. Mutations occurring under the null were simulated by sampling with replacement a number of residues equal to the total observed mutations,

$$M_{sim}^{k} \sim Uniform(R^{k}, Size = \sum M_{r}^{k}) \tag{5.5}$$

where M_{sim}^{k} is the simulated missense counts for all residues in k. The procedure was modified slightly for protein complexes, which contain multiple protein chains that originate from a single gene product (e.g., a homodimer). I accounted for this non-independence by running identical simulations simultaneously on multiple duplicated protein chains. Duplicate chains were identified based on either having same PDB chain letter and/or the same chain description. The mutation density for simulated mutations was calculated in the same manner as the observed mutations. The procedure was repeated for 10,000 iterations on each structure.

Based on the empirical null distribution established from simulations, I cal-

culated the one-tailed p-value for each residue's mutation density being equal or larger,

$$P_r^k = \frac{\#\{d : d \geq D_r^k, d \in D_{sim}^k\}}{\#\{d \in D_{sim}^k\}} \tag{5.6}$$

where D_{sim}^k is the set of all simulated mutation densities and P_r^k is the p-value for residue r in k. Since there may be many structures and/or models that cover the same corresponding portion of the genome, multiple p-values were collapsed by taking the minimum p-value among residues that mapped to the same genomic codon. These unique genomic-level p-values were then corrected for multiple hypotheses by the Benjamini-Hochberg method [109] and deemed significant at a q-value of 0.01. I selected the very conservative q=0.01 empirically, to minimize the number of false discoveries in our study. Identifying the corresponding significant residues at the structure (or model) level was backtracked by using the supremum of significant p-values at the codon level as a cutoff,

$$P^* = \sup\{P_c : q_c < 0.01, \forall c\} \tag{5.7}$$

$$R_{signif}^k = \{r : P_r^k \leq P^*, \forall r\} \tag{5.8}$$

where P_c and q_c are the genomic p-value and q-value, respectively, for codon c, P^* is the p-value cutoff adjusted for multiple hypotheses, R_{signif}^k is the set of

significant residues for k.

5.2.3 Constructing hotspot regions

3D mutation hotspot regions were identified as groupings of significant residues, according to the principle of maximum parsimony. Specifically, I found the minimum number of non-contiguous hotspot regions that explained all significant residues. I first constructed a neighbor graph amongst significant residue positions, where edges were created if two residues could be considered as neighbors, defined as within 10 angstroms (1nm), which is the order of magnitude for the length of an amino acid residue side chain. 3D mutation hotspot regions for each k were then found as the connected components of the neighbor graph using breadth-first search. Our results were not very sensitive to small perturbations of this parameter (8Å, 9Å, 11Å, 12Å). The 10Åmaximum distance identified 85% of the hotspot residues identified at the four other threshold values.

5.3 Mapping mutations to protein structure

5.3.1 Mutational data set

Mutation annotation format (MAF) file data for 23 tumor types from The Cancer Genome Atlas (TCGA) was downloaded by the Xena data store (`https://genome-cancer.soe.ucsc.edu/proj/site/xena/hub/`) using their API.

5.3.2 Protein structure

PDB structures were obtained from the Worldwide Protein Data Bank (PDB) (10/17/2015). Only structures solved by x-ray crystallography and containing at least one human protein chain were used. To avoid computation on crystal-packing artifacts that are common in PDB multi-domain protein structures and proteins in complex with other proteins or DNA/RNA structures, I used PDB biological assemblies that model how proteins exist in vivo (`ftp://ftp.wwpdb.org/pub/pdb/data/biounit`). Additionally, single-domain, theoretical protein structure models constructed based on homology to non-human proteins were included to increase coverage over a greater proportion of genes. Theoretical models were obtained from the ModPipe human 2013 dataset (`ftp://salilab.org/databases/modbase/projects/genomes/H_sapiens/2013/`),

built with Modeller 9.11 [123]. In addition to criteria required by ModPipe (ModPipe Protein Quality Score > 1.1), theoretical models were further filtered to increase the quality of structures used in our assessment, requiring that: 1) models had a minimum length of 75 residues. 2) The sequence of the target human protein and the sequence of the non-human homolog used for homology modeling were $\geq$10% identical. 3) The "loop" content of the protein model was $\leq$30%. 4) Compactness score C (see Equation 5.9) was $\leq$1Å/residue. The compactness score was based on the protein radius of gyration (R_g), and was employed to reject overly extended or unfolded structures. All thresholds were selected by visual inspection of structures meeting each of the four criteria.

Theoretical protein structure compactness score filter:

$$C = \frac{4R_g}{N} \tag{5.9}$$

$$\text{where } R_g = \sqrt{\frac{\sum_i m_i(\vec{r}_i - \vec{r}_c)^2}{\sum_i m_i}} \tag{5.10}$$

where N is total number of residues. m_i is the mass of the i'th atom, $\vec{r}_i$ is the center of the i'th atom, and $\vec{r}_c$ is the protein center of geometry.

5.3.3 Mapping algorithm

Mapping of genome coordinates was done using a modified version of the TransMap algorithm (Figure 5.2), previously described in [124]. In a minority of cases mutations did not have a one-to-one mapping within a protein structure (0.6% of mutations analyzed in this study were impacted). Any hotspot region residue positions with ambiguous mappings were dropped from the final analysis. Protein sequences in the UniProt database (SwissProt curated only) [125] were aligned to all transcripts in RefSeq, CCDS and Ensembl databases with tBLASTn [126]. Transcripts were then aligned to human genome assembly GRCh37 (hg19) with BLAT [127]. BLAT was also used to align the UniProt protein sequences with PDB SEQRES amino acid residue sequences (Figure 5.2). For theoretical models, ModPipe provided a RefSeq or Ensembl transcript identifier and translation of each transcript into protein sequence, eliminating the need for the tBLASTn step to align protein sequence to transcript.

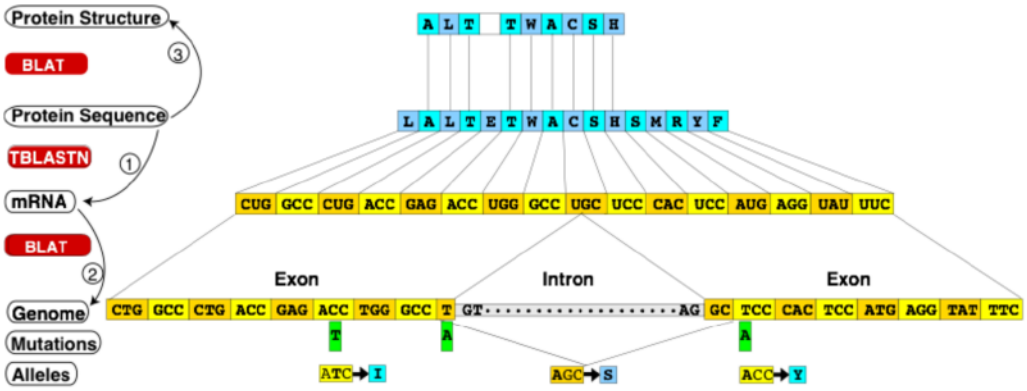

Figure 5.2: The mapping is done with three pairwise alignment steps, using tBLASTn and BLAT. Projection of protein sequence coordinates to mRNA transcript coordinates (1) and finally genomic coordinates (2) is done "top down". The process enables handling of split codons, such as the "AGC" shown. Protein sequence coordinates are subsequently projected into the PDB coordinate system of protein structure (3).

5.4 3D mutation hotspot regions are important in cancer

5.4.1 3D hotspot regions are enriched in well-known cancer genes

Among the set of genes with available protein structure or models (n = 15,697), the genes harboring a 3D hotspot region are enriched for OGs and TSGs (P = 6.1E30 for OGs and P = 2.4E13 for TSGs; one-tailed Fisher exact test). They are also enriched for genes in the CGC list (P = 1.4E30; one-tailed Fisher exact test). The subset of these genes harboring only a 3D hotspot re-

gion not detectable in 1D is also significantly enriched (P = 4.3E09 for OGs, P = 7.9E12 for TSGs, P = 8.0E11 for CGC genes; one-tailed Fisher exact test). An additional 23 genes that are proposed OGs, TSGs, and/or drug targets or hereditary cancer genes contained at least one 3D hotspot region. This enrichment of known and candidate driver genes supports my claim that many of the regions are biologically relevant and not simply artifacts. While regions were detected in only approximately 18% of established cancer genes, I expect that many of these genes harbor drivers other than missense mutations, some are drivers in tumor types not represented in our study and many lack structural coverage.

5.4.2 Mutations in 3D hotspot regions are different from other somatic mutations in cancers

I examined whether the amino acid residue positions and the missense mutations in the 3D hotspot regions had distinctive features suggestive of a special biologic importance, when compared with the remaining mutations in our study. Four candidate distinguishing features were tested: (i) vertebrate evolutionary conservation; (ii) occurrence at a protein-protein interface, which increases the potential for a missense mutation to disrupt protein-protein in-

teractions; (iii) *in silico* cancer driver scores generated with the CHASM algorithm [39]; and (iv) in silico pathogenicity scores generated with the VEST algorithm [106], which are predictors of increased missense mutation impact (Figure 5.3). In comparison with mutated residues not in 3D hotspot regions, vertebrate evolutionary conservation was higher and protein-protein interface occurrence was higher in the 3D hotspot regions (conservation P = 2.9E29, Mann-Whitney U test; protein interface P = 5.2E13, one-tailed Fisher exact test). In silico driver scores and pathogenicity scores were higher for missense mutations in 3D hotspot regions (driver score P = 3.0E47, pathogenicity score P = 3.0E16; Mann-Whitney U-test) than for the remaining mutations (Figure 5.3).

5.4.3 3D hotspot regions are different in oncogenes and tumor suppressor genes

The catalog contains 37 regions stratified by tumor type in bonafide tumor suppressor genes and 77 in bonafide oncogenes (114 regions in 30 genes), using as a benchmark the classifications of Vogelstein and colleagues (landscapes benchmark; [14]). I used these data to explore possible differences between tumor suppressor gene and oncogene regions at amino acid resolution. I found that in tumor suppressor genes, 3D hotspot regions were larger than in onco-

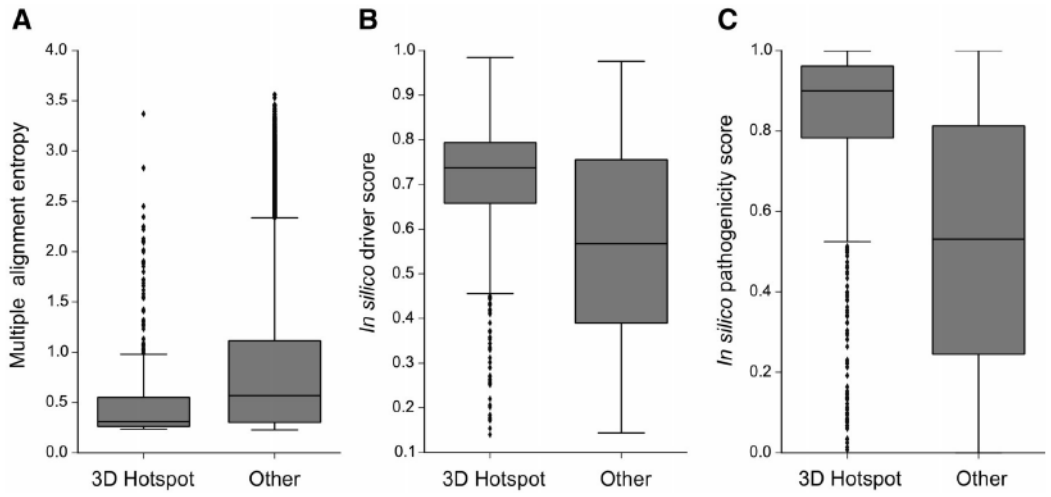

Figure 5.3: Three distinguishing features of HotMAPS regions. A, HotMAPS-mutated residues are more conserved in vertebrate evolution than mutated residues not in hotspot regions, as shown by lower multiple alignment entropy (P = 1.2E29; Mann-Whitney U test). Multiple alignment entropy is calculated as the Shannon entropy of protein-translated 46-way vertebrate genome alignments from UCSC Genome Browser, which is lowest for the most conserved residues. B and C, HotMAPS missense mutations have higher in silico cancer driver scores from the CHASM algorithm (P = 5.3E47; Mann-Whitney U test) than those mutations not in hotspot regions (B) and higher in silico pathogenicity scores from the VEST algorithm (P = 7.0E162; Mann-Whitney U test; C). Finally, HotMAPS-mutated residues occur more frequently at protein-protein interfaces (P = 1.3E11; one-tailed Fisher exact test).

genes (region size P = 9.6E06; Mann-Whitney U test). They were also more mutationally diverse (mutational diversity P = 2.1E07; Mann-Whitney U test). In addition, oncogene 3D hotspot regions were more conserved in vertebrate evolution and more solvent accessible in protein structure, meaning that they tend to occur at the protein surface (evolution P = 4.7E07, solvent accessible P = 1.5E06; Mann-Whitney U test). Hotspot regions in tumor suppressor genes harbored increased net change in hydrophobicity (P = 3.3E07; Mann-Whitney U test) and net change in volume (P = 2.2E07; Mann-Whitney U test), suggesting that their impact on protein function could be due to decreased stability. The in silico missense mutation cancer driver scores were higher for oncogene regions (P = 0.003; Mann-Whitney U test). I also tested differences between in silico pathogenicity scores and occurrence at protein-protein interfaces between OG and TSG regions, but these were not significant (pathogenicity scores P = 0.37, protein interface P = 0.34; Mann-Whitney U test).

The fact that these differences between oncogene and tumor suppressor gene regions were statistically significant suggested that they might have predictive value. Principal components analysis (PCA) of the six significant features indicated some separation (Figure 5.4A). Next, I trained a Naive Bayes machine learning classifier to discriminate between oncogene and tumor suppressor gene hotspot regions, using region size, mutational diversity, vertebrate conservation, residue solvent accessibility, mutation net hydrophobicity

change, and residue volume change as features. A rigorous gene-level holdout protocol was used to avoid overfitting. A Naive Bayes score closer to 1.0 indicates that the hotspot region is likely in an OG while a score closer to 0.0 indicates that it is in a TSG. Area under receiver operating characteristic (ROC) curve or AUC, a standard measure of classifier performance, was 0.84 out of 1.0, a result that supports my claim that 3D hotspot regions in oncogenes and tumor suppressor genes have distinctive characteristics (Figure 5.4B). AUC of a classifier with random performance is 0.5. Performance did not improve when the other features were included in the classifier. The ROC performance and PCA plot support my claim that characteristic differences between oncogenes and tumor suppressor genes hotspots can be quantified.

5.4.4 What is gained by 3D hotspot region detection versus 1D?

The larger size and mutational diversity of hotspot regions in tumor suppressor genes (TSGs) versus oncogenes (OGs) suggest that they could be more difficult to detect and perhaps they have been underreported by 1D approaches. OG hotspot regions consisting of recurrent missense mutations at one or two residues can be seen by eye with lollipop plots and are straightforward to detect computationally based on 1D primary sequence. I hypothesized that detection

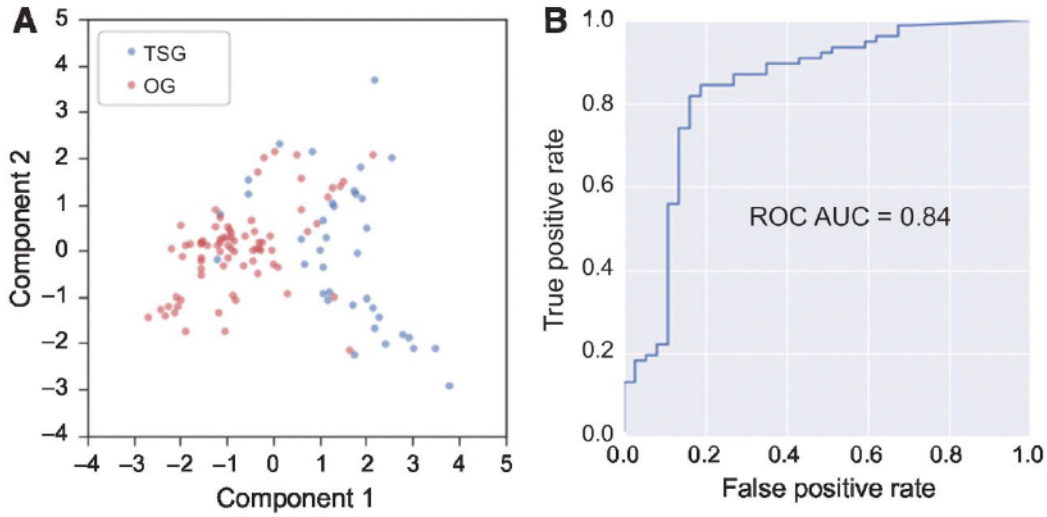

Figure 5.4: A, PCA plot shows a clustering pattern in hotspot regions identified in oncogenes (OG=red) and tumor suppressor genes (TSG=blue). Each point is a region represented by six numeric features, projected into two dimensions. The features are region size, mutational diversity, vertebrate evolutionary conservation, residue relative solvent accessibility, mutation net change in hydrophobicity, and mutation net change in residue volume. B, OG and TSG HotMAPS regions can be discriminated with machine learning, based on six features. A Gaussian Naive Bayes classifier trained with the landscapes benchmark provides a reasonable separation between the two classes with AUC = 0.84 out of 1.0. Performance of a random classifier is AUC = 0.5. ROC, receiver operating characteristic; AUC, area under the ROC curve.

of many TSG hotspot regions might require a 3D algorithm. To maximize the interpretability of this analysis, regions that occurred in multiple tumor types were merged so that each region was represented only once in each gene.

For a well-controlled comparison of 3D and 1D hotspot region detection, I applied a 1D version of our method to the protein chain sequences of the same set of PDB protein bioassemblies and theoretical protein structure models to detect nonuniform clustering patterns on primary protein sequence. Seventy-two percent of hotspot regions identified in 3D were identifiable in 1D.

Next, I compared the number of hotspot regions identified in OGs and TSGs. I considered regions identified in 3D only, in both 3D and 1D, and in 1D only. Using the bona fide OGs and TSGs (Table 5.1), there were significantly more OG regions that TSG regions identified by the 1D algorithm ($P = 0.03$; one-sided Fisher exact test). The 1D-only version of the algorithm detected 5 OG and 2 TSG regions; 1D further detected an additional 25 OG and 7 TSG regions that were also identified by the 3D algorithm. The 3D algorithm identified an additional 4 OG and 6 TSG regions. To increase our power, I repeated this test again using the bona fide OGs and TSGs plus additional regions in five candidate OGs and TSGs reported in the literature (OGs were FSIP2, MTOR, RANBP2, CHEK2, and MAPK1; TSGs were RASA1, SMARCA2, KEAP1, CUL1, TGFBR2; all are listed and cited in Table 5.2), yielding increased statistical significance ($P = 0.009$, one-sided Fisher exact test). The results suggest that 1D

detection methods may be better suited to detecting regions in OGs rather than TSGs.

A further problem with sequence-based 1D hotspot region detection is that larger regions detectable in 3D may be only partially characterized and/or split into multiple pieces. Figure 5.5 shows an example of a TSG hotspot region in FBXW7 found in 3D by HotMAPS that has been split into two pieces by the 1D algorithm. In 1D protein sequence, residue 465 is not close enough to residues 502 and 505 to be identified in one hotspot region. On the 3D protein structure of FBXW7 (PDB code 2OVQ), the three residues are spatially close and a single hotspot region is detected.

5.5 3D hotspot regions may increase interpretability of driver mechanisms

Three-dimensional consideration of hotspot regions in protein structure can potentially provide researchers with a rich source of hypothesis generation about driver mechanisms. While gene- or domain-level mutation enrichment analysis can point to potential protein functions, interactions, biologic processes, and pathways important for cancer etiology and progression, more detailed information may be available once a specific set of mutated amino acid residues has been identified as significant.

Gene	Landscapes Annotation	Cancer Gene Census (CGC)	TCGA Tumor Type(s)
FGFR3	OG	Dom	BLCA
SF3B1	OG	Dom	BRCA, BLCA
FGFR2	OG	Dom	BRCA, UCEC
KRAS	OG	Dom	CESC, UCS, PAAD, STAD, BLCA, UCEC, LUAD, BRCA
PIK3CA	OG	Dom	ESCA, CESC, UCS, LUSC, GBM, STAD, LGG*, BLCA, UCEC, PRAD, LUAD, KIRC, BRCA, HNSC
NFE2L2	OG	Dom	ESCA, HNSC, BLCA, UCEC, LUSC
IDH1	OG	Dom	GBM, LGG, SKCM
IDH2	OG	Dom	LGG
PTPN11	OG	Dom	LGG
MAP2K1	OG	Dom	LUAD*, SKCM
GNAS	OG	Dom	PAAD
BRAF	OG	Dom	THCA, GBM, LUAD, SKCM, PRAD*
HRAS	OG	Dom	THCA, PCPG, BLCA, HNSC, LUSC*
NRAS	OG	Dom	THCA, SKCM
PPP2R1A	OG	Dom?	UCS, UCEC
SPOP	OG	Rec	PRAD
ERBB2	OG		ESCA*, BRCA, BLCA
EGFR	OG		GBM, LGG, LUAD
RET	OG		PCPG
PIK3R1	TSG	Rec	BRCA*, GBM, UCEC*, LGG*
FBXW7	TSG	Rec	CESC*, UCS, LUSC*, STAD, BLCA, UCEC, HNSC
TP53	TSG	Rec	ESCA, UCS, PAAD, LUSC, GBM, STAD, LGG, BLCA, UCEC, PRAD, LUAD, OV, BRCA, HNSC
CIC	TSG	Rec	LGG
SMARCA4	TSG	Rec	LGG*
BCOR	TSG	Rec	UCEC
PTEN	TSG		BRCA, GBM*, UCEC
CDKN2A	TSG		ESCA*
VHL	TSG		KIRC*
NOTCH1	TSG		LGG*
SMAD4	TSG		STAD*
RHOA		Dom	BLCA*, HNSC, STAD
RAC1		Dom	HNSC, SKCM
ERBB3		Dom	STAD

Table 5.1: Cancer genes with 3D HotMAPS regions identified in TCGA tumor types and in landscapes benchmark or cancer gene census

Gene	TCGA Tumor Type(s)	Gene Details
AP2B1	ESCA	Involved in FGFR signaling. Knockdown promotes the formation of matrix degrading invadopodia, adhesion structures linked to invasive migration in cancer cells (Pignatelli 2012).
CAND1	BLCA*	Component of many protein complexes involved in proteasome-dependent protein degration via ubiquitination and neddylation. CAND1 binding to the complexes inactivates ubiquitin ligase activity and may block adaptor and NEDD8 conjugation sites. (Bosu 2008). May play a role in PLK4-mediated centriole overduplication and Disrupted in prostate cancer (Korzeniewski 2012).
CHEK2	ESCA, LGG, BLCA, HNSC, PRAD, LUAD, PCPG, KIRC	Checkpoint kinase involved in DNA damage response signaling. Significantly mutated gene and candidate **OG** in papillary thyroid carcinoma (PTC) cohort of 296 patients (TCGA 2014 #85). **Breast cancer susceptibility gene** (inherited germline variants) (Vogelstein 2013)
CUL1	BLCA	Candidate **TSG**. SCF complex E3 ubiquitin ligase scaffold protein. Suppressor of centriole multiplication through regulation of PLK4 level (Korzeniewski 2009)
ERCC2	BLCA, LGG*	DNA-repair (Nucleotide excision repair) protein. Significantly mutated in cisplatin-responders vs. non-responders in cohort of 50 patients with muscle-invasive urothelial carcinoma (MIUC). ERCC2 mutation status may inform cisplatin-containing regimen usage in MIUC (Van Allen 2014). Recurrently mutated in cohort of 17 patients with urothelial bladder cancer (UBC) (Balbas-Martinez 2013). **Xeroderma pigmentosum susceptibility gene** (inherited germline variants) (Vogelstein 2013)
FSIP2	ESCA*	Candidate **OG**. Recurrently amplified in testicular germ cell tumors (TGCTs)(Litchfield 2015).
GNA13	BLCA	Significantly mutated in cohort of 55 patients with diffuse large B-cell lymphoma (DLBCL) (Lohr 2012)
GTF2I	UCEC	Highly recurrent missense mutation in Thymic epithelial tumors and associated with increased patient survival (Petrini 2014).
HDAC4	ESCA	Histone de-aceytlation enzyme. **Drug target**. Overexpression shown to promote growth of colon cancer cells via p21 repression. Regulator of colon cell profliferation. (Wilson 2008). May regulate cancer cell response to hypoxia via its regulates HIF1a acetylation and stability (Geng 2011)
HLA-A	BLCA, HNSC, LGG, PRAD	Immune system. Encodes MHC-Class 1A protein, which presents antigens for T cell recognition. Somatic mutations previously suggested to contribute to tumor immune escape (Shukla 2015).
KEAP1	LUAD*	Candidate **TSG**. Inhibits NRF2 (aka NFE2L2). In cohort of 76 non-small cell lung cancer (NSCLC) patients, KEAP1 found mutated in 2 patients with advanced adenocarcinoma and smoking history. KEAP1 mutation was mutually exclusive of EGFR, Kas, ERBB2 and NFE2L2 mutation in the cohort and KEAP1 mutation status proposed as marker for personalized therapy selection. (Sasaki 2013) Proposed TSG in lung squamous cell carcinomas (Hast 2014) Proposed as therapeutic target for thyroid-transcription-factor-1 (TTF1)-negative lung adenocarcinoma (LUAD) (Cardnell 2015).
KLF5	BLCA*	Transcription factor that promotes breast cancer cell proliferation, survival, migration and tumour growth. Upregulates TNFAIP2, which interacts with the two small GTPases Rac1 and Cdc42, thereby increasing their activities to change actin cytoskeleton and cell morphology (Jia 2015). Proposed as playing dual role as both **TSG** when acetylated and **OG** when de-acetylated in prostate cancer (Atala 2015). Recurrently mutated in mucinous ovarian carcinoma (Ryland 2015)
MAPK1	CESC*, HNSC	Kinase involved in cell proliferation, differentiation, transcription regulation, and development; key signaling component of the toll-like receptor pathway. Candidate **OG** in pancreatic cancer (Furukawa 2006), laryngeal squamous cell carcinoma cell lines (Kostrzewska-Poczekaj 2010). Significantly mutated in cohort of 91 chronic lymphocytic leukemia CLL patients.(Wang 2011).
MSN	ESCA*	Protein homolog of TSG NF2 (Merlin) (Golovnina 2005). Member of the Ezrin-Radixin-Moesin (ERM) protein family. Links membrane and cytoskeleton involved in contact-dependent regulation of EGFR (Chiasson-MacKenzie 2015). Regulates the motility of oral cancer cells via MT1-MMP and E-cadherin/p120-catenin adhesion complex. Cytoplasmic expression of MSN correlates with nodal metastasis and poor prognosis of oral squamous cell carcinomas (OSCCs), may be potential candidate for targeted gene therapy for OSCCs (Li 2015).
MTOR	KIRC	Candidate **OG**. Serine/threonine protein kinase regulates cell growth, proliferatin and survival. Frequently activated in human cancer and a major therapeutic target. Randomly selected mutants in HEAT repeats and kinase domain induced transformation in NIH3T3 cells and rapid tumor growth in nude mice (Mueugan 2013)
NBPF10	BLCA*	Somatic missense mutation reported in prostate cancer cohort of 141 patients (Manson-Bahr 2015). In gene family with numerous tandem repeats and pseudogenes, possible read alignment and mutation calling errors.
PARG	GBM, LGG, BLCA, HNSC, PRAD, LUAD, PCPG, KIRC*	Involved in DNA damage repair (with PARP1). Cells deficient in these proteins are sensitive to lethal effects of ionizing radiation and alkylating agents (17). Potential **Drug target** for BRCA2-deficient cancers (Fathers 2012).
RANBP2	ESCA	Candidate **OG** (Gylfe 2013). A large multimodular and pleiotropic protein with SUMO E3 ligase function. (Zhu 2015) Interacts with mTOR (to regulate cell growth and proliferation via cellular anabolic processes) (Kazyken 2014). Hot spot mutation previously found in MSI colorectal cancer (CRC). Hot spot suggested as useful for personalized tumor profling and therapy in CRC. (Gylfe 2013)
RASA1	HNSC*	Identified as **TSG** in another squamous cell cancer, cutaneous squamous cell skin cancer (cSCC) (Pickering 2014)
RGPD3	BLCA*, UCEC, PAAD	Component of ubiqutin E3 ligase complex. Named for similarity to RANBP2.
SIRPB1	HNSC, PRAD	Ig-like cell-surface receptor. Negatively regulates RTK processes. Related to FGFR signaling.
SMARCA2	BLCA*	Actin-dependent regulator of chromatin. Its ATPase domain named as **Drug target** in SWI/SNF mutant cancers (e.g., lung, synovial sarcoma, leukemia, and rhabdoid tumors) (Vangamudi 2015). Proposed **TSG**, and synthetic lethal target in SMARCA4 (aka BRG1) -deficient cancers.(Hoffman 2014)
TGFBR2	HNSC	TSG in HNSC (Rothenberg,2012) MSI CRC (Biswas 2008), epithelial transformation and invasive squamous cell carcinoma in the mouse forestomach (Yang 2014).

Table 5.2: Genes with HotMAPS regions identified in TCGA tumor types

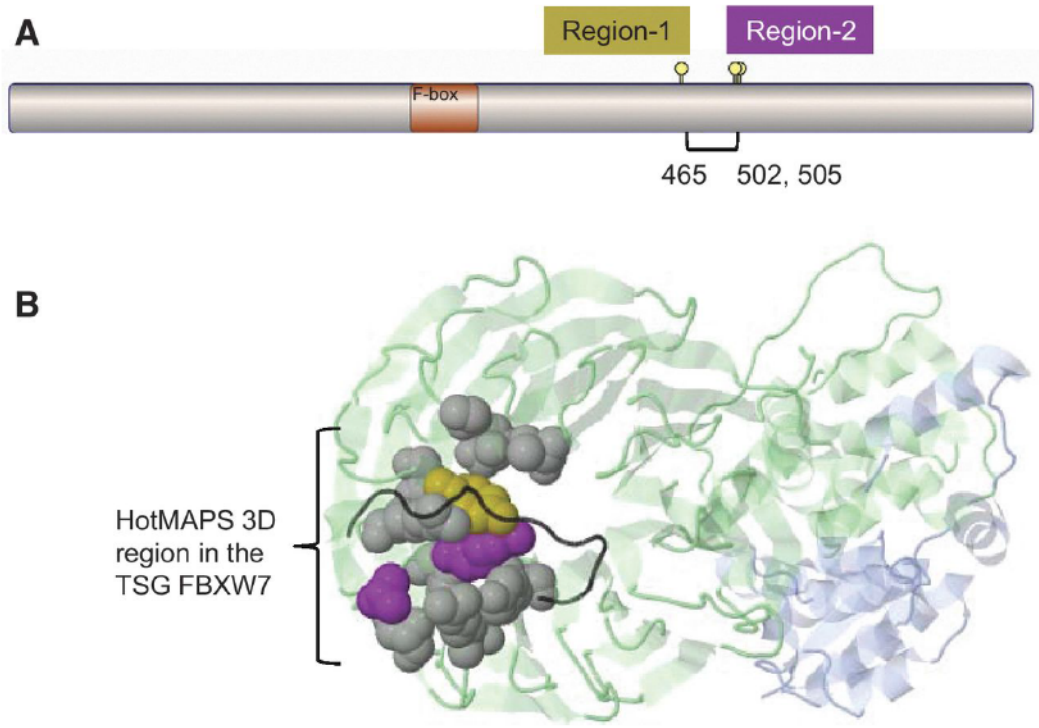

Figure 5.5: Comparison of hotspot detection in the TSG FBXW7 in 1D and 3D. A, a simplified 1D version of HotMAPS found two regions in FBXW7. The 3D version of HotMAPS found a single larger region, encompassing both regions. Diagram shows protein sequence of FBXW7, which contains a single F-box functional domain. Region-1, residue 465 (left lollipop); Region-2, residues 502 and 505 (right lollipops). B, HotMAPS identifies a single 3D hotspot region in FBXW7. Structure of SCFFbw7 ubiquitin ligase complex (PDB 2OVQ), containing FBXW7 (green), SKP1 (blue), and CCNE1 fragment (degron peptide; black). Residue coloring: 1D Region-1, gold; 1D Region-2, purple. Residues missed by 1D detection but included in HotMAPS 3D, gray. Although the 1D regions are far in the primary protein sequence, residues 505 and 465 spatially contact at the interface with CCNE1. Protein structure figures were generated by JSMol in MuPIT (`http://mupit.icm.jhu.edu/`).

For many of the 3D hotspot regions found by HotMAPS, the literature contains evidence that they are in direct contact with or proximal to amino acid residues of known functional importance. Figure 5.6 shows six cancer-associated proteins in which the hotspot region is either overlapping or proximal to important functional sites.

5.5.1 RAC1 hotspot in squamous head and neck cancer

RAC1 is a Rho GTPase important in signaling systems that regulate the organization of actin cytoskeleton and cell motility. The hotspot overlaps the GTP/GDP-binding site and could impact regulation of normal RAC1 cycling between GTP- and GDP-bound states (Figure 5.6A). It contains a previously identified recurrent mutation in melanoma (P29S), which dysregulates RAC1 by a fast cycling mechanism [128].

5.5.2 SPOP hotspot in prostate cancer (PRAD)

SPOP is the substrate recognition component of a cullin3-based E3 ubiquitin-protein ligase complex, which targets multiple substrates for proteasomal degradation. The hotspot overlaps with a binding groove harboring five residue positions (pink) where mutagenesis has strongly reduced affinity for the substrate

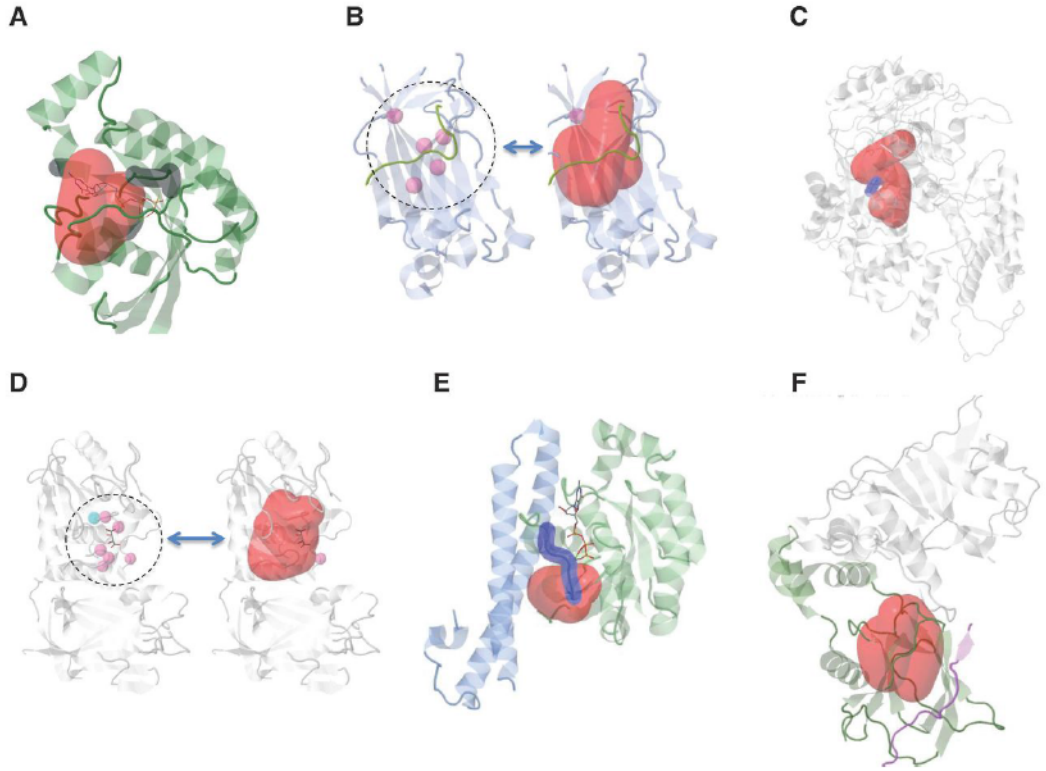

Figure 5.6: HotMAPS hotspot regions overlap and are proximal to important functional sites. A, HNSCC hotspot region (red) in RAC1 (green) and GTP/GDP-binding residues (dark gray; PDB 2FJU). B, PRAD hotspot region (red) in SPOP-substrate complex (PDB 3HGH) with SPOP (blue) and H2AFY substrate (green). Left, five residues (pink) that when mutated show strongly reduced affinity for substrate. C, BLCA hotspot region (red) in ERCC2 (gray) shown on theoretical model of ERCC2 helicase ATP-binding domain. The hotspot is proximal to the DEAH box (blue), a highly conserved motif containing residues that interact with Mg2+ and are critical for ATP-binding and helicase activity. D, UCEC hotspot region (red) in PTEN (PDB 1D5R) with active site phosphocysteine residue (blue), residues when mutated annotated to reduce phosphatase activity (pink). E, STAD hotspot region (red) in RHOA with a GTP analog bound (sticks; PDB 1CXZ). GTP-binding residues and effector region, dark blue. F, KIRC hotspot region (red) in VHL-TCEB1-TCEB2 complex, bound to HIF1A peptide (PDB 4AJY). Proximity to the interaction site of VHL (green) and HIF1A (blue) suggests possible decreased ubiquitination of HIF1A, resulting in increased protein expression of HIF1A. TCEB1 and TCEB2, gray.

(annotated in the UniProtKB) (Figure 5.6B).

5.5.3 ERCC2 hotspot in bladder cancer

ERCC2 is an ATP-dependent helicase that is part of the protein complex TFIIH involved in RNA polymerase II transcription and nucleotide excision repair (NER). I identified a hotspot region, proximal to the DEAH box, a highly conserved motif containing residues that interact with Mg2+ and are critical for ATP binding and helicase activity (Figure 5.6C). This proximity suggests that the hotspot mutations could disrupt ATPase activity and yield defective NER [129].

5.5.4 PTEN hotspot

PTEN is a phosphatase for both proteins and phosphoinositides, and it removes a phosphate from PIP3, critical for signaling to AKT. The hotspot region identified in endometrial cancer (UCEC) spans two functionally important loops in the protein (P and WPD loops) at the boundary of the active site pocket (Figure 5.6D). Residues in these loops are critical for catalysis (blue dot) and are important for the P-loop's conformation. Mutagenesis of residues in the WPD loop reduces phosphatase activity and increases colony formation in cell culture [130]. Pink dots show residues that impact phosphatase activity.

5.5.5 RHOA hotspots

RHOA is a small GTPase oncogene, and like RAC1 is a member of the Ras superfamily [131]. I identified hotspot regions in bladder cancer (BLCA), head and neck squamous cell cancer (HNSCC), and stomach adenocarcinoma (STAD). The hotspot regions overlap with the RHOA effector region, a highly conserved motif that is involved in Ras superfamily signaling with downstream effector proteins (Figure 5.6E). The regions are immediately proximal to a magnesium ion, which has been implicated in regulating the kinetics of Rho family GTPases [132].

5.5.6 VHL hotspot (KIRC)

VHL is a component of an E3 ubiquitin protein ligase complex, and it ubiquitinates the OG transcription factor HIF1A, targeting it for proteasomal degradation [133]. One impact of VHL loss of function with failure to ubiquitinate HIF1A is increased protein expression of HIF1A. The hotspot region is proximal to its interaction site with HIF1A and could potentially have an impact on this interaction (Figure 5.6F). The TCGA kidney cancer (KIRC) samples were stratified on the basis of their missense mutation status: VHL hotspot, non-hotspot, or no missense (WT). HIF1A protein expression was not significantly different between VHL non-hotspot and VHL WT groups (P = 0.5; Mann-

Whitney U test), but was significantly higher between VHL hotspot and VHL WT groups (P = 0.03; Mann-Whitney U test). This result is consistent with a special role for VHL hotspot missense mutations in regulating HIF1A protein expression. However, increased HIF1A expression in these KIRC samples is likely impacted by additional genetic and other factors. I might see a substantially lower P value if VHL hotspot mutations were the only cause of the observed increase. Also, there are many VHL missense mutations outside of the hotspot region, and it is likely that several of these also have a functional impact. In particular, several of them are at the interface of VHL and the TCEB1 and TCEB2 in the complex and could impact VHL/TCEB binding.

5.6 Conclusions

I systematically identified 3D missense hotspot regions using TCGA somatic mutation data from 6,594 samples in 23 tumor types. HotMAPS identified 107 unique regions and 216 cancer type-specific regions. This catalog enables assessment of how the specific missense mutations in a hotspot contribute to cancer-associated molecular mechanisms. Unlike many machine learning algorithms, the visualization of HotMAPS region with protein structure allows model interpretability by biologists with domain knowledge of a particular protein.

At the time of publication, several other algorithms were published which also supported the notion that mutational clustering in protein structure was advantageous [41, 45, 46]. In a comparison from [41], HotMAPS had performance equivalent to other top methods on discriminating likely driver missense mutations from an *in vivo* experiment. The HotMAPS algorithm does have similarities with the DBSCAN algorithm [118], which is also based on using density estimates for clustering. However, DBSCAN does not have a statistically principled criterion for controlling false discoveries.

Although recurrent missense mutations have long been known to occur in both oncogenes and tumor suppressor genes [116], they have been observed more frequently in oncogenes. Here I showed that there are systematic differences in hotspot regions found in oncogenes and tumor suppressor genes. Oncogene regions are smaller, less mutationally diverse, more evolutionarily conserved, and more solvent accessible than tumor suppressor gene regions. Tumor suppressor gene regions are more likely to harbor mutations that may impact protein stability through changes in hydrophobicity or volume. Potential explanations for these differences are that there are more ways to lose the function of a protein than to gain function [134]. Loss-of-function tumor suppressor mutations can occur at many residue positions and involve many types of amino acid residue substitutions, while oncogene mutations will occur at a few functionally important positions and involve fewer substitution types.

Chapter 6

CHASMplus: enhanced context reveals the scope of somatic missense drivers in human cancers

6.1 Introduction

In previous chapters, I have shown large-scale sequencing studies of patient cohorts have enabled identification of many genes or regions that when mutated *can* act as cancer drivers. However, not every mutation in a driver gene or region is necessarily a driver of cancer; thus, requiring methods to

discriminate whether an individual mutation is a driver or passenger.

The most common approach has been to apply machine learning to predict the cancer driver status of individual missense mutations by leveraging features characterizing a mutation, e.g., inter-species evolutionary conservation, features of the local protein environment, molecular function annotations, and biophysical characterizations of the amino acid substitution. Cancer-focused machine learning methods have previously tried to enhance performance by training cancer type specific models [39,63] or boosting data with synthetic passenger missense mutations [39]. Unfortunately, a recent systematic study comparing 15 such methods concluded that none of them were sufficiently reliable for experimental or clinical follow-through [135,136]. I and others have hypothesized that determining the impact of missense mutations requires proper context [67], which has not been sufficiently leveraged in a comprehensive manner in the current generation of methods. Context includes both prior knowledge about the functional importance of genes or gene subregions in which a mutation occurs, and mutational patterns that are now evident from cancer sequencing studies of many thousands of patients.

In this chapter, I present a new driver missense mutation prediction method, CHASMplus, that uses machine learning to integrate missense mutation context at multiple scales. The new CHASMplus consistently outperforms comparable methods, including the original CHASM, on eight different benchmark

sets – including in vitro experiments, in vivo experiments and literature benchmarks. Encouraged by these results, I applied CHASMplus to 8,657 The Cancer Genome Atlas (TCGA) samples from 32 cancer types to systematically identify driver missense mutations.

6.2 CHASMplus algorithm

6.2.1 Overview

CHASMplus uses the Random Forest algorithm to discriminate somatic missense mutations (referred to hereafter as missense mutations) that are drivers of human cancers from passenger missense mutations. A Random Forest is an ensemble of many randomized decision trees (see chapter 3) [102,103]. Each tree is trained on a random selection of training set examples and candidate features, via a recursive splitting process [104](Figure 6.1A). CHASMplus is trained using somatic mutation calls from The Cancer Genome Atlas (TCGA) covering 8,657 samples in 32 cancer types. Because there is no gold standard set of driver and passenger missense mutations, I developed a semi-supervised approach to assign class labels to missense mutations, taking advantage of Random Forest robustness to noisy class labels. Briefly, class labels are assigned so as to enrich the positive class for driver missense mutations

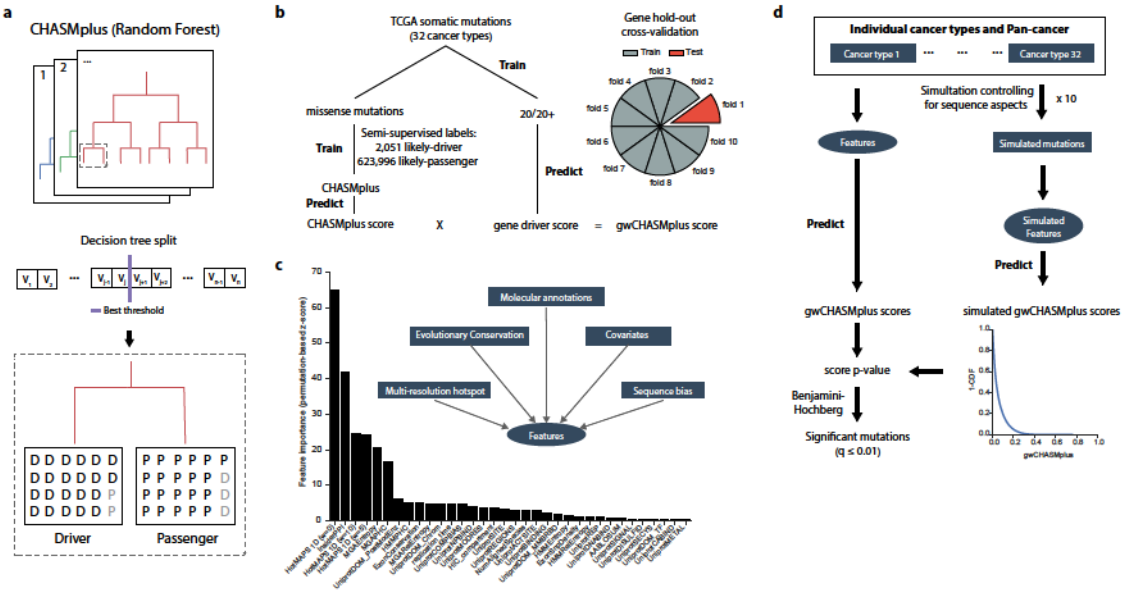

Figure 6.1: Overview of CHASMplus algorithm. a) CHASMplus predicts driver somatic missense mutations by using a random forest algorithm, consisting of an ensemble of decision trees. Each decision tree is constructed by selecting a random set of examples and features and recursively splitting examples by the best split criterion. b) Diagram of training and testing procedure by CHASMplus. c) Features with a net-positive feature importance by CHASMplus according to a permutation adjusted z-score. Boxed text indicates broad feature categories that were important. d) Diagram of how CHASMplus identifies statistically significant driver somatic missense mutations in each of the 32 cancer types individually and in aggregate (pan-cancer).

(Figure 6.1B). CHASMplus training is done with a rigorous gene holdout cross-validation protocol to avoid overfitting, by ensuring all mutations within a gene are within the same fold [132, 137]. Therefore, missense mutations are never scored by a Random Forest trained on any missense mutation harbored by the same gene. Finally, predicted scores from CHASMplus are weighted by the 20/20+ driver gene score, producing gene-weighted (gwCHASMplus) scores (Figure 6.1B).

6.2.2 Semi-supervised training labels

Using the TCGA mutation dataset, I established training labels with a semi-supervised approach, designed to minimize bias (Figure 6.2A). The positive class (likely-driver missense mutations) was selected by the following criteria: 1) missense mutations had to occur in a curated set of 125 pan-cancer driver genes [14]; 2) for each of the 32 TCGA cancer types, missense mutations found in that cancer type had to occur in a significantly mutated gene for that cancer type according to MutSigCV v1.4 [69]. I ran MutSigCV using recommended settings and a full sequencing coverage file (`http://archive.broadinstitute.org/cancer/cga/mutsig`). Importantly, MutSigCV v1.4 only assess the total number of mutations in a gene, and not any characteristics of those mutations; thus, I avoid making strong assumptions about the properties of a particular driver mutation; 3) missense mutations had to occur in samples with relatively low mutation rate (less than 500 mutations, half the minimum hypermutator threshold). This filter was intended to limit the number of passenger mutations mislabeled as drivers. The negative class (likely-passenger missense mutations) consisted of the remaining missense mutations in the TCGA mutation set. For training purposes, I only used unique mutations to avoid double counting a mutation seen more than once. If, however, the same mutation consequence observed in different cancer types had contradictory labels, I regarded the mutation as a driver because mutation recurrence is often

cited as supportive evidence for a cancer driver role. This established a set of 2,051 likely-driver missense mutations and 623,996 likely-passenger missense mutations, for which I found sufficient annotation to compute our selected features.

6.2.3 Features

CHASMplus scores benefit from representation of missense mutation context at multiple scales. The Random Forest was trained on 95 features, and the 34 features with a net positive feature importance are shown in Figure 6.1C. Important features assess five broad categories: multi-resolution missense mutation hotspots (HotMAPS 1D algorithm [43]), evolutionary conservation/human germline variation, molecular function annotations (e.g., protein-protein interface annotations from [138]), sequence biased regions, and gene-level covariates (e.g., replication timing). Missense mutation context is further represented by the 20/20+ driver score of the gene harboring the missense mutation and the specific cancer type in which it was observed. While gene-level features have been previously applied to missense mutation driver prediction [62], to my knowledge, this is the first time that gene-level and missense mutation-level driver scores have been coupled in a cancer type-specific manner.

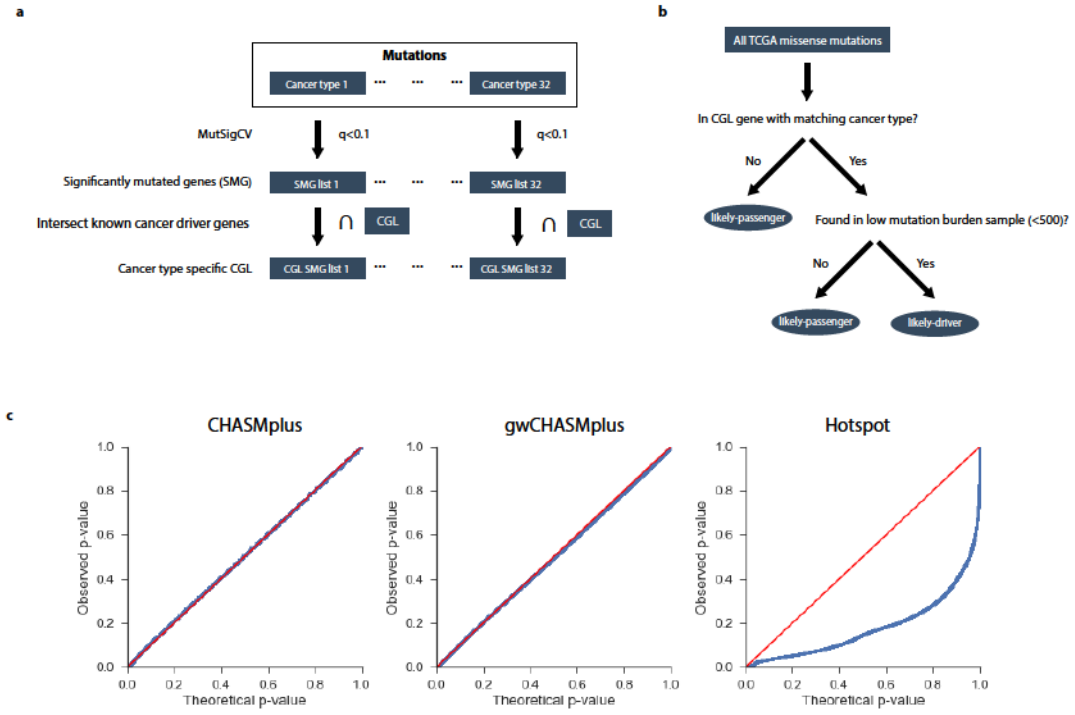

Figure 6.2: Training set labeling procedure and calibration of statistical model. a) Diagram demonstrating how the cancer type specificity of Cancer Genome Landscape (CGL) genes were determined. b) Somatic missense mutations were labeled either as "likely-passenger" or "likely-driver" based on a semi-supervised approach using two steps: overlap with previously known genes from CGL in a cancer type specific manner and samples with low mutation burden. c) QQ plot of observed p-values for a method (blue line) compared to theoretically expected under the null hypothesis (red line). All mutations in genes found in the Cancer Gene Census were removed to eliminate possible driver mutations in this comparison. CHASMplus represents unweighted CHASMplus scores, gwCHASMplus represents gene weighted CHASMplus scores, and Hotspot is a previous codon-level mutation hotspot detection method.

6.2.4 Statistical significance

CHASMplus can also evaluate the statistical significance of cancer type-sepcific predictions for each of 32 cancer types from The Cancer Genome Atlas (TCGA), and pan-cancer predictions for all TCGA cancer types in aggregate (Figure 6.1D). Because Random Forests do not intrinsically include hypothesis testing techniques, I used simulated mutations to assess the statistical significance of scores. P-values were estimated from a simulated null distribution, controlling for sequence composition, and corrected for multiple testing with the Benjamini-Hochberg method (see section 2.3). The resulting P-value distributions suggest our statistical model is well calibrated (Figure 6.2B). Well-calibrated P-values enable quantitative estimates of false discovery rate and thus inform a user about how to select a suitable score threshold for predicted driver missense mutations.

6.3 CHASMplus dramatically improves identification of missense mutation drivers

I next sought to compare the performance of CHASMplus on seven mutation-level benchmarks with respect to 12 comparable methods: VEST [106], CADD [139], FATHMM cancer [64], SIFT [121], MutationAssessor [140], REVEL [60], MCAP [61], ParsSNP [62], CHASM [39], Polyphen2 [141], transFIC [142] and CanDrA [63]. Scores were obtained by means made available by each of the methods.

My benchmarks fall under three broad categories: in vitro experiments, high throughput in vivo screens, and curation from published literature. Each of these categories has weaknesses, but, in aggregate, they span multiple scales of evaluation and amount of supportive evidence (Figure 6.3A). For example, several benchmarks are limited to one or a few well-established driver genes, while others are exome-wide, but lack experimental support. A range of benchmarks is critical because missense mutations with the most established experimental support for a driver role tend to be in a few well-understood cancer driver genes. However, limiting benchmarking to these genes makes it difficult to assess the generalizability of a method's performance to missense mutations

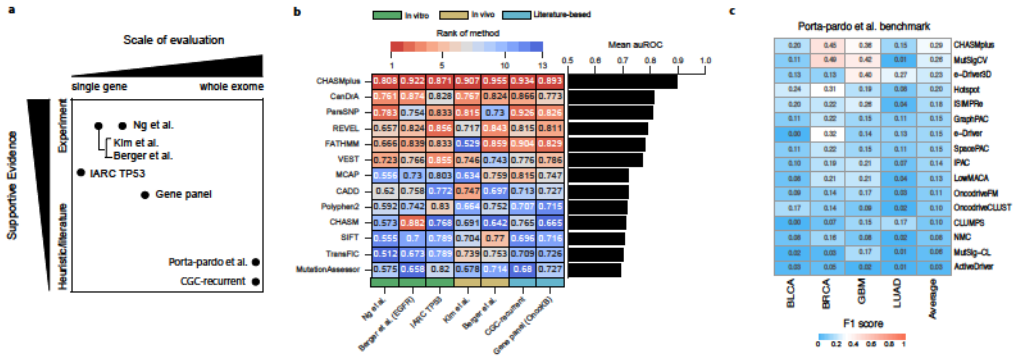

Figure 6.3: Cancer driver prediction benchmark. a) Conceptual diagram of how 8 benchmarks compare in terms of the scale of evaluation and amount of supportive evidence. b) A heatmap showing performance measured by the area under the Receiver Operating Characteristic Curve (auROC) on the 7 mutation-level benchmarks (shown in text). The color scale from red to blue indicates methods ranked from high to low performance. Benchmarks are categorized by in vitro (green), in vivo (yellow), and literature-based benchmarks (turquoise). The bar graph shows the mean auROC across the benchmarks. c) Heatmap showing performance (F1 score) on a cancer type specific benchmark. The overall performance on four cancer types (BLCA, BRCA, GBM, and LUAD) is measured by the average F1 score (right column).

in other genes. All benchmark evaluations used the area under the Receiver Operating Characteristic Curve (auROC) as a metric (Figure 6.3B). Overall, CHASMplus had a mean auROC of 0.09 higher than the next best method. This common metric is used in machine learning to describe how well predictions separate two classes without a priori selecting a score threshold, which for many methods is not well defined [143]. In our assessment, the two classes represent likely driver and passenger missense mutations. In general, auROC values range from 0.5 (random prediction performance) to 1.0 (perfect).

I used three benchmarks based on in vitro experiments. The first was a set of missense mutations assessed by an assay of cell viability in two growth-

factor dependent cell lines, Ba/F3 and MCF10A (pro-B and breast epithelium cell lines), covering 747 mutations in 48 genes [144]. CHASMplus had significantly higher performance than the next best performing method (ParsSNP) ($p<0.05$, delong test). In the second benchmark, an in vitro assay of EGFR resistance to erlotinib from missense mutations observed in lung adenocarcinoma [145], CHASMplus (auROC=0.92) outperformed all other methods, with the next best method (CanDrA) having an auROC of 0.87. CHASMplus auROC was significantly better than that of 7 of the methods tested ($p<.05$, delong test). For the remaining 5 methods, the improvement was not significant, possibly due to lack of power given the small number of mutations (n=75) tested in the assay. In the third benchmark, an assay of reduced transactivation (<50% WT, median of 8 targets) in TP53 from the IARC database (n=2,314 mutations) [146], CHASMplus significantly outperformed the next best method (REVEL) ($p=0.02$, delong test).

To investigate whether CHASMplus would also perform well when compared to results of in vivo experiments, I considered two benchmarks based on pooled in vivo screens in mice that assessed mutation driver status by fitness in a competition assay. The first was performed from mutations observed in lung cancers (44 missense mutations) [145] and the second from mutations observed in 27 cancer types (71 missense mutations) [147]. CHASMplus had the highest auROC of the 13 tested methods on both benchmarks, with an increase

in auROC by 0.09 and 0.1, respectively, compared to the next best methods (ParsSNP in the first benchmark and FATHMM in the second). The increase was significant in the second larger, benchmark (p=0.03, delong test, n=72), but not in the first, which may be the result of the smaller sample size. In the first benchmark, CHASMplus was significantly better than 9 out of 12 tested methods ($p<0.05$, delong test, n=44).

Experimental testing of mutations across large number of genes or the whole exome is currently not feasible. Therefore, evaluation of CHASMplus at larger scales relied on two benchmarks based on literature and database curation. The first benchmark in this category labeled recurrent missense mutations within genes in the Cancer Gene Census [112] as drivers. I found that the gene weighted CHASMplus scores (auROC=0.934) were substantially better at this whole exome-wide prioritization task compared to the unweighted CHASMplus scores (auROC=0.893) ($p<2.2e-16$, delong test). CHASMplus scores were also significantly better than the next best method (ParsSNP) (p=0.001, delong test). The second benchmark was derived from a large driver gene panel (MSK-IMPACT, 414 genes) and 10,130 sequenced cancer patients [148]. Missense mutations were labeled as drivers if they were annotated as such in OncoKB [149], a knowledge-base that aggregates known literature. CHASMplus significantly outperformed all other methods, the nearest being ParsSNP (p=7e-14, delong test).

6.4 CHASMplus improves identification of cancer type specific driver genes

I evaluated the performance of CHASMplus on identifying cancer-type specific driver genes, using a previously published benchmark and assessment of 15 computational methods designed for this purpose [30]. The 15 methods are: Hotspot [42], NMC [150], OncodriveCLUST [55], MutSig-CL [69], iSiM-PRe [151], iPAC [54], GraphPAC [52], SpacePAC [53], CLUMPS [58], e-Driver [152], e-Driver3D [48], ActiveDriver [72], LowMACA [47], OncodriveFM [40], and MutSigCV [33]. Genes were labeled by their designations in the Cancer Gene Census as a cancer driver gene for a specific cancer type. Out of the 4 cancer type cohorts assessed (BLCA, BRCA, GBM, and LUAD), CHASMplus had the highest average F1 score, a balance between precision and recall that was used as a performance metric by [30] (Figure 6.3C). I additionally note that of the methods tested, CHASMplus was the only one not primarily designed to predict driver genes that had high recall (average recall=.45) while maintaining precision (average precision=.23).

6.5 CHASMplus identified both common and rare cancer drivers

Certain cancer driver mutations primarily occur in a specific cancer type, while others appear in many cancer types. The power to detect driver mutations, which occur at low frequency in many cancer types, is increased when many cancer types are aggregated, known as a pan-cancer analysis. Conversely, driver mutations, which are specific to a particular cancer type, are best identified when cancer types are analyzed individually [80]. Using CHASMplus, I identified 3,527 unique missense mutations as statistically significant drivers by pan-cancer analysis at an estimated false discovery rate of 1%. When applied to each cancer type individually, the number found significant varied substantially from 8 in thymoma to 572 in bladder urothelial carcinoma with a median of 78 (Figure 6.4A). The median overlap with literature-based oncogenicity annotation from OncoKB was 53%, suggesting 47% of the driver missense mutations identified by CHASMplus either have not been previously characterized or not sufficiently characterized for inclusion in OncoKB. While OncoKB missense mutation annotations are not cancer-type specific, the genes with highest frequencies of cancer-type specific driver missense mutations identified by CHASMplus have well-known roles in cancer [21] (Figure 6.4B).

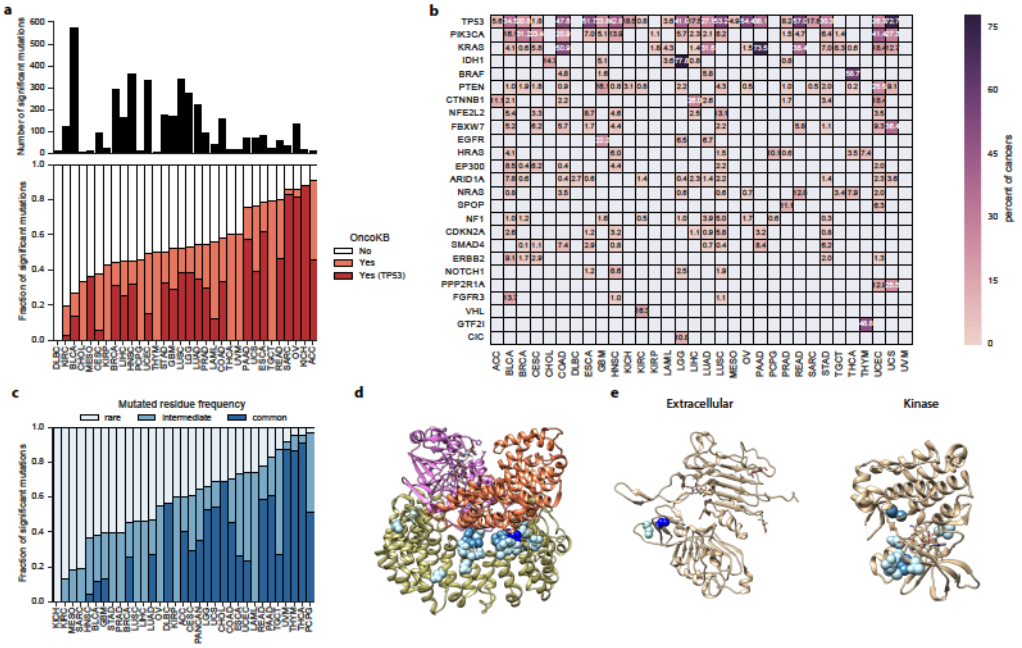

Figure 6.4: Discovery of driver missense mutations. a) Bar graphs showing the number of unique driver somatic missense mutations (top) and the proportion previously known in OncoKB, a literature curated database (bottom). b) Heatmap of the top 25 genes containing the most frequent driver somatic missense mutations in TCGA across the cancer type specific analyses. Shown are the percentage of samples that are mutated. c) Proportion of overall frequency of driver somatic missense mutations found in rare (<1% of samples or singleton mutations), intermediate (1-5%), and common (>5%) driver somatic missense mutations. Correspondingly shown as light to dark blue. d) Structure of the Phosphatase 2A holoenzyme (PDB 2IAE). e) Structures of the ERBB2 extracellular domain (left, PDB 2A91) and kinase domain (right, PDB 3PP0).

The long tail hypothesis, proposed from examining overall mutation frequency of driver genes [74,75], suggests there are many rare drivers. However, the overall mutation frequency of a gene does not account for the confounding presence of passenger mutations within a driver gene. From our mutation-level analysis, I observed that the relative prevalence of rare (<1% of samples), intermediate (1-5%), and common (>5%) driver missense mutations varied substantially among cancer types (Figure 6.4C). For example, uveal melanoma was dominated by common driver missense mutations (88%), while head and neck squamous cell carcinoma (HNSC) was dominated by rare driver missense mutations (63%). Interestingly, from the pan-cancer analysis, the overall proportion of driver missense mutations considered rare was only slightly smaller than for common drivers (35.4% and 35.5%, respectively), but 4-fold greater than found by a previous method (8%, P¡2.2e-16, Fishers exact test) [9].

Rare driver missense mutations exist not only in rare driver genes, but also may be spatially proximal in protein structure to common driver missense mutations. For example, the protein phosphatase PPP2R1A, which has been implicated as a tumor suppressor gene in many tumor types [59], contained common driver missense mutations in our pan-cancer analysis at residue positions 179 and 183, which is located at the protein interface composing the phosphatase 2A complex (Figure 6.4D). It also had a much broader set of rare drivers throughout the protein interface, such as R105Q and R459C. Similarly,

CHASMplus identified common driver missense mutations (S310A/F/Y) in the extracellular domain of the well-known oncogene ERBB2, but also finds rare driver missense mutations in both the extracellular and kinase domain (e.g., L313V and R678Q) (Figure 6.4E). This is supportive of previous experimental work implicating rare cancer driver mutations in common cancer driver genes [19].

6.6 Mutation hotspot detection has limited power

A codon or small region of protein sequence or structure where recurrent mutations are observed is known as a hotspot. Similar to statistical methods for driver gene detection, hotspot detection identifies an excess number of mutations compared to expectation. using a large number of cancer samples. I asked whether, given current cohort sizes, codon-based hotspot detection had sufficient statistical power to identify rare driver mutations. I assessed the number of samples required to detect driver mutations across a range of frequencies (proportion of samples in which a mutation occurs) and somatic background mutation rates. In Figure 6.5A, each of the 32 TCGA cancer types is placed according to its sample size and background mutation rate, relative to six curves which represent the required sample size to detect driver mutations

of a certain frequency, with 90% power, using hotspot detection. For example, the TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) cohort has 274 samples and a background mutation rate of 3.5 mutations/Mb. This sample size is sufficient to detect driver mutations that occur in 2% of the samples with 90% power.

At current TCGA sample sizes, I found codon-based hotspot detection approaches were not well powered to identify driver mutations that occurred at less than 1% frequency in most cancer types. Exceptions were thyroid carcinoma (THCA), low grade glioma (LGG) and breast cancer (BRCA), which are seen to lie above (or close to) the curve representing 1% frequency (Figure 6.5A). Notably, these cohorts had large numbers of samples and low-to-medium background mutation rates. I also found that when cancer types were aggregated in pan-cancer analysis, power to detect codon-based hotspots improved substantially, but only when the recurrent mutations were shared in more than one cancer type. For these mutations, pan-cancer analysis using 10,000 TCGA samples should enable detection of driver mutations at frequency as low as 0.1%.

In our pan-cancer analysis, CHASMplus had greater sensitivity to detect putatively oncogenic missense mutations than a recently published codon-based hotspot detection method (Figure 6.5B). I compared the missense mutations in the TCGA pan-cancer cohort that were called statistically significant by

CHASMplus and those called by a hotspot method described by [9] ($q \leq 0.01$ for both methods). For each method, I computed the overlap with well-curated

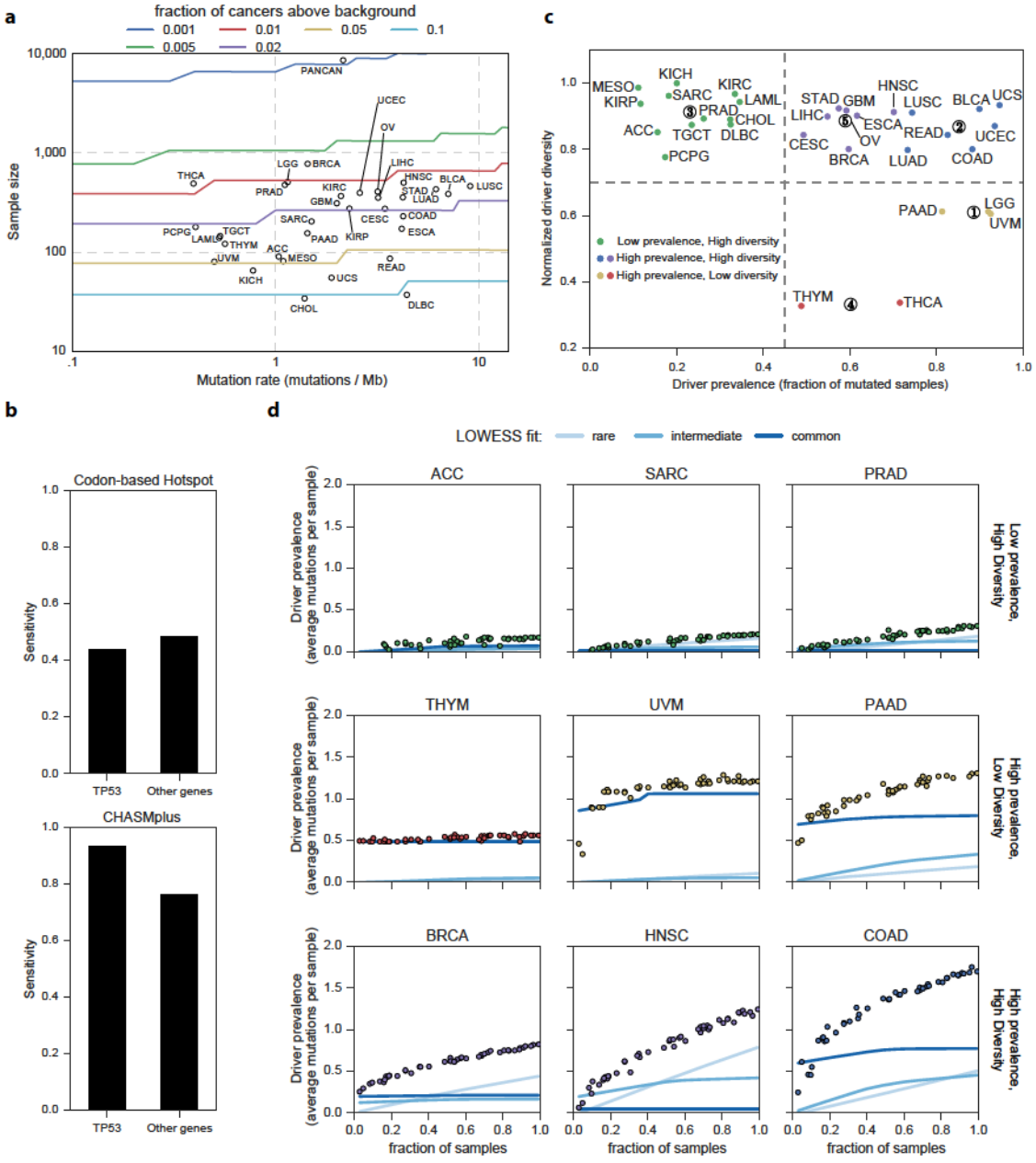

Figure 6.5: (Caption next page.)

Figure 6.5: (Figure: previous page) Saturation and characteristics of driver somatic missense mutations. a) Statistical power to detect significantly elevated non-silent mutations for individual codons as a function of sample size and mutation rate. Circles represent each cancer type from the TCGA, and is placed according to sample size and median mutation rate. Curves are colored by the effect size of the driver mutations (fraction of non-silent mutated cancer samples above the background mutation rate). b) Bar graph comparing sensitivity to detect labeled oncogenic driver missense mutations from OncoKB between CHASMplus and a hotspot detection approach. c) Plot displaying normalized driver diversity and driver prevalence (fraction of samples mutated) for driver somatic missense mutations in 32 cancer types. K-means clustering identified 5 clusters with centroids shown as numerically designated circles. d) Prevalence of driver somatic missense mutations as a function of sample size. Lines represent LOWESS fit to different rarities of driver somatic missense mutations.

oncogenic mutations in the OncoKB database. CHASMplus sensitivity to detect the OncoKB-labeled mutations was 0.83. The sensitivity of the hotspot method (0.46) was significantly lower ($p<2.2e-16$, McNemar's test, n=896). To minimize gene bias, I also repeated the analysis after excluding all 389 TP53 mutations, yielding sensitivity of 0.76 for CHASMplus and 0.49 for hotspot detection, a difference which is still statistically significant (p¡2.2e-16, McNemar's test, n=507). Moreover, these results are also reflected in the number of significant predictions of the two methods. The codon-based hotspot method only identified 360 unique codons as significant in our TCGA data set, while CHASMplus found significant missense mutations in 2,588 codons. I believe that the increased sensitivity is the result of CHASMplus using a broad range of important features, including multi-resolution hotspot detection and weighting by driver gene scores (Figure 6.5C). Importantly, my increased sensitivity

did not come at the cost of low specificity, as evidenced by our p-value calibration and extensive ROC analysis across seven benchmarked datasets, which measures a balance of sensitivity and specificity.

6.7 Characterizing cancer types and the trajectory of discovery

The diversity and prevalence of driver missense mutations varied considerably across TCGA cancer types (Figure 6.5C). I defined diversity with respect to the distribution of driver missense mutations across codons and prevalence with respect to the frequency of the mutations in tumor samples. High diversity indicated mutations were broadly distributed across codons, while high prevalence indicated driver missense mutations that occurred in a large number of tumor samples. Using K-means clustering, I found that cancer types grouped into high diversity and low prevalence (12 cancer types), high diversity and high prevalence (15 cancer types), and low diversity and high prevalence (5 cancer types). These differences were not associated with intra-tumor heterogeneity or normal contamination, as assessed by mean variant allele fraction (VAF) of a cancer type ($p>0.05$, correlation test). The differences also could not be associated only with TCGA sample size for a particular cancer type. For example, while both pancreatic ductal adenocarcinoma (PAAD) and

sarcoma (SARC) had similar sample sizes (n=155, n=204 respectively), PAAD had high prevalence and low diversity, while SARC had low prevalence and high diversity. After adjusting for sample size, I observed that the average mutation burden for a cancer type positively correlated with the prevalence of rare (but not common) driver missense mutations (R=0.63, P=4.7e-5, likelihood ratio test).

Are there substantially more cancer driver missense mutations yet to be discovered? If discovery was measured by the number of unique driver missense mutations identified, subsampling analysis showed all cancer types had a linear increase ($R^2 > 0.5$) with no evidence of saturation at current sample sizes (Figure 6.6). However, I did observe substantial variability in trajectories if discovery was measured by driver prevalence (average number of driver missense mutations per cancer sample) (Figure 6.5D), a metric which goes directly to utility of driver discovery in clinical practice (Discussion). For sarcoma (SARC), adrenocortical carcinoma (ACC), and prostate adenocarcinoma (PRAD), driver prevalence remained minimal as sample size increased. While, in contrast, thymoma (THYM), uveal melanoma (UVM), and pancreatic ductal adenocarcinoma (PAAD) contained common driver missense mutations that could be detected by using only a few samples from the cohort, e.g., GTF2I L424H in THYM. Due to a lack of rare or intermediate driver missense mutations, I observed THYM and UVM saturated discovery as sample size in-

creased. Although PAAD did show a growing set of intermediate/rare driver missense mutations, the overall driver prevalence exhibited a diminishing rate of discovery. In contrast, breast (BRCA), head and neck squamous (HNSC), and colon cancers (COAD) harbored a full spectrum of driver missense mutations, with rare drivers increasing substantially as a function of sample size.

6.8 Discussion

CHASMplus was designed to better represent the context in which missense mutations occur, by coupling prior information about a mutation's likely functional importance and mutational patterns evident from large cancer sequencing studies. I compared CHASMplus with 27 other computational methods, including the original CHASM, on eight benchmarks covering in vivo experiments, in vitro experiments, and literature curation, CHASMplus had the best performance at predicting drivers at each scale of evaluation - a whole exome, a targeted gene panel, and within a single gene. Individually, none of the benchmarks was ideal. For example, mutations in the in vitro or in vivo benchmarks, were selected by complicated study inclusion criteria and limited by resource constraints. However, I believe that application of multiple independent benchmarks spanning a wide array of genes is the current best practice.

The long tail hypothesis [74, 75] posits that there are many rare driver mu-

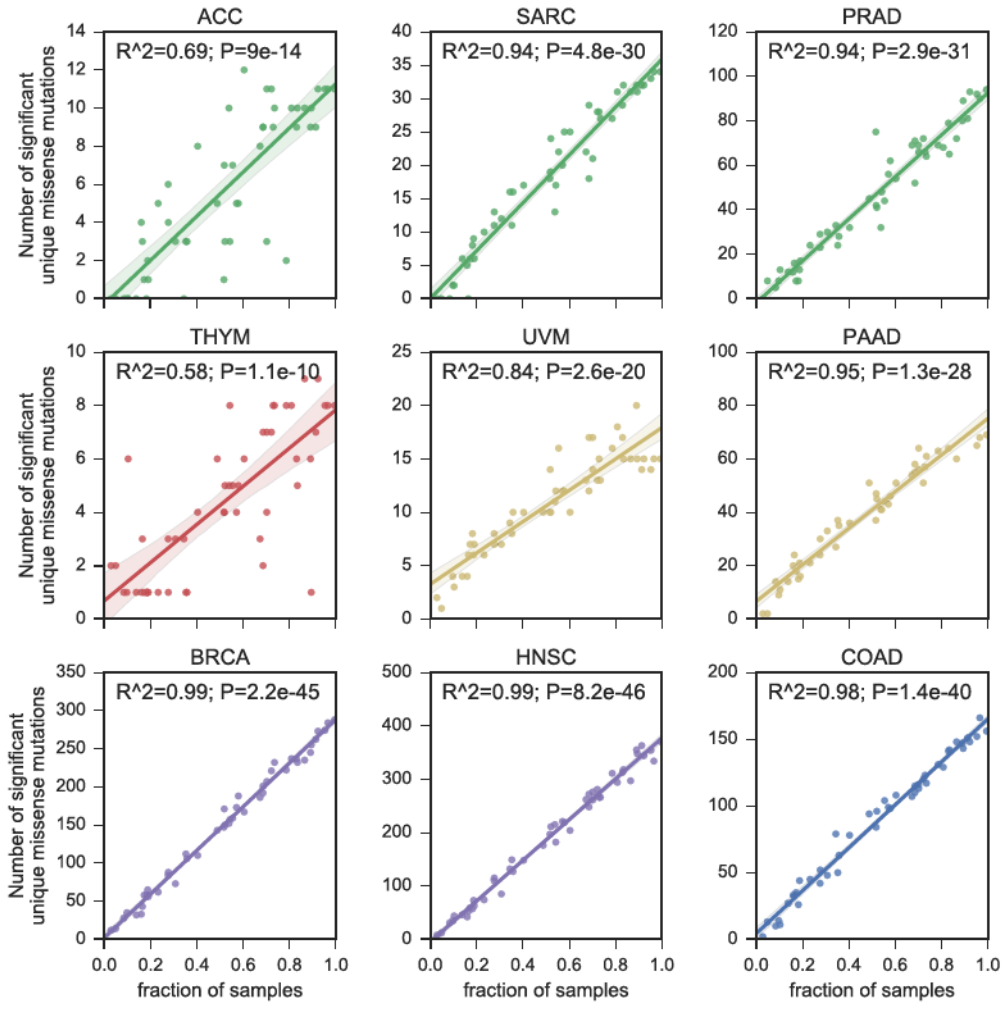

Figure 6.6: Subsampling analysis of unique driver somatic missense mutations by CHASMplus. The number of driver somatic missense mutations identified as significant by CHASMplus ($q \leq 0.01$) as a function of sample size. CHASMplus was ran on random subsets of various sizes (fraction of samples) of the full data.

tations in human cancers. To assess this hypothesis, I leveraged the improvements made in CHASMplus to systematically predict driver missense mutations in 8,657 samples from the TCGA. Although individually rare, I found that rare driver missense mutations played a prominent role in aggregate, consistent with the long tail hypothesis. This result supports the critical role of assessing the prevalence of driver mutations – failure to capture and identify rare driver mutations, which occur in aggregate at reasonable prevalences, may result in crucial missed opportunities. Because high-throughput functional validation studies of missense mutations are not yet widespread, computational methods, like CHASMplus, are needed to prioritize mutations for low- and medium-throughput studies. A key advantage of CHASMplus is that I can precompute a score for every possible missense mutation, forming an *in silico* saturation mutagenesis across all genes to capture rare driver mutations yet seen mutated.

To my knowledge, mine is the first study to show that the prevalence and diversity of driver missense mutations is highly variable across the cancer types represented in the TCGA. I observed that mutation burden for a cancer type positively correlated with prevalence of rare (but not common) driver missense mutations, even after correcting for sample size, suggesting that accumulating a greater number of mutations in a cancer may increase the competitiveness of rare drivers. More research into the origins of rare driver mutations is war-

ranted, because differences in the rarity of driver missense mutations could arise from a variety of factors, including the driver mutation's strength, dependence on genetic or environmental factors, competition from other types of tumor-derived alterations, or role in cancer subtypes.

Chapter 7

Comprehensive discovery of driver genes and mutations in cancer

Over the past decade, The Cancer Genome Atlas (TCGA) has coordinated a monumental enterprise of data generation and genomic investigation across 33 cancer types, and numerous notable findings have emerged from this project (`https://cancergenome.nih.gov/publications`). The individual TCGA projects also motivated the development of many bioinformatic algorithms oriented toward discovery, characterization, and prioritization of cellular processes driving cancer based on pathways [153], genes [31], or individual variations [154]. However, despite this remarkable progress, algorithms do not

entirely agree on certain candidate cancer driver genes and mutations, necessitating continued expert curation to filter likely false positive findings. Moreover, previous PanCancer analyses [70] have been limited to fewer cancer types and have largely avoided nominating rare driver mutations. This chapter is work done as a co-leading analyst in the Driver's group within the TCGA PanCanAtlas (hereafter referred to as driver's group).

7.1 Material and methods

7.1.1 Mutation calling quality control

A publicly available MAF file (`https://synapse.org/MC3`) was recently compiled by the MC3 Working Group and is annotated with filter flags to highlight potential artifacts or discrepancies. This dataset represents the most uniform attempt to systematically provide mutation calls for TCGA tumors. The MC3 effort provided consensus calls from 7 software packages [155]. Flagged artifacts include: non-exonic regions, whole-genome amplified (WGA) samples, exclusion lists, blood/tumor derived pairs, strand-bias, contamination estimations, oxo-guanine artifacts, low normal read depth, polymorphisms common in EXAC [156], mutations present in a panel of normal samples, non-preferred tumor normal pairs, and mutations outside the regions of interest for any caller.

If a mutation was not assigned any flag and was called by 2 or more variant calling software packages, it received a 'PASS' identifier. I restricted our analysis to PASS calls with the exception of samples from OV and LAML, which were some of the earliest sequenced by TCGA. Preparations for these samples utilized whole genome amplified (WGA) DNA, an important factor in that the WGA process can induce artefactual mutations. Of the 412 OV and 141 LAML samples present in our data 347 (84%) and 141 (100%), respectively, had variants derived from WGA DNA. In order to maintain sample sizes and uniformity in mutation calling, I did not filter mutations containing only 'wga' filter tags from these two cancer types. I recognize multiple limitations of this mutation call set including the lack of structural variants and copy number alterations, as well as variability in sequencing depth and tumor purity. The above limitations may lead to variability in mutation detection; however, the MC3 dataset reflects the state-of-the-art in consensus mutation detection.

I also excluded highly mutated samples. These hypermutators were defined as samples with a mutation count exceeding Tukey's outlier condition, i.e. greater than 1.5 times the interquartile range above the third quartile in their respective cancer types (3Q + 1.5*IQR). Designation as a hypermutator also required the number of mutations in a sample to exceed 1000, a heuristic that limited the number of discarded samples in low mutation rate cancer types. LUAD, SKCM, and UCEC had hypermutator thresholds greater than

1000 mutations (1047, 2122, and 2545 respectively). I also excluded samples that were flagged by the analysis-working group based on pathology, but allowed "RNA degradation" samples to remain, as this factor is not particularly relevant for most driver prediction tools based on mutations. The final driver-discovery dataset consisted of 9,079 samples having a total of 791,637 missense mutations, 323,884 silent mutations, 96,196 3'UTR mutations, 57,900 nonsense mutations, 42,251 intronic mutations, 42,251 Frame shift deletions, 34,266 5' UTR, 21,804 splice site mutations, 19,856 RNA mutations, 11,305 frame shift insertions, 7,622 3' flanking mutations, 6,419 5' flanking mutations, 6,144 in-frame deletions, 1,362 translation start site mutations, 964 nonstop mutations, and 632 in-frame insertions.

7.1.2 Driver gene discovery approach

Using multiple tools can overcome numerous technical issues that confound individual statistical analyses to find driver genes, such as heterogeneous mutation rate across the genome [33], inflated significance for long genes [157], and false positive calls in cancers with high mutation rates [85]. In the first phase, 8 different tools comprising algorithms based on mutation frequency (MuSiC2 [73] and MutSig2CV [69]), features (20/20+ [85], CompositeDriver(in preparation) and OncodriveFML [99]), clustering (OncodriveCLUST [55]), and externally defined regions (e-Driver [152] and ActiveDriver [71]) were used

(Figure 7.1A). Each tool reported gene or mutation level scores and/or p-values along with a brief description of recommended cutoff thresholds or filters.

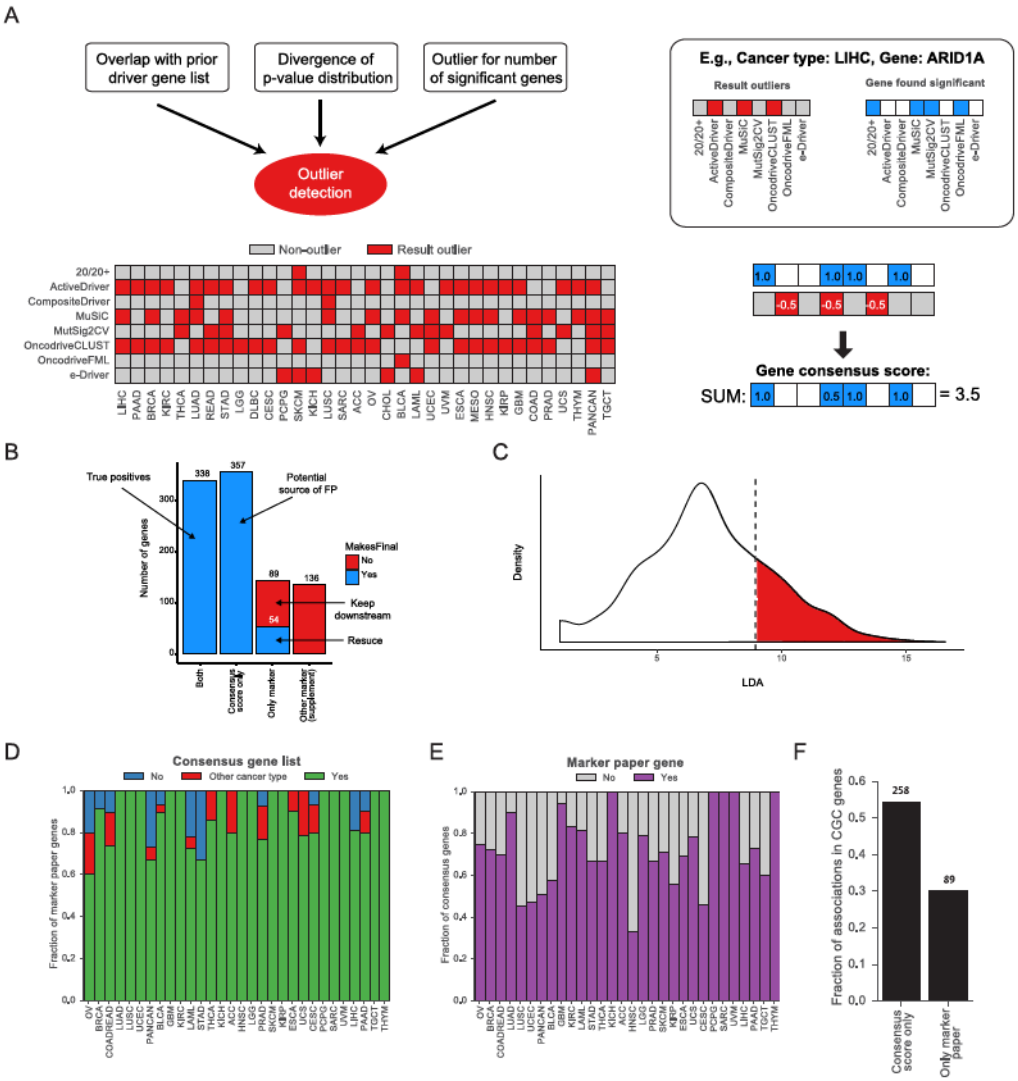

Figure 7.1: (Caption next page.)

Figure 7.1: (Figure previous page) Consensus Gene scores and SMG filtering. (A) Left, outlier detection was performed on a per analysis and method basis. Outliers were marked (red) based on the quasi-majority of three criteria: (1) low concordance with known cancer genes from Vogelstein et al (lower than median); (2) high divergence of p-value distribution from theoretical expectation (higher than median); and (3) abnormally high number of significant genes (>1.5x the interquartile range above the third quartile). The first two criteria were assessed based on the other tools within a single analysis, while the third criterion was assessed based on the same tool's results over all the individual cancer types (excluding the PanCancer analysis). Right, example calculation of the gene consensus score for ARID1A in the cancer type LIHC. A result from an outlier is down weighted, receiving a weight of 0.5 instead of 1.0. The gene consensus score is the sum of weights for tools finding that gene as significant. (B) Overlap of consensus gene list with prior TCGA marker papers. (C) Likely false positives were detected with a high Linear Discriminant Analysis (LDA) score threshold representing 90% sensitivity for keeping associations found in Cancer Gene Census genes. LDA was trained to distinguish common false positives in exome sequencing from previous TCGA PanCancer marker papers. The LDA threshold was only applied to the potential source of false positive genes. (D) Fraction of marker paper genes highlighted in the main text that were also found in our consensus gene list. (E) Fraction of our consensus gene list found in previous TCGA marker papers. (F) Fraction of associations found in the Cancer Gene Census (CGC) that were either found only in the consensus gene list or TCGA marker paper.

7.1.2.1 Consensus methodology

I identified a preliminary total of 2,101 potential drivers by taking the union of genes predicted by the eight driver-gene discovery tools. As illustrated in Figure 7.1A, the increased number of false positive genes is likely due to any individual tool's capability to maintain sound statistical properties that handle a complex set of factors such as tumor heterogeneity, increased mutation rates, and variable sample sizes. I refined this list by calculating, for each gene predicted in each cancer type, a consensus score that compensated for outlier

results and correlation among tools (Figure 7.1). The consensus score was defined as a weighted sum of the number of tools that predicted the gene to be a driver in each cancer type (see subsubsection 7.1.2.2). I required a minimum of two tools to agree, where both could not be outliers (score≥1.5). Although it is difficult to distinguish the overall performance improvement on a small number of held out CGC genes (Figure 7.2A), the weighting strategy did have higher specificity (p=4.3e-8, McNemar test), which is preferable given concerns of false positives. Regardless, the consensus score performance on identifying CGC genes (Figure 7.2A) support previous reports that merging the results from different algorithms improve cancer driver discovery [70].

To maximize the coverage of our analysis and ensure the accuracy of our final list, previous findings were reviewed in 31 individual cancer types and PanCancer-12 from TCGA. For cancer types not yet having a TCGA publication, the relevant analysis working groups were consulted (LIHC, TGCT, UVM, SARC, PAAD, and THYM). I included in our final consensus list all those genes that were previously described as drivers by experts in the cancer-specific analysis of TCGA datasets and were also identified by at least one of the eight algorithms, even if they did not meet our consensus score threshold (≥1.5)(Figure 7.3A). This resulted in an additional 54 gene-cancer pairs, such as *ATR*, *CHEK2*, *IDH2*, and *ERCC2* in the PanCancer dataset and FOXA1 in BLCA, *HRAS* in SKCM, and *MET* in LUAD (Figure 7.1B-F). The majority of

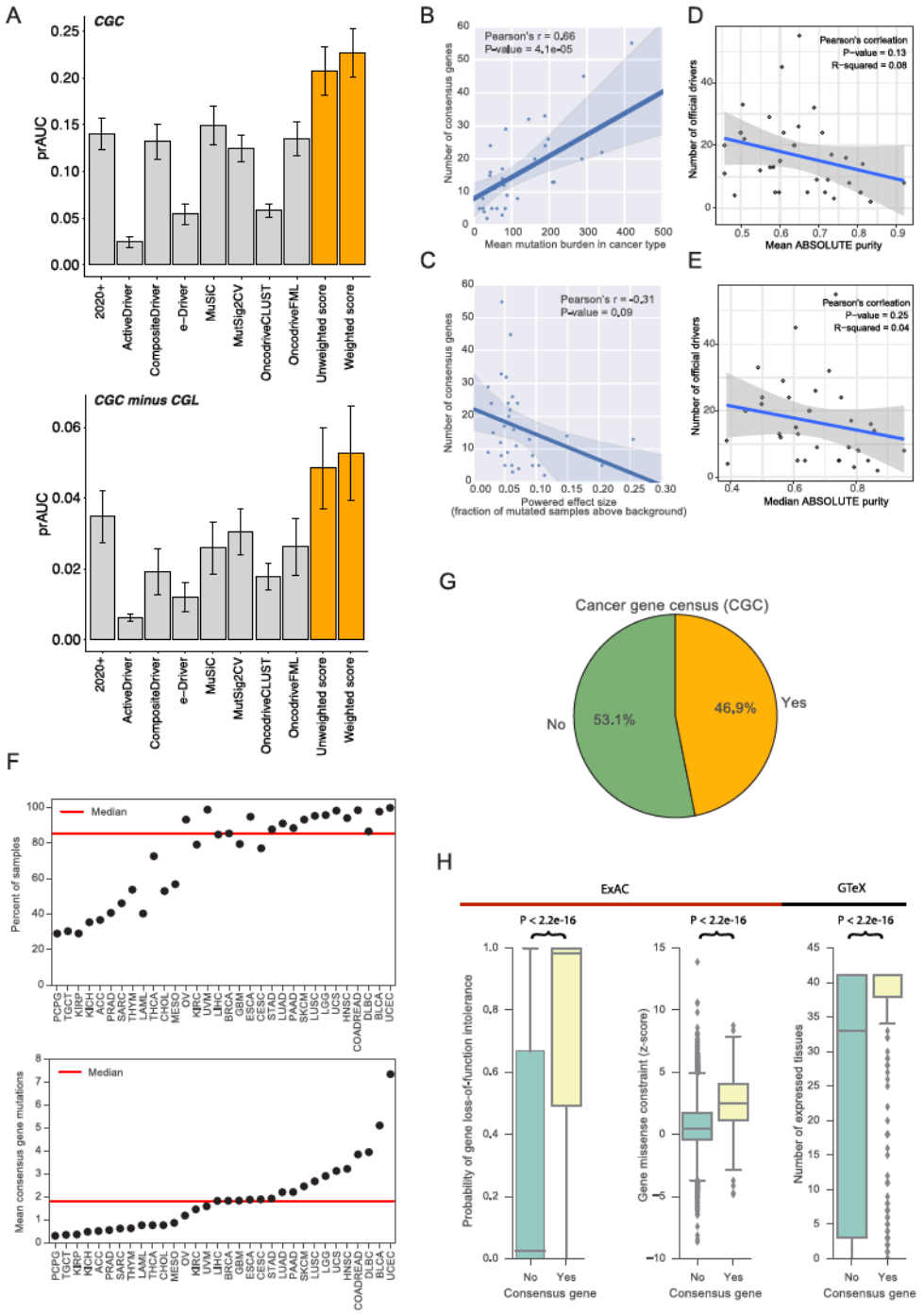

Figure 7.2: (Caption next page.)

Figure 7.2: (Figure previous page) Characteristics of consensus genes. (A) Predictive power of each individual driver gene detection method (in gray) and of the weighted and weighted scores (in orange). The predictive power was measured as prAUC, using all the genes in the Cancer Gene Census and a set that additionally excludes Cancer Genome Landscape genes used in outlier detection. Error bars, calculated by bootstrapping, indicate one standard deviation. (B) The number of consensus genes in each cancer type positively correlated with the average mutation burden. Shaded area indicates 95% bootstrapped confidence interval. (C) Given the variability in powered effect size (fraction of mutated samples above background with 90% power) in this study, there is a negative but not significant correlation with the number of consensus genes in each cancer type. COAD and READ were excluded because analysis was performed separately, but the final consensus genes were merged. (D) Pearson correlation between the number driver genes identified and median purity was calculated and plotted. (E) Pearson correlation between the number driver genes identified and mean purity was calculated and plotted. Summary statistics for p-value and r-squared value are reported in the top right corner of panels D and E. (F) Percent of samples containing a non-silent mutation stratified by cancer type. The red line indicates the median across cancer types (left) and average number of non-silent mutations in consensus genes per sample (right). (G) A pie chart showing the percent of consensus genes which are found in the Cancer Gene Census with annotations for small somatic mutations (missense, splice site, indel, and nonsense) (H) Consensus genes showed a higher probability for loss-of-function intolerance and missense mutation constraint of germline mutations based on ExAC, and were expressed (RPKM>1) in a wider number of tissues from GTeX (version 6). Given the high correlation of gene expression in the 11 brain regions assessed from GTEx, we took the median of multiple brain tissues, as done in Lek et al., 2016.

this effort resulted in linking cancer genes identified by our strategy to additional cancer types based on previous literature (32/54).

To limit false positives in the expanded list, linear discriminant analysis was applied (Figure 7.1C). 45 genes were identified and removed from the consensus as they are likely false positives. These included *CACNA1E* in Pan-Cancer, *COL11A1* in LUAD, *DST* in GBM, and *TTN* in SKCM. The consensus

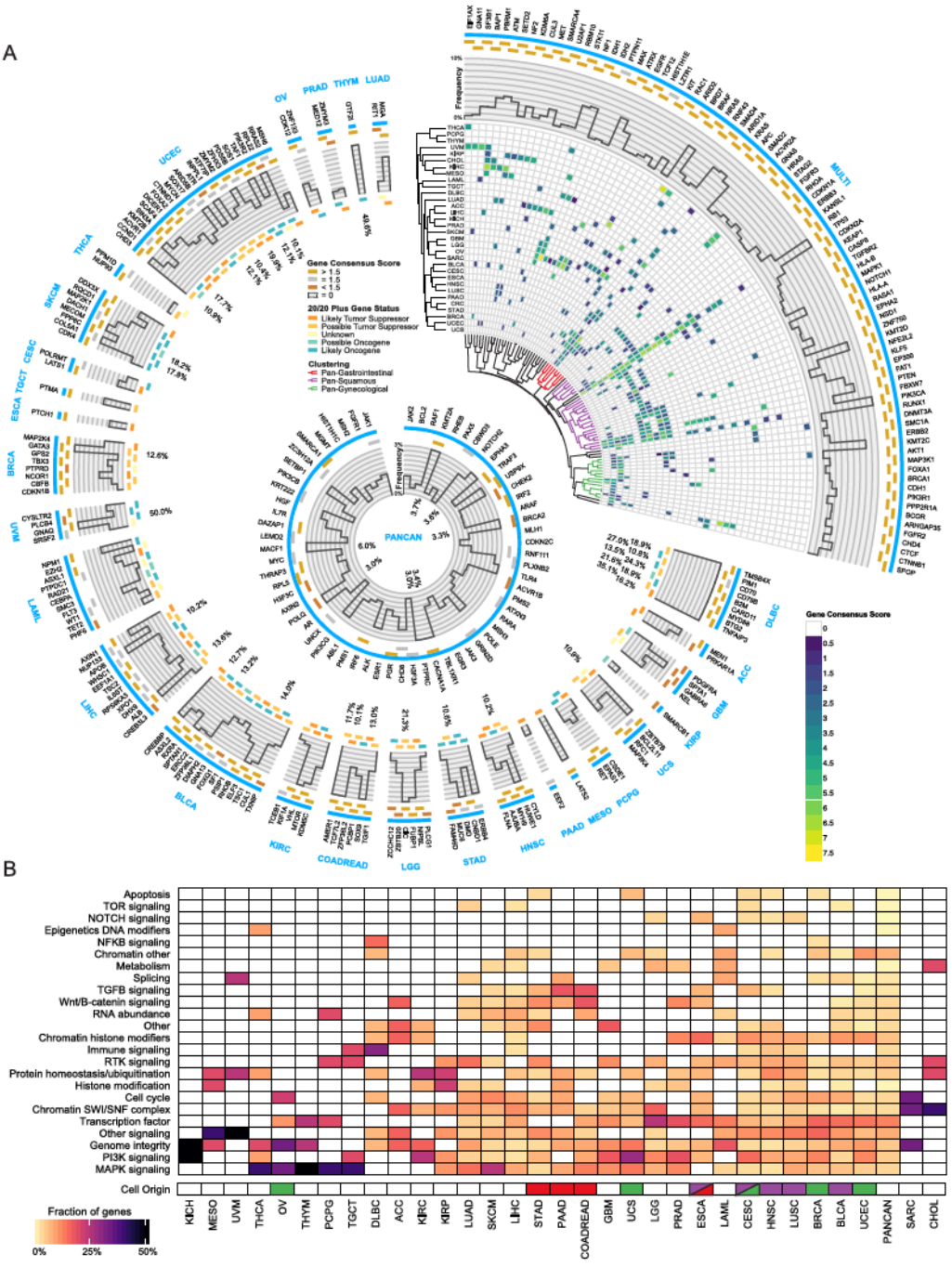

Figure 7.3: (Caption next page.)

Figure 7.3: (Figure previous page) Cancer driver gene discovery: (A) Circos [158] plot displays 299 cancer genes. Each sector indicates a unique cancer type (text in blue) with predicted drivers unique to that cancer type listed (gene name in black). Only tissues with at least one unique driver gene are shown. The top right sector shows all genes found significant in multiple cancer types. Next, a categorical score of gold, silver, or bronze is assigned to each gene based on the highest consensus score. If a gene was not scored and required rescue then the field is empty. The next ring illustrates the mutation frequency of a gene in our dataset. For the top right wedge, the PanCancer frequency is used, while cancer-type-specific frequencies are used in the remaining sectors. Where frequencies exceed the y-axis limit of 10%, the innermost label indicates the frequency. The final ring uses a 5-point scale from orange to teal to represent each gene from likely tumor suppressor to likely oncogene, respectively, by the 20/20+ algorithm. Finally, in the top right slice we show hierarchical clustering of the gene consensus scores for genes that were found in more than one cancer type (note: CRC refers to the COADREAD cancer type). Additionally, significant gene clusters (permutation test) identified Pan-Gastrointestinal (red), Pan-Squamous (purple), and Pan-Gynecological tissues (green). The middle ring illustrates all genes that were only found using PanCancer results, or were otherwise rescued. (B) Heatmap showing clustering of different cancer types by pathway / biological process affected by associated consensus driver genes. Cell of origin for pan-gynecological, pan-gastrointestinal, and pan-squamous are colored as above.

list from the above systematic approach consisted of 258 unique genes. The average number of non-silent mutations per sample in our consensus gene list varied substantially by cancer type ranging from ¡1 in 12 cancer types (ACC, CHOL, KICH, KIRP, LAML, MESO, PCPG, PRAD, SARC, TGCT, THCA, and THYM) to 7.3 in UCEC. A median of 85% of tumors harbored non-silent mutations in consensus genes across cancer types (Figure 7.2F).

Given the limitations of a systematic approach, 41 genes were manually rescued. In the rescue attempt, I started with a list of genes identified from previous TCGA marker papers but not found from our systematic approach. Genes

were rescued with supportive evidence from the following sources: hypermutator phenotype related genes (since we excluded hypermutated samples in our systematic discovery; 6 genes), established cancer genes from LAML because of low quality variant calling originating from liquid tumor contamination of the normal samples (6 genes), genes supported by omic network tools (DriverNet and OncoIMPACT; 25 genes), and a gene supported by all three approaches from the driver mutation discovery (1 gene). Addition of genes to the final list was subjected to expert manual curation (3 genes).

The final consensus gene list consisted of 299 unique genes across 33 cancer types and the PanCancer dataset (Figure 7.3A). The list captures most previously described driver genes for the majority of cancer types. I overlapped the cancer driver genes obtained from the consensus approach without manual curation with those from 5 independent studies in 4 cancer types (BRCA, PRAD, PAAD, and LIHC) of which one is whole-genome sequencing. The consensus approach always had a greater inter-study overlap, with an average increase of 26% over only using a single tool, either MuSiC2 or MutSig2CV [59, 159–163]. Among the 299 genes, 59 novel genes were not previously identified in 6 previous PanCancer publications [10, 14, 59, 69, 70, 164, 165] or the cancer gene census list (`http://cancer.sanger.ac.uk/census/`) [112].

7.1.2.2 Weighting strategy

Tools predicting cancer genes were weighted according to their performance in each cancer type, receiving half the weight if a result was deemed an outlier, thereby obligating additional tool agreement. Specifically, I examined quality metrics across tools and within the same tool, which allowed us to identify outlier results. I marked outliers based on the quasi-majority of three criteria: low concordance with known cancer genes, high divergence of p-value distribution from theoretical expectation, and abnormally high number of significant genes. The first criterion evaluated the fraction overlap of significant genes with a previously manually curated set of driver genes from [14] compared with the median across all tools. The second criterion examined whether the divergence of observed p-values from those theoretically expected by the Mean Log Fold Change (MLFC) [85] was greater than the median of all tools, which may indicate a tool's statistical assumptions may not be well satisfied. The third criterion examined whether a tool's prediction for particular cancer types appeared as an outlier in terms of the number of significant genes compared against all of the results for that tool (Tukey's outlier criterion: number significant $>$ 3Q + 1.5*IQR). I calculated a gene consensus score by summing the tools that declared the gene as being significant, with a weight of 1 for non-outlier results and 0.5 for outlier results.

7.1.3 Driver mutation approach

To maximize the coverage of our analysis I used 12 tools that look for three distinct hallmarks of "driverness". The collection was comprised of 8 mutation-level algorithms (SIFT [25], PolyPhen2 [26], MutationAssessor [140], trans-FIC [40], fathmm [64], CHASM [39], CanDrA [63] and VEST [106]), and 4 structure-based (HotSpot3D [46], HotMAPS [43], 3DHotSpots.org [41] and e-Driver3D [48]). In order to combine the predictions from the sequence-based approaches I used principal component analysis to develop a Combined Tool Adjusted Total (CTAT) scores for both, population-based and cancer-specific scores. Principal component analysis has been previously shown successful in a similar task of prioritizing germline mutations [166]. I also combined the results from three-dimensional tools by adding the number of tools that predicted a specific position as belonging to a cancer-mutation cluster. Finally, to limit the number of false positives, I focused our analysis on the genes of our consensus driver list.

The CTAT score combines multiple individual tools that prioritize missense mutations. To normalize each score, I calculated the z-score by subtracting the mean score and then dividing by the standard deviation. I then performed principal component analysis (PCA) using ScikitLearn v0.18.0 and used the score along the first principal component as our CTAT score, representing the scalar projection onto the first eigenvector. Only missense mutations that had

no missing values for each of the combined tools were used in generating the principal component analysis. I performed this procedure on two distinct categories of tools, "population-based" tools that distinguish damaging/pathogenic germline missense variants from common polymorphisms (SIFT, PolyPhen2, VEST, and MutationAssessor), and "cancer-focused" tools designed to distinguish somatic missense mutations that are drivers from passengers (CHASM, CanDrA, fathmm, and transFIC). To score the remaining missense mutations that did have a missing score, I imputed missing scores of the individual tool with the mean for the method. Imputation was only performed for the cancer-focused tools as the population-based tools had too many missing values.

To define the CTAT score thresholds, I used the maximum balanced accuracy when predicting OncoKB mutations "oncogenic" or "likely oncogenic". This yielded a threshold of 1.2 for CTAT-population and 2.4 for CTAT-cancer. For the structural algorithms, I report a mutation as likely driver if at least 2 algorithms identify it within a cluster. Finally, I evaluated the performance of each CTAT score using mutations from OncoKB labeled as "likely oncogenic" or "oncogenic" as true-positives.

7.2 Results

7.2.1 Mutational data set

Mutation calls were produced by the Multi-Center Mutation Calling in Multiple Cancers (MC3) working group by harmonizing results of 7 algorithms [155]. To reduce the false positive rate for driver gene discovery I implemented three strategies addressing known issues affecting driver detection and data quality (see Mutation calling quality control). The driver detection dataset ultimately consisted of 9,079 samples having 1,457,702 total mutations, where the number of mutations per sample was widely distributed across cancer types and was consistent with previous publications [33, 66, 70].

7.2.2 The landscape of cancer driver genes

The final consensus list consists of 299 unique genes: 258 genes obtained from a systematic approach and 41 additional genes recovered after manual curation of previous TCGA marker papers with the majority (26 out of 41, 63%) supported by additional -omic network tools (DriverNet and OncoIMPACT) not used in original SMG detection. Note that, for the rest of the analyses presented here, I will focus on the 258 genes set, but I acknowledge the limitations of a systematic approach by including the 41 genes rescued by manual curation

in our final list to achieve comprehensiveness.

The list recovers most of the previously described driver genes for the majority of cancer types. In fact, in 20 out the 31 cancer types included in our study that had either been previously published or for which I had an internal list of known cancer driver genes, the recovery rate is 80% or higher (Figure 7.1D and Figure 7.1E). The most significant outliers are STAD and the previous PanCancer study, for which I only recovered around 70% of the previously described genes (Figure 7.1D). The consensus list also includes 59 novel genes that had not been described previously and other known drivers not previously associated with a given tissue. Predictions of known cancer driver genes in new tissues include *ATRX* in ACC, *KMT2C*, *CTNNB1* and *PTEN* in BLCA, and *ARID1A* and *KRAS* in BRCA. Entirely novel predictions include *GNA13* in BLCA (a homologue of the known drivers *GNAQ* and *GNA11*), *RRAS2* in UCEC (with shared homology in *KRAS* and *HRAS*), and *KIF1A* in HNSC (a kinesin of the same family of the cancer driver *KIF5B*).

The number of detected cancer driver genes varies among cancer types, with KICH having the fewest (2 genes) and UCEC having the most (55 genes). Furthermore, the ratio of predicted tumor suppressor genes and oncogenes vary widely by tissue (Figure 7.4). I observed a significant positive correlation (Pearson R=0.66, P value=4.1e-5) between the average mutation burden in a cancer type and the number of identified consensus genes (Figure 7.2B). Study-based

calculations for powered effect size in each cancer type did not entirely explain this phenomenon (Pearson R=-0.31, P value=0.09) (Figure 7.2C). Regarding the associations of driver genes with different cancer types, many genes (142 out of 258) are associated with a single cancer, whereas 87 genes have driver roles in two or more cancer types, with an additional 29 genes uniquely identified using all samples combined-PanCancer approaches. As expected, TP53 is the most extreme case, as it is associated with 27 cancer types, followed by PIK3CA, KRAS, PTEN and ARID1A, each of which is associated with 15 or more tissue types (Figure 7.3A).

I clustered the different cancer types according to the consensus scores of their associated genes. Remarkably, some cancer types grouped according to their tissue of origin, such as LGG and GBM; while others according to their cell of origin. The most significant of the cell origin clusters spans all the squamous cancer types (BLCA, CESC, ESCA, HNSC and LUSC, (permutation test, adjusted $p < 0.01$) and includes several transcription factors (*ZNF750*, *NFE2L2* or *KLF5*), chromatin and histone modifiers (*KMT2D*, *EP300*, or *NSD1*), and various PI3K pathway genes (*PIK3CA*, *PTEN* or *MAPK1*). I found two additional significant clusters (permutation test, adjusted $p < 0.05$) that group gynecological (UCS, CESC, UCEC, OV, and BRCA) as well as gastrointestinal cancers (COADREAD, PAAD, ESCA and STAD) (Figure 7.3A).

Finally, I classified the consensus driver genes according to the cancer-

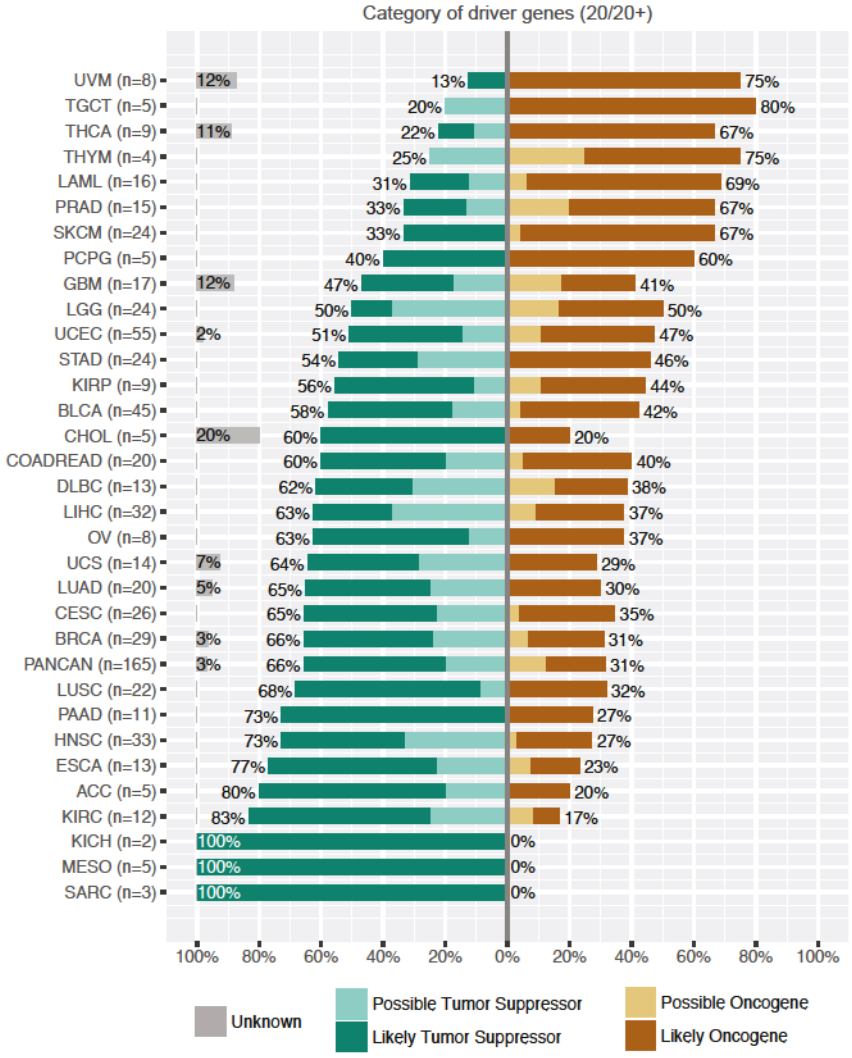

Figure 7.4: Balance of oncogenes and tumor suppressor genes. Percentage of consensus genes predicted as either oncogene (brown), tumor suppressor gene (green), or unknown (gray) by the 20/20+ algorithm, an improved version of the 20/20 rule. The 20/20+ algorithm uses a supervised-learning approach (random forests) and bases predictions on the mutational patterns observed within a gene. "Likely" and "Possible" statuses were determined at a threshold of 0.05 for q-value (Benjamini-Hochberg method) and p-value, respectively. Consensus genes were designated as "Unknown" if they did not meet these thresholds. N represents the number of significant genes in each cancer type.

related biological processes and pathways with which they were associated (Figure 7.3B). For most genes, the categories (excluding "other" and "other signaling") clearly reflect known processes involved in carcinogenesis, as they are "transcription factor" (39 genes), "RTK signaling" (16) and "RNA abundance" (15), "protein homeostasis/ubiquitination" (15), "chromatin histone modifiers" (15), "genome integrity" (14), "chromatin other" (14) and, remarkably, "immune signaling" (10). The last group is of particular interest, given the connection between driver genes and immune response. In terms of cancer types, most have at least one cancer driver that belongs to either genome integrity (28 out of 33 cancer types) or the MAPK or PI3K signaling pathways (24 and 22 cancer types, respectively). Interestingly, the squamous cancer types were again grouped together when looking at which processes and pathways associated with their driver genes, having higher proportions of genes involved in chromatin histone modification as well as receptor-tyrosine kinase and immune signaling.

7.2.3 Discovery of driver mutations

Not all mutations in a cancer driver gene have the same impact on its function [167]. Their consequences frequently depend on which position within the protein is affected and what amino acid change is induced [39]. Here, I sought to explore this topic across the entire PanCancer dataset, classifying 751,876

unique missense mutations by examining the 299 cancer driver genes that I identified, according to their predicted oncogenic effect. I combined the output of three different categories of tools into consensuses approaches: (I) tools that distinguish between benign and pathogenic mutations using sequence-based features (CTAT-population); (II) tools that distinguish between driver and passenger mutations using sequence-based features (CTAT-cancer); and (III) tools that discover statistically significant three-dimensional clusters of missense mutations (Structure-based); these identified 10,098 (1.3% of the total missense mutations), 4,595 (0.6%), and 1,469 (0.2%) unique amino acid substitutions, respectively (Figure 7.5A). The differences in the number of predicted driver mutations for each approach are likely due to the design and requirements of the tools, i.e., dependence of structural clustering tools on available three-dimensional protein structures (either experimental or homology-based) yields fewer predicted driver mutations. Nevertheless, structural tools may provide additional molecular biological context for the identified mutations, which can be particularly relevant for variants of unknown significance (VUS) [168].

When benchmarked against OncoKB [149]-a manually curated dataset of cancer mutations annotated according to likely oncogenic effect, cancer-focused algorithms had higher predictive value than algorithms that distinguished between benign and pathogenic mutations. In addition, the CTAT-cancer score

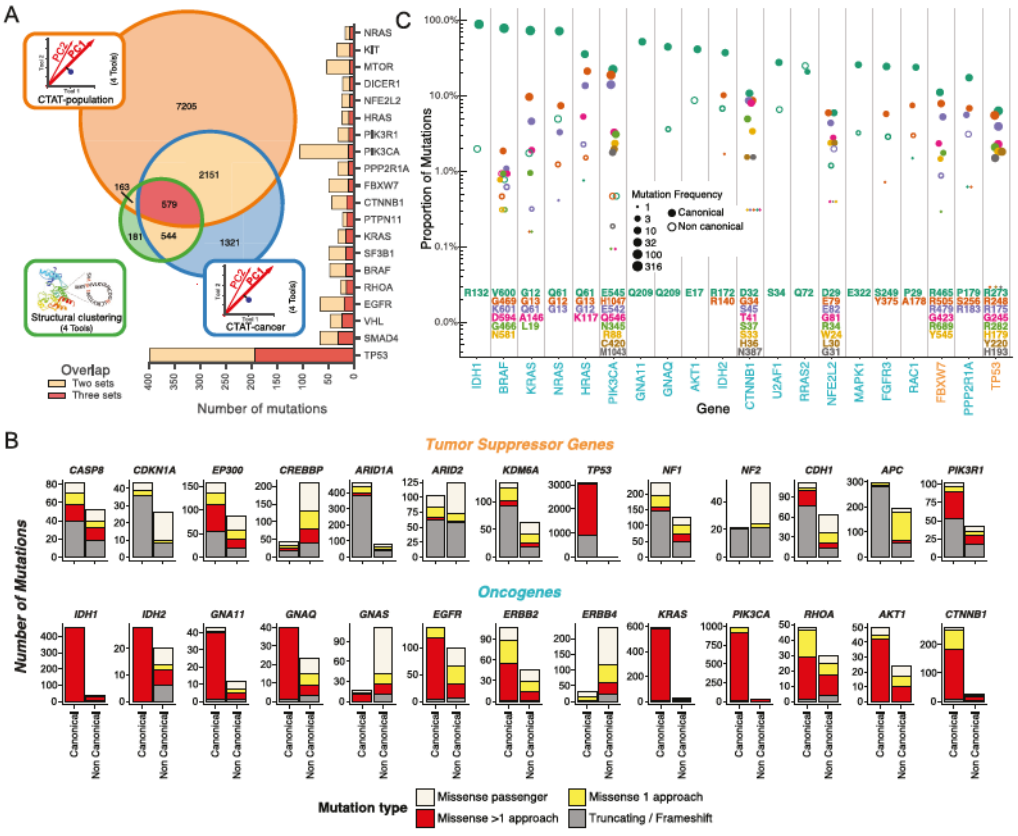

Figure 7.5: Driver mutation discovery approaches, overview, overlap, and contrasts: (A) Venn diagram indicates total number of mutations overlapping among three consensus approaches-CTAT-population, CTAT-cancer, and structural clustering. Adjacent bar chart indicates the top 20 genes sorted by 3-set intersecting mutation counts. (B) Driver gene discovery identified gene-tissue pairs (canonical genes) in tumor suppressors and oncogenes. However, some gene-tissue pairs were not identified in driver discovery (non-canonical). Mutation frequency from canonical and non-canonical cancer genes are displayed and divided among 4 mutation classes: truncation/frameshift mutations (grey); missense mutations uniquely identified by only one approach (yellow, see Panel A); missense mutations identified by multiple approaches (red, see Panel A); and missense passenger mutations not identified by any approach (off white). (C) Mutation percentage out of all missense and truncating/frameshift mutations within a gene is shown on the y-axis (log scale). Point size is log scaled and represents amino acid position frequency. The top 23 genes ordered by increasing mutational diversity (normalized entropy) and only the 9 most frequently mutated amino acid positions for each gene are shown.

outperformed all individual sequence-based approaches.

Overall, there are 9,919 predicted cancer driver mutations in our cohort (3,437 unique mutations) identified by 2 or more approaches from CTAT-population, CTAT-cancer, or structural clustering. These mutations affect 5,782 tumor samples. I observed that these missense driver mutations represent a greater fraction of the total mutations in oncogenes than in tumor suppressors (Figure 7.5B). In this latter group, most mutations seem to be truncating or frameshift, a result in agreement with previous observations [169]. Nevertheless, there are also tumor suppressor genes having high numbers of missense driver mutations, such as *EP300*, *CREBBP*, *CASP8*, *PIK3R1* and *TP53* (Figure 7.5B). An interesting example is *CDH1*, which is mostly affected by truncating or frameshift mutations in BRCA (75 out of 85 mutations), but mostly targeted by missense driver mutations in STAD (21 out of 25 mutations). This could suggest different roles for *CDH1* in these two cancer types.

I was also intrigued by missense driver mutations detected in cancer types where the gene was not predicted to be a driver. This subset is particularly important for genotype-driven clinical trials [170]. Overall, there are 1,719 of tissue-unmatched likely driver mutations (19% of the total) in 1,431 patients (16%) and there are 502 patients whose only predicted missense driver mutations affect genes not yet known to play a role in that cancer type. For example, I identified 28 patients with predicted EGFR driver mutations in can-

cer types where EGFR is not yet identified as a common driver gene, such as HNSC, STAD, LUSC, UCEC, ESCA and LIHC. In some extreme cases, such as *ERBB4* or *GNAS*, these mutations actually represent the majority of predicted driver missense mutations in the gene (Figure 7.5B). Additionally, 2% (10/457) of *IDH1* missense events that occur at the amino acid position R132 are found in tissues not typically known to carry such mutations i.e. BLCA (n=2), BRCA (2), COADREAD (2), LUAD (2), PCPG (1), and THYM (1) (Figure 7.5C). Furthermore, I observed that *RRAS2* Q72, a predicted oncogene in UCEC (n=5 samples) with strong homology to *KRAS* Q61 and *HRAS* Q61, was also mutated in cancer types in which it was not predicted to be an oncogene - UCS (n=1), LUSC (1), LUAD (1), PRAD (1), HNSC (1), and TCGT (1). Any analysis focusing only on common driver genes and mutations known in that cancer type would very likely miss presumed driver mutations for those patients. These results emphasize the advantage of PanCancer panels of driver mutations in order to maximize the coverage of driver-detection analyses.

7.2.4 Structure-guided discovery

Results were compared to an independent dataset of 1,049 experimentally tested somatic mutations to validate our driver mutation predictions [144]. Briefly, SNVs were introduced to two cancer cell lines, Ba/F3 and MCF10A, and were evaluated for their oncogenicity based on survival and growth. In total,

160 mutations from 19 genes were validated in this dataset. The percentage of functionally validated mutations increased from 60% predicted with CTAT-population, to 61% for those found by CTAT-cancer, and 78% for Structure-based analysis (Figure 7.6A). Among the 579 mutations predicted by all three approaches, 39 of the 46 that were tested (85%) were also validated. Further, the sensitivity and specificity of identifying driver mutations annotated by OncoKB suggests performance is generalizable to a larger set of genes. These results support the value of the prediction algorithms used in our study and the advantage of combining multiple tools. Also, I would like to note that this approach only addresses true positive findings and represents a floor estimate for computational predictions.

Structural-based mutations clustered on 66 proteins, including one cluster on KLF5, a gene not previously identified in PanCancer studies and ranked among the top 30 clusters by PanCancer mutation frequency (Figure 7.6B). I sought to examine in more detail the predictions of the three approaches in various well-established cancer driver genes, such as PIK3CA/PIK3R1, BRAF, and KEAP1/NFE2L2 (Figure 7.6C-4H). The interface between PIK3CA and PIK3R1 contains a cluster of mutations that were found by at least 2 of the approaches and includes both mutations that were validated and those not tested. D560G, N564D, and K567E are validated mutations that cluster closely to non-tested mutations R577P/Q, S565R, and P568T in PIK3R1. Similarly,

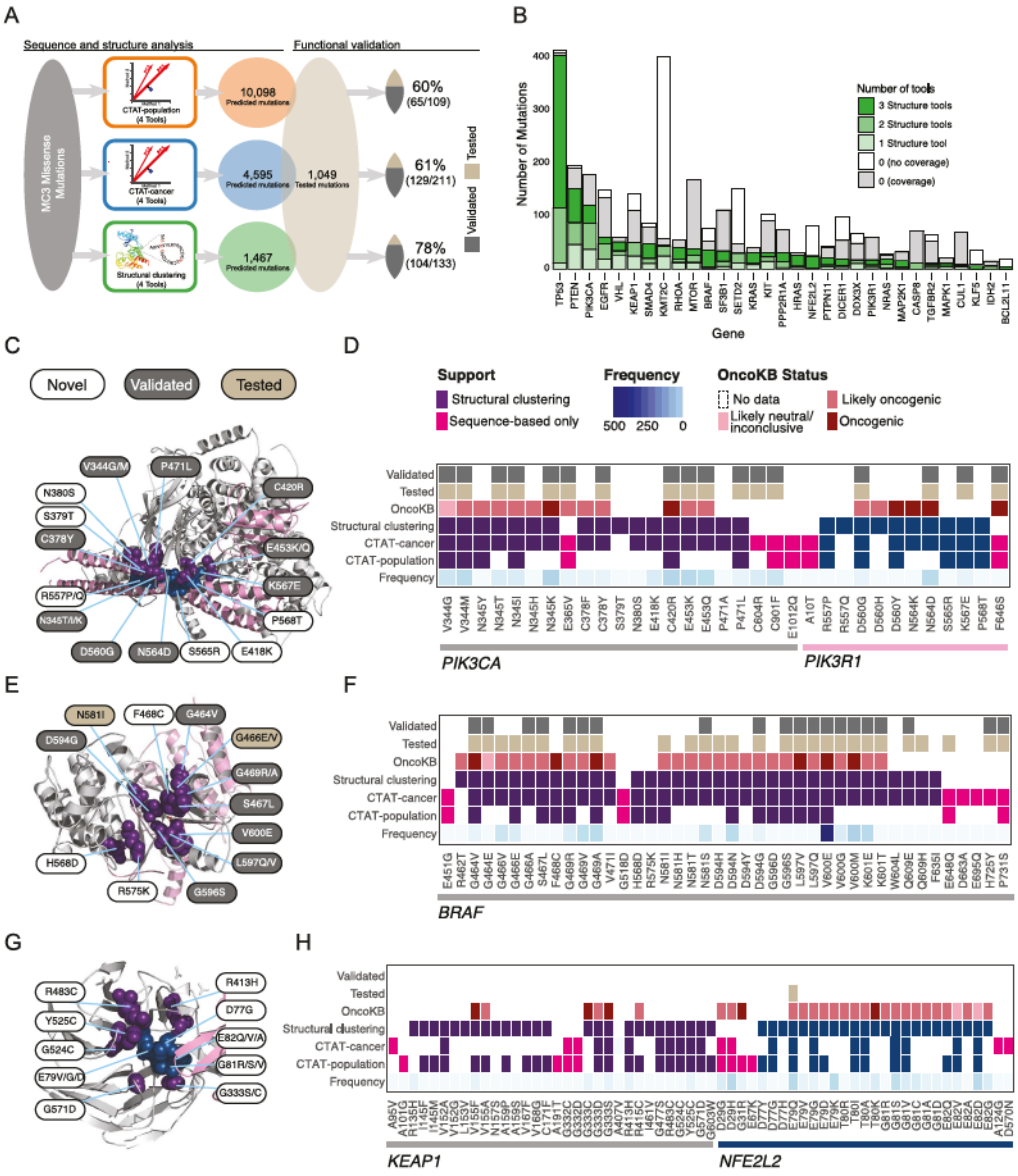

Figure 7.6: (Caption next page.)

Figure 7.6: (Figure: previous page) Driver mutation discovery and validation: (A) This schematic displays the steps taken to assess consensus among mutation-level predictions using sequence-based and structural clustering tools and comparing them to an orthogonal set of functionally validated mutations. From left to right: the grey box represents the missense mutations that were processed by 12 tools from 3 categories (population-based, cancer-focused, and structural clustering tools) and combined into three consensus approaches (CTAT-population, CTAT-cancer, and structural clustering). Finally, the total number and percentage of functionally validated/tested mutations is shown. (B) The number of mutations (y-axis) found by structural tools for each gene (x-axis) are shaded according to support by structural tools (green). Those mutations without support are distinguished by two categories, with (grey) and without (white) available protein structure. Heatmaps (D, F, H) coupled with protein structure (C, E, G) are shown in panels for the proteins PIK3CA/PIK3R1 (PDB ID: 4OVU), BRAF (4MBJ), and KEAP1/NFE2L2 (3ZGC), respectively, and display whether a particular mutation was detected by sequence-based (CTAT-population or CTAT-cancer) or structure-based approaches (at least two structural tools). Purple/teal colors distinguish proteins (PIK3CA/PIK3R1 and KEAP1/NFE2L2 pairs) for mutations found by structure-based approaches, while pink boxes indicate mutations found only by sequence-based approach. Additionally, for each mutation, frequency (blue gradient), OncoKB status (red gradient), testing status (tan), and validation status (grey) are provided. All mutations found by structure-based approaches in each of the 3 genes are shown with a few additional mutations that are only found by sequence-based approaches. Key mutations are highlighted from the heatmaps and labeled with white, grey, and tan labels referring to novel, validated, and tested (not validated) mutations, respectively.

PIK3CA contains validated mutations C378Y, V344G/M, N345T/I/K, P471L, C420R, and E418K clustering with non-tested mutations S379T, N380S, and E418K. These non-tested mutations are excellent candidates for further experimental validation due to their close proximity to known validated driver mutations as well as support from sequence-based approaches (Figure 7.6C and 7.6D). BRAF also contains clusters similar to this PIK3CA/PIK3R1 cluster, with a mixture of validated and novel mutations (Figure 7.6E and Figure 7.6F).

Additionally, there are many genes that contain mutations found by all three approaches but that were not tested experimentally, including *KEAP1*, *NFE2L2*, *RHOA*, *MTOR*, *MAP2K1*, and *VHL*. Nevertheless, many of these driver mutations have orthogonal evidence from OncoKB. For example, the mutations G333D/S in KEAP1 have an OncoKB status of likely oncogenic and oncogenic, respectively (Figure 7.6G and 7.6H). There are also NFE2L2 mutations that cluster closely with the KEAP1 mutations along the protein-protein interface (D77, E82, G81, E79) and were not experimentally validated but have an OncoKB status of either likely-oncogenic or oncogenic. Other KEAP1 mutations in the same cluster found by all three approaches are R483C, Y525C, G524C, G571D, and R413H. However, none of these mutations were tested in our dataset, nor have evidence from OncoKB. Given their proximity to the validated KEAP1 sites and the bioinformatic evidence that I found, these mutations are ideal candidates for follow-up validation experiments.

Overall, this analysis reinforces the notion that sequence-based approaches and structure-based approaches ought to be used in conjunction and tend to be complementary. For example, E365V, C604R, and C901F in PIK3CA, F646S in PIK3R1, and H725Y and P731S in BRAF were found by sequence-based approaches but not the structure-based approach and are experimentally validated (Figures 7.6D and 7.6F). Conversely, R462T in BRAF was only found through a structural approach and not sequence-based approaches and is an-

notated as likely oncogenic in OncoKB (Figures 7.6F and 7.6H). Finally, I note that, while looking at mutations detected by all 3 approaches provides high confident driver mutations, there may still be important driver mutations that were missed.

7.3 Discussion

As a lead analyst member, the driver's group has done a PanCancer and PanSoftware analysis on one of the largest available cancer genomics datasets at the moment, allowing us to identify 299 cancer driver genes. The gene list is limited in that the study focused on point mutations and small indels, but did not consider drivers affected by copy-number variations [171], genomic fusions [172], or methylation events [173]. Nevertheless, it represents the most comprehensive effort thus far to identify cancer driver genes and will serve the community as an important research asset well into the future.

Another important result is the dataset of 3,442 predicted driver mutations from both sequence-based and three-dimensional structure-based approaches. To emphasize a previous point, it is evident that not all mutations in driver genes are actually drivers themselves, so identifying the true-driver mutation subset will be a key challenge in the coming years. Also results were compared using an external independent experimental dataset to successfully validate

the predictions from three different approaches that predict cancer driver mutations. These results suggest that cancer-specific sequence-based approaches outperform those aimed at detecting pathogenic variants in general. Likewise, the structure-based approaches are more specific than sequence-based approaches at predicting driver mutations, but at a slight cost of reduced sensitivity. While functional validation confirmed true positive predictions, it gives no information regarding false negatives. Thus, what is reported here represents a lower bound. The assay is unable to capture other factors relevant to positive selection, such as tumor microenvironment, metastasis, or interactions with treatment or the immune system. While caution must be taken when extrapolating, these observations are consistent with other functional studies on individual proteins or a subset of the proteome that have shown that mutations affecting the same three-dimensional functional regions are likely to have similar phenotypes [174]. However, there were several instances in which sequence-based approaches captured driver mutations that were missed by structure-based approaches. Considering both approaches as complementary can improve prediction sensitivity.

The findings reported here and by the larger TCGA enterprise represent early steps toward a new era in cancer research and ultimately in cancer treatment. Studies will move beyond focusing on individual genes toward systematically integrating the myriad aspects of the cancer genome, including the

interrelationships among its somatic and germline variations [175], the tumor microenvironment and the immune system. Although this study represents the largest cancer gene and mutation study to date, the driver's group is mindful that the corpus of cancer driver genes and mutations may still be incomplete. However, it is likely that the community is nearing the beginning of the end of this phase of research, as larger cohorts continue to be examined with longer-range and longer-read sequencing technologies.

Chapter 8

Concluding remarks

The first confirmed human cancer driver gene, HRAS, was identified in 1982 [1, 7]. The past decade has seen the list of likely cancer driver genes grow rapidly. Although partly reflecting the growth in size of studies due to advances in next generation sequencing, it also reflects improvements in computational techniques. Computational methods are starting to become more robust with realistic models of how somatic mutations accumulate in cancer. Moreover, studies are now moving to understanding cancer drivers at increasing resolution – moving from genes to individual mutations.

The first part of my dissertation (Chapter 2) focused on how to appropriately statistically model the accumulation of somatic mutations in cancer. The typical choice of modeling the background mutation rate is problematic because it is highly variable at multiple scales. However, a key insight is that covariates

usually modulate mutation rate at the scale of megabases within the genome, but nearly all genes span <1MB. By statistically conditioning on the total number of mutations within a gene while simulating mutations, nuisance factors influencing mutation rate, which are not always measured or known, are substantially lessened. This approach allows substantial flexibility in comprehensively modeling the many mutational patterns indicative of positive selection in cancer.

Here, I introduced several new computational methods to analyze cancer drivers at different levels – such as the gene (20/20+, Chapter 3), region (HotMAPS, Chapter 5), and mutation (CHASMplus, Chapter 6). I used these methods to interrogate fundamental questions regarding cancer driver mutations, such as their cancer type specificity, commonness or rarity, the balance and characteristics of oncogenes and tumor suppressor genes, and the likely future trajectory of cancer driver discovery. 20/20+ identified that the balance of oncogenes and tumor suppressor genes (TSG) varies considerably by cancer type, some having all TSGs while others having mostly oncogenes. Also, CHASMplus found significantly more rare cancer driver mutations than previously understood, which is supportive of the long-tail hypothesis. The high prevalence of rare driver mutations suggests interpretation of a cancer genome will need to be increasingly personalized, since a patients driver mutation may have not been previously observed.

Due to the lack of a gold-standard for cancer drivers, I also developed a benchmark for cancer driver gene prediction (Chapter 4). This included five components: number of significant genes, overlap with previous literature, overlap with other methods, divergence of p-values from expectation, and consistency. I found that some methods did not accurately model the heterogeneity of accumulation of mutations by chance. As new computational methods are developed, it will be critical to effectively benchmark them against existing state-of-the-art.

I also used computational methods in an attempt to comprehensively discover driver genes and mutations in nearly 10,000 human cancers (Chapter 7). This was done as a consensus across institutions from The Cancer Genome Atlas, and by extension a consensus of computational methods. This idea of a consensus is not new, and has been used successfully in many other domains [176, 177]. The analysis revealed 299 cancer driver genes across 33 cancer types. Many of the cancer driver genes were cancer type specific, but others were found in multiple cancer types that had a common cell of origin. Analysis also found 3,400 unique missense mutations as likely cancer drivers, with high validation rates compared to an in vitro assay. The results from CHASMplus, however, suggest that in some cancer types the rate of discovery is starting to exhibit diminishing returns.

Although the landscape of cancer drivers in primary tumors are progres-

sively getting more fully explored, there remains many aspects that are still poorly understood. Sequencing untreated primary tumors gives an understanding of one time point in the natural evolutionary history of cancers. Understanding the full heterogeneity and dynamics of cancer will require sequencing both before (pre-cancerous lesions) and after (metastases). This will provide more understanding of such questions as: do cancers need particular gatekeeper drivers to initiate tumorigenesis?, and how much does the temporal ordering of driver mutations matter as opposed to overall driver mutation burden? In addition, the role of drivers in the non-coding region of the genome is only beginning to be characterized, but early studies indicate that >90% of driver point mutations may actually reside in coding regions [178]. The discrepancy of why other common diseases are estimated to have the majority of heritability in non-coding regions [179, 180] remains to be understood.

A complete catalog of all cancer drivers, in it of it self, will not suffice. Cancer driver mutations will need to be related to their functional consequence, interaction with the microenvironment, and to other driver mutations. Given the prevalence of rare driver mutations, the scale of experiments designed to functionally characterize or validate potential driver mutations will need to match pace. On the basis of this, a mechanistic insight will be critical for a rational understanding of the effects of targeted drugs and optimization of drug combinations. Lastly, driver mutations may not exert their effect in isolation,

but rather epistatically interact with other driver mutations. For instance, co-mutations of *KRAS*, *ATM*, *STK11*, and *KEAP1* in lung adenocarcinoma define a cancer subtype with different biology and immune signatures [181].

In summary, cancer was first understood as a genetic disease by observing large changes in chromosomes through a microscope [5]. Now, computational methods, like those developed here, are serving as a mathematical microscope into understanding driver mutations in cancer. The tools from statistics allow control of false discoveries, which prevents errant mistakes. While machine learning translates many biological features indicative of positive selection into concrete predictions of driver mutations. The convergence of this and large-scale cancer sequencing has turned many aspects of cancer research into a data science. As future research focuses on greater precision required for interpreting individual mutations in a patient's cancer, the trend of data science in cancer research will likely continue.

Appendix A

Glossary of Terms

cancer driver gene A gene that contains at least one cancer driver mutation.

cancer driver mutation A mutation that increases the net growth advantage of a cancer cell* under the specific microenvironmental context encounter *in vivo* in humans. *=cancer driver mutations may occur prior to a cell clone officially becoming cancer and has the theoretical possibility that the mutation is no longer advantageous for the cancer cell at a later time. However, a cancer driver mutation should be advantageous at least

at certain parts towards the development or progression of cancer.

cancer drivers A genetic loci that contains driver mutations, without reference specifically to whether it is a gene, region, or a mutation.

decision tree A decision tree is a set of questions asked recursively (in a tree-like manner) to predict either a categorical value (classification) or continuous value (regression).

driver mutation Shorthand for cancer driver mutation.

intra-tumor heterogeneity Heterogeneity among cells within a tumor. Here, specifically referring to different usage of cancer driver mutations.

microenvironment The local environment surrounding and within the tumor.

mutation hotspot A genomic or protein loci with highly localized mutations.

OG oncogene.

oncogene A cancer driver gene that is activated upon a driver mutation.

pan-cancer An analysis considering all or many types of cancer together.

random forest A supervised machine learning algorithm consisting of an ensemble of randomized decision trees.

somatic mutation starting from embryogenesis, mutations that occur in the cells of the body (excluding germ cells), and therefore are not inherited.

TSG tumor suppressor gene.

tumor suppressor gene A cancer driver gene that is inactivated upon a driver mutation.

Appendix B

Pan-cancer mutation dataset

The pancancer dataset consists of 729,205 small somatic variants encompassing 7,916 distinct samples from 34 specific cancer types by merging data in published whole-exome or whole-genome sequencing studies used by TUSON (Dataset) [89] and Mutsig (Tumor portal) [69] and removing duplicate samples in both studies. Any studies that did not report silent mutations were removed. Data in refs. [89] and [69] originated from The Cancer Genome Atlas, International Cancer Genome Consortium, the Catalogue of Somatic Mutations in Cancer database [112], and dbGAP. We did not see evidence of batch effects by data source in the number of variants per tumor type, single-nucleotide mutation spectra, or specific mutation consequence types. We further applied quality control to this data by filtering out hypermutated samples (¿1,000 intragenic small somatic variants) [14], and regions prone to mutation calling ar-

tifacts [any sequencing read mappability warning cataloged in the University of California, Santa Cruz (UCSC), Genome Browser]. The cleaned pancancer dataset is here. The CRAVAT webserver (version 3.0) was used to automatically retrieve the mappability warning codes. Gene names were standardized to HUGO Gene Nomenclature Committee through converting previous symbols and synonyms to the accepted gene name (downloaded January 29, 2015: here).

Bibliography

[1] L. F. Parada, C. J. Tabin, C. Shih, and R. A. Weinberg, "Human ej bladder carcinoma oncogene is homologue of harvey sarcoma virus ras gene," *Nature*, vol. 297, no. 5866, pp. 474–8, 1982. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/6283357

[2] 2017. [Online]. Available: https://seer.cancer.gov/statfacts/html/all.html

[3] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–74, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21376230

[4] D. M. Hyman, B. S. Taylor, and J. Baselga, "Implementing genome-driven oncology," *Cell*, vol. 168, no. 4, pp. 584–599, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28187282

[5] T. Boveri, "Concerning the origin of malignant tumours by theodor boveri. translated and annotated by henry harris," *Journal of Cell*

Science, vol. 121, no. Supplement 1, p. 1, 2008. [Online]. Available: http://jcs.biologists.org/content/121/Supplement_1/1.abstract

[6] D. Stehelin, H. E. Varmus, J. M. Bishop, and P. K. Vogt, "Dna related to the transforming gene(s) of avian sarcoma viruses is present in normal avian dna," *Nature*, vol. 260, no. 5547, pp. 170–3, 1976. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/176594

[7] C. J. Tabin, S. M. Bradley, C. I. Bargmann, R. A. Weinberg, A. G. Papageorge, E. M. Scolnick, R. Dhar, D. R. Lowy, and E. H. Chang, "Mechanism of activation of a human oncogene," *Nature*, vol. 300, no. 5888, pp. 143–9, 1982. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/6290897

[8] J. Knudson, A. G., "Mutation and cancer: statistical study of retinoblastoma," *Proc Natl Acad Sci U S A*, vol. 68, no. 4, pp. 820–3, 1971. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/5279523

[9] W. K. Cavenee, T. P. Dryja, R. A. Phillips, W. F. Benedict, R. Godbout, B. L. Gallie, A. L. Murphree, L. C. Strong, and R. L. White, "Expression of recessive alleles by chromosomal mechanisms in retinoblastoma," *Nature*, vol. 305, no. 5937, pp. 779–84, 1983. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/6633649

[10] S. J. Baker, E. R. Fearon, J. M. Nigro, S. R. Hamilton, A. C.

Preisinger, J. M. Jessup, P. vanTuinen, D. H. Ledbetter, D. F. Barker, Y. Nakamura, R. White, and B. Vogelstein, "Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas," *Science*, vol. 244, no. 4901, pp. 217–21, 1989. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2649981

[11] C. H. Yeh, M. Bellon, J. Pancewicz-Wojtkiewicz, and C. Nicot, "Oncogenic mutations in the fbxw7 gene of adult t-cell leukemia patients," *Proc Natl Acad Sci U S A*, vol. 113, no. 24, pp. 6731–6, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27247421

[12] A. Papa, L. Wan, M. Bonora, L. Salmena, M. S. Song, R. M. Hobbs, A. Lunardi, K. Webster, C. Ng, R. H. Newton, N. Knoblauch, J. Guarnerio, K. Ito, L. A. Turka, A. H. Beck, P. Pinton, R. T. Bronson, W. Wei, and P. P. Pandolfi, "Cancer-associated pten mutants act in a dominant-negative manner to suppress pten protein function," *Cell*, vol. 157, no. 3, pp. 595–610, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24766807

[13] M. Santarosa and A. Ashworth, "Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way," *Biochim Biophys Acta*, vol. 1654, no. 2, pp. 105–22, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15172699

[14] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, J. Diaz, L. A., and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–58, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23539594

[15] C. Tomasetti, L. Marchionni, M. A. Nowak, G. Parmigiani, and B. Vogelstein, "Only three driver gene mutations are required for the development of lung and colorectal cancers," *Proc Natl Acad Sci U S A*, vol. 112, no. 1, pp. 118–23, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25535351

[16] E. R. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell*, vol. 61, no. 5, pp. 759–67, 1990. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2188735

[17] N. Cancer Genome Atlas, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–7, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22810696

[18] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes,

K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y. E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M. H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, no. 7132, pp. 153–8, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17344846

[19] T. Sjoblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu, "The consensus coding sequences of human breast and colorectal cancers," *Science*, vol. 314, no. 5797, pp. 268–74, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16959974

[20] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary,

D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein, "The genomic landscapes of human breast and colorectal cancers," *Science*, vol. 318, no. 5853, pp. 1108–13, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17932254

[21] S. Jones, X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S. M. Hong, B. Fu, M. T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler, "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses," *Science*, vol. 321, no. 5897, pp. 1801–6, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18772397

[22] D. W. Parsons, S. Jones, X. Zhang, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, J. Diaz, L. A., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. Marie, S. M. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler, "An integrated genomic analysis of human glioblastoma multiforme," *Science*, vol. 321, no. 5897, pp. 1807–12, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18772396

[23] N. Cancer Genome Atlas Research, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–8, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18772890

[24] E. R. Mardis, L. Ding, D. J. Dooling, D. E. Larson, M. D. McLellan, K. Chen, D. C. Koboldt, R. S. Fulton, K. D. Delehaunty, S. D. McGrath, L. A. Fulton, D. P. Locke, V. J. Magrini, R. M. Abbott, T. L. Vickery, J. S. Reed, J. S. Robinson, T. Wylie, S. M. Smith, L. Carmichael, J. M. Eldred, C. C. Harris, J. Walker, J. B. Peck, F. Du, A. F. Dukes, G. E. Sanderson, A. M. Brummett, E. Clark, J. F. McMichael, R. J. Meyer, J. K. Schindler, C. S. Pohl, J. W. Wallis, X. Shi, L. Lin, H. Schmidt,

Y. Tang, C. Haipek, M. E. Wiechert, J. V. Ivy, J. Kalicki, G. Elliott, R. E. Ries, J. E. Payton, P. Westervelt, M. H. Tomasson, M. A. Watson, J. Baty, S. Heath, W. D. Shannon, R. Nagarajan, D. C. Link, M. J. Walter, T. A. Graubert, J. F. DiPersio, R. K. Wilson, and T. J. Ley, "Recurring mutations found by sequencing an acute myeloid leukemia genome," *N Engl J Med*, vol. 361, no. 11, pp. 1058–66, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19657110

[25] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M. L. Lin, G. R. Ordonez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, no. 7278, pp. 191–6, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20016485

[26] E. D. Pleasance, P. J. Stephens, S. O'Meara, D. J. McBride, A. Meynert,

D. Jones, M. L. Lin, D. Beare, K. W. Lau, C. Greenman, I. Varela, S. Nik-Zainal, H. R. Davies, G. R. Ordonez, L. J. Mudie, C. Latimer, S. Edkins, L. Stebbings, L. Chen, M. Jia, C. Leroy, J. Marshall, A. Menzies, A. Butler, J. W. Teague, J. Mangion, Y. A. Sun, S. F. McLaughlin, H. E. Peckham, E. F. Tsung, G. L. Costa, C. C. Lee, J. D. Minna, A. Gazdar, E. Birney, M. D. Rhodes, K. J. McKernan, M. R. Stratton, P. A. Futreal, and P. J. Campbell, "A small-cell lung cancer genome with complex signatures of tobacco exposure," *Nature*, vol. 463, no. 7278, pp. 184–90, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20016488

[27] N. Cancer Genome Atlas Research, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–15, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21720365

[28] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–8, 1976. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/959840

[29] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–13, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22258609

[30] E. Porta-Pardo, A. Kamburov, D. Tamborero, T. Pons, D. Grases,

A. Valencia, N. Lopez-Bigas, G. Getz, and A. Godzik, "Comparison of algorithms for the detection of cancer drivers at subgene resolution," *Nat Methods*, vol. 14, no. 8, pp. 782–788, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28714987

[31] L. Ding, M. C. Wendl, J. F. McMichael, and B. J. Raphael, "Expanding the computational toolbox for mining cancer genomes," *Nat Rev Genet*, vol. 15, no. 8, pp. 556–70, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25001846

[32] I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell, "Universal patterns of selection in cancer and somatic tissues," *Cell*, vol. 171, no. 5, pp. 1029–1041 e21, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29056346

[33] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca,

A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23770567

[34] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, I. Australian Pancreatic Cancer Genome, I. B. C. Consortium, I. M.-S. Consortium, I. PedBrain,

J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, pp. 415–21, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23945592

[35] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jonsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. Langerod, A. Tutt, J. W. Martens, S. A. Aparicio, A. Borg, A. V. Salomon, G. Thomas, A. L. Borresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, M. R. Stratton, and C. Breast Cancer Working Group of the International Cancer Genome, "Mutational processes molding the genomes of 21 breast cancers," *Cell*, vol. 149, no. 5, pp. 979–93, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22608084

[36] G. Parmigiani, S. Boca, J. Ding, and L. Trippa, "Statistical

tools and r software for cancer driver probabilities," *Methods Mol Biol*, vol. 1101, pp. 113–34, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24233780

[37] E. Hodis, I. R. Watson, G. V. Kryukov, S. T. Arold, M. Imielinski, J. P. Theurillat, E. Nickerson, D. Auclair, L. Li, C. Place, D. Dicara, A. H. Ramos, M. S. Lawrence, K. Cibulskis, A. Sivachenko, D. Voet, G. Saksena, N. Stransky, R. C. Onofrio, W. Winckler, K. Ardlie, N. Wagle, J. Wargo, K. Chong, D. L. Morton, K. Stemke-Hale, G. Chen, M. Noble, M. Meyerson, J. E. Ladbury, M. A. Davies, J. E. Gershenwald, S. N. Wagner, D. S. Hoon, D. Schadendorf, E. S. Lander, S. B. Gabriel, G. Getz, L. A. Garraway, and L. Chin, "A landscape of driver mutations in melanoma," *Cell*, vol. 150, no. 2, pp. 251–63, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22817889

[38] C. L. Araya, C. Cenik, J. A. Reuter, G. Kiss, V. S. Pande, M. P. Snyder, and W. J. Greenleaf, "Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations," *Nat Genet*, vol. 48, no. 2, pp. 117–25, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26691984

[39] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin, "Cancer-specific high-throughput

annotation of somatic mutations: computational prediction of driver missense mutations," *Cancer Res*, vol. 69, no. 16, pp. 6660–7, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19654296

[40] A. Gonzalez-Perez and N. Lopez-Bigas, "Functional impact bias reveals cancer drivers," *Nucleic Acids Res*, vol. 40, no. 21, p. e169, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22904074

[41] J. Gao, M. T. Chang, H. C. Johnsen, S. P. Gao, B. E. Sylvester, S. O. Sumer, H. Zhang, D. B. Solit, B. S. Taylor, N. Schultz, and C. Sander, "3d clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets," *Genome Med*, vol. 9, no. 1, p. 4, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28115009

[42] M. T. Chang, S. Asthana, S. P. Gao, B. H. Lee, J. S. Chapman, C. Kandoth, J. Gao, N. D. Socci, D. B. Solit, A. B. Olshen, N. Schultz, and B. S. Taylor, "Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity," *Nat Biotechnol*, vol. 34, no. 2, pp. 155–63, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26619011

[43] C. Tokheim, R. Bhattacharya, N. Niknafs, D. M. Gygax, R. Kim, M. Ryan, D. L. Masica, and R. Karchin, "Exome-scale discovery of hotspot mutation regions in human cancer using 3d protein structure,"

Cancer Res, vol. 76, no. 13, pp. 3719–31, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27197156

[44] T. Chen, Z. Wang, W. Zhou, Z. Chong, F. Meric-Bernstam, G. B. Mills, and K. Chen, "Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types," *BMC Genomics*, vol. 17 Suppl 2, p. 394, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27356755

[45] M. J. Meyer, R. Lapcevic, A. E. Romero, M. Yoon, J. Das, J. F. Beltran, M. Mort, P. D. Stenson, D. N. Cooper, A. Paccanaro, and H. Yu, "mutation3d: Cancer gene prediction through atomic clustering of coding variants in the structural proteome," *Hum Mutat*, vol. 37, no. 5, pp. 447–56, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26841357

[46] B. Niu, A. D. Scott, S. Sengupta, M. H. Bailey, P. Batra, J. Ning, M. A. Wyczalkowski, W. W. Liang, Q. Zhang, M. D. McLellan, S. Q. Sun, P. Tripathi, C. Lou, K. Ye, R. J. Mashl, J. Wallis, M. C. Wendl, F. Chen, and L. Ding, "Protein-structure-guided discovery of functional mutations across 19 cancer types," *Nat Genet*, vol. 48, no. 8, pp. 827–37, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27294619

[47] G. E. Melloni, S. de Pretis, L. Riva, M. Pelizzola, A. Ceol, J. Costanza,

H. Muller, and L. Zammataro, "Lowmaca: exploiting protein family analysis for the identification of rare driver mutations in cancer," *BMC Bioinformatics*, vol. 17, p. 80, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26860319

[48] E. Porta-Pardo, L. Garcia-Alonso, T. Hrabe, J. Dopazo, and A. Godzik, "A pan-cancer catalogue of cancer driver protein interaction interfaces," *PLoS Comput Biol*, vol. 11, no. 10, p. e1004518, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26485003

[49] F. Yang, E. Petsalaki, T. Rolland, D. E. Hill, M. Vidal, and F. P. Roth, "Protein domain-level landscape of cancer-type-specific somatic mutations," *PLoS Comput Biol*, vol. 11, no. 3, p. e1004147, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25794154

[50] M. L. Miller, E. Reznik, N. P. Gauthier, B. A. Aksoy, A. Korkut, J. Gao, G. Ciriello, N. Schultz, and C. Sander, "Pan-cancer analysis of mutation hotspots in protein domains," *Cell Syst*, vol. 1, no. 3, pp. 197–209, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27135912

[51] P. Jia, Q. Wang, Q. Chen, K. E. Hutchinson, W. Pao, and Z. Zhao, "Msea: detection and quantification of mutation hotspots through mutation set enrichment analysis," *Genome Biol*, vol. 15, no. 10, p. 489, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25348067

[52] G. A. Ryslik, Y. Cheng, K. H. Cheung, Y. Modis, and H. Zhao, "A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations," *BMC Bioinformatics*, vol. 15, p. 86, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24669769

[53] G. A. Ryslik, Y. Cheng, K. H. Cheung, R. D. Bjornson, D. Zelterman, Y. Modis, and H. Zhao, "A spatial simulation approach to account for protein structure when identifying non-random somatic mutations," *BMC Bioinformatics*, vol. 15, p. 231, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24990767

[54] G. A. Ryslik, Y. Cheng, K. H. Cheung, Y. Modis, and H. Zhao, "Utilizing protein structure to identify non-random somatic mutations," *BMC Bioinformatics*, vol. 14, p. 190, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23758891

[55] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes," *Bioinformatics*, vol. 29, no. 18, pp. 2238–44, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23884480

[56] C. M. Joyce and T. A. Steitz, "Function and structure relationships in dna polymerases," *Annu Rev Biochem*, vol. 63, pp. 777–822, 1994. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/7526780

[57] A. D. McLachlan, "Protein structure and function," *Annual Review of Physical Chemistry*, vol. 23, no. 1, pp. 165–192, 1972. [Online]. Available: https://doi.org/10.1146/annurev.pc.23.100172.001121

[58] A. Kamburov, M. S. Lawrence, P. Polak, I. Leshchiner, K. Lage, T. R. Golub, E. S. Lander, and G. Getz, "Comprehensive assessment of cancer missense mutation clustering in protein structures," *Proc Natl Acad Sci U S A*, vol. 112, no. 40, pp. E5486–95, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26392535

[59] J. C. Gonzalez-Sanchez, F. Raimondi, and R. B. Russell, "Cancer genetics meets biomolecular mechanism-bridging an age-old gulf," *FEBS Lett*, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29364530

[60] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C. L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N.

Thibodeau, A. S. Whittemore, and W. Sieh, "Revel: An ensemble method for predicting the pathogenicity of rare missense variants," *Am J Hum Genet*, vol. 99, no. 4, pp. 877–885, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27666373

[61] K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, and G. Bejerano, "M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity," *Nat Genet*, vol. 48, no. 12, pp. 1581–1586, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27776117

[62] R. D. Kumar, S. J. Swamidass, and R. Bose, "Unsupervised detection of cancer driver mutations with parsimony-guided learning," *Nat Genet*, vol. 48, no. 10, pp. 1288–94, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27618449

[63] Y. Mao, H. Chen, H. Liang, F. Meric-Bernstam, G. B. Mills, and K. Chen, "Candra: cancer-specific driver missense mutation annotation with optimized features," *PLoS One*, vol. 8, no. 10, p. e77945, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24205039

[64] H. A. Shihab, J. Gough, D. N. Cooper, I. N. Day, and T. R. Gaunt, "Predicting the functional consequences of cancer-associated amino acid

substitutions," *Bioinformatics*, vol. 29, no. 12, pp. 1504–10, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23620363

[65] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Res*, vol. 11, no. 5, pp. 863–74, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11337480

[66] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, pp. 333–339, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24132290

[67] B. J. Raphael, J. R. Dobson, L. Oesper, and F. Vandin, "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine," *Genome Med*, vol. 6, no. 1, p. 5, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24479672

[68] R. D. Kumar, A. C. Searleman, S. J. Swamidass, O. L. Griffith, and R. Bose, "Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data," *Bioinformatics*, vol. 31, no. 22, pp. 3561–8, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26209800

[69] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz, "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature*, vol. 505, no. 7484, pp. 495–501, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24390350

[70] D. Tamborero, A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, C. Kandoth, J. Reimand, M. S. Lawrence, G. Getz, G. D. Bader, L. Ding, and N. Lopez-Bigas, "Comprehensive identification of mutational cancer driver genes across 12 tumor types," *Sci Rep*, vol. 3, p. 2650, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24084849

[71] J. Reimand, O. Wagih, and G. D. Bader, "The mutational landscape of phosphorylation signaling in cancer," *Sci Rep*, vol. 3, p. 2651, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24089029

[72] J. Reimand and G. D. Bader, "Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers," *Mol Syst Biol*, vol. 9, p. 637, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23340843

[73] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, "Music: identifying mutational significance in

cancer genomes," *Genome Res*, vol. 22, no. 8, pp. 1589–98, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22759861

[74] L. A. Garraway and E. S. Lander, "Lessons from the cancer genome," *Cell*, vol. 153, no. 1, pp. 17–37, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23540688

[75] L. Ding, M. C. Wendl, D. C. Koboldt, and E. R. Mardis, "Analysis of next-generation genomic data in cancer: accomplishments and challenges," *Hum Mol Genet*, vol. 19, no. R2, pp. R188–96, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20843826

[76] W. Poole, K. Leinonen, I. Shmulevich, T. A. Knijnenburg, and B. Bernard, "Multiscale mutation clustering algorithm identifies pan-cancer mutational clusters associated with pathway-level changes in gene expression," *PLoS Comput Biol*, vol. 13, no. 2, p. e1005347, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28170390

[77] A. Blackford, O. K. Serrano, C. L. Wolfgang, G. Parmigiani, S. Jones, X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary, J. R. Eshleman, M. Goggins, E. M. Jaffee, C. A. Iacobuzio-Donahue, A. Maitra, J. L. Cameron, K. Olino, R. Schulick, J. Winter, J. M. Herman, D. Laheru, A. P. Klein, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, and R. H. Hruban, "Smad4 gene mutations are associated with poor prognosis in

pancreatic cancer," *Clin Cancer Res*, vol. 15, no. 14, pp. 4674–9, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19584151

[78] E. M. Van Allen, K. W. Mouw, P. Kim, G. Iyer, N. Wagle, H. Al-Ahmadie, C. Zhu, I. Ostrovnaya, G. V. Kryukov, K. W. O'Connor, J. Sfakianos, I. Garcia-Grossman, J. Kim, E. A. Guancial, R. Bambury, S. Bahl, N. Gupta, D. Farlow, A. Qu, S. Signoretti, J. A. Barletta, V. Reuter, J. Boehm, M. Lawrence, G. Getz, P. Kantoff, B. H. Bochner, T. K. Choueiri, D. F. Bajorin, D. B. Solit, S. Gabriel, A. D'Andrea, L. A. Garraway, and J. E. Rosenberg, "Somatic ercc2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma," *Cancer Discov*, vol. 4, no. 10, pp. 1140–53, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25096233

[79] J. G. Paez, P. A. Janne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F. J. Kaye, N. Lindeman, T. J. Boggon, K. Naoki, H. Sasaki, Y. Fujii, M. J. Eck, W. R. Sellers, B. E. Johnson, and M. Meyerson, "Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science*, vol. 304, no. 5676, pp. 1497–500, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15118125

[80] J. Phallen, M. Sausen, V. Adleff, A. Leal, C. Hruban, J. White, V. Anagnostou, J. Fiksel, S. Cristiano, E. Papp, S. Speir, T. Reinert,

M. W. Orntoft, B. D. Woodward, D. Murphy, S. Parpart-Li, D. Riley, M. Nesselbush, N. Sengamalay, A. Georgiadis, Q. K. Li, M. R. Madsen, F. V. Mortensen, J. Huiskens, C. Punt, N. van Grieken, R. Fijneman, G. Meijer, H. Husain, R. B. Scharpf, J. Diaz, L. A., S. Jones, S. Angiuoli, T. Orntoft, H. J. Nielsen, C. L. Andersen, and V. E. Velculescu, "Direct detection of early-stage cancers using circulating tumor dna," *Sci Transl Med*, vol. 9, no. 403, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28814544

[81] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–24, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19360079

[82] T. Helleday, S. Eshtad, and S. Nik-Zainal, "Mechanisms underlying mutational signatures in human cancers," *Nat Rev Genet*, vol. 15, no. 9, pp. 585–98, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24981601

[83] M. L. Hoang, C. H. Chen, V. S. Sidorenko, J. He, K. G. Dickman, B. H. Yun, M. Moriya, N. Niknafs, C. Douville, R. Karchin, R. J. Turesky, Y. S. Pu, B. Vogelstein, N. Papadopoulos, A. P. Grollman, K. W. Kinzler, and T. A. Rosenquist, "Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing," *Sci*

Transl Med, vol. 5, no. 197, p. 197ra102, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23926200

[84] C. R. Boland and A. Goel, "Microsatellite instability in colorectal cancer," *Gastroenterology*, vol. 138, no. 6, pp. 2073–2087 e3, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20420947

[85] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, "Evaluating the evaluation of cancer driver genes," *Proc Natl Acad Sci U S A*, vol. 113, no. 50, pp. 14 330–14 335, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27911828

[86] R. Sabarinathan, L. Mularoni, J. Deu-Pons, A. Gonzalez-Perez, and N. Lopez-Bigas, "Nucleotide excision repair is impaired by binding of transcription factors to dna," *Nature*, vol. 532, no. 7598, pp. 264–7, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27075101

[87] F. Supek and B. Lehner, "Differential dna mismatch repair underlies mutation rate variation across the human genome," *Nature*, vol. 521, no. 7550, pp. 81–4, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25707793

[88] B. Schuster-Bockler and B. Lehner, "Chromatin organization is a major influence on regional mutation rates in human cancer cells,"

Nature, vol. 488, no. 7412, pp. 504–7, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22820252

[89] T. Davoli, A. W. Xu, K. E. Mengwasser, L. M. Sack, J. C. Yoon, P. J. Park, and S. J. Elledge, "Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome," *Cell*, vol. 155, no. 4, pp. 948–62, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24183448

[90] M. R. Chernick and C. Y. Liu, "The saw-toothed behavior of power versus sample size and software solutions," *The American Statistician*, 2012.

[91] E. M. Rettig, J. Talbot, C. C., M. Sausen, S. Jones, J. A. Bishop, L. D. Wood, C. Tokheim, N. Niknafs, R. Karchin, E. J. Fertig, S. J. Wheelan, L. Marchionni, M. Considine, C. Fakhry, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, P. K. Ha, and N. Agrawal, "Whole-genome sequencing of salivary gland adenoid cystic carcinoma," *Cancer Prev Res (Phila)*, vol. 9, no. 4, pp. 265–74, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26862087

[92] G. Parmigiani, S. Boca, J. Lin, K. W. Kinzler, V. Velculescu, and B. Vogelstein, "Design and analysis issues in genome-wide somatic mutation studies of cancer," *Genomics*, vol. 93, no. 1, pp. 17–21, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18692126

[93] B. A. Logsdon, A. J. Gentles, C. P. Miller, C. A. Blau, P. S. Becker, and S. I. Lee, "Sparse expression bases in cancer reveal tumor drivers," *Nucleic Acids Res*, vol. 43, no. 3, pp. 1332–44, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25583238

[94] D. Bertrand, K. R. Chng, F. G. Sherbaf, A. Kiesel, B. K. Chia, Y. Y. Sia, S. K. Huang, D. S. Hoon, E. T. Liu, A. Hillmer, and N. Nagarajan, "Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles," *Nucleic Acids Res*, vol. 43, no. 7, p. e44, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25572314

[95] J. P. Hou and J. Ma, "Dawnrank: discovering personalized driver genes in cancer," *Genome Med*, vol. 6, no. 7, p. 56, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25177370

[96] D. Tamborero, N. Lopez-Bigas, and A. Gonzalez-Perez, "Oncodrive-cis: a method to reveal likely driver genes based on the impact of their copy number changes on expression," *PLoS One*, vol. 8, no. 2, p. e55489, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23408991

[97] S. Ng, E. A. Collisson, A. Sokolov, T. Goldstein, A. Gonzalez-Perez, N. Lopez-Bigas, C. Benz, D. Haussler, and J. M. Stuart, "Paradigm-shift predicts the function of mutations in multiple cancers using pathway

impact analysis," *Bioinformatics*, vol. 28, no. 18, pp. i640–i646, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22962493

[98] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah, "Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer," *Genome Biol*, vol. 13, no. 12, p. R124, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23383675

[99] L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, and N. Lopez-Bigas, "Oncodrivefml: a general framework to identify coding and non-coding regions with cancer driver mutations," *Genome Biol*, vol. 17, no. 1, p. 128, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27311963

[100] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C. Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, and P. J. Campbell, "Cosmic: somatic cancer genetics at high-resolution," *Nucleic Acids Res*, vol. 45, no. D1, pp. D777–D783, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27899578

[101] M. P. Schroeder, C. Rubio-Perez, D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "Oncodriverole classifies cancer driver genes

in loss of function and activating mode of action," *Bioinformatics*, vol. 30, no. 17, pp. i549–55, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25161246

[102] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[103] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, 1997. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.7.1545

[104] L. Breiman, "Classification and regression trees," 1984.

[105] W. C. Wong, D. Kim, H. Carter, M. Diekhans, M. C. Ryan, and R. Karchin, "Chasm and snvbox: toolkit for detecting biologically important single nucleotide mutations in cancer," *Bioinformatics*, vol. 27, no. 15, pp. 2147–8, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21685053

[106] H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, and R. Karchin, "Identifying mendelian disease genes with the variant effect scoring tool," *BMC Genomics*, vol. 14 Suppl 3, p. S3, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23819870

[107] G. E. A. P. A. Batista, A. L. C. Bazzan, and M. C. Monard, "Balancing training data for automated annotation of keywords: a case study," *In Proceedings of the Second Brazilian Workshop on Bioinformatics*, pp. 35–43, 2003.

[108] pp. 935–942, 2007. [Online]. Available: http://delivery.acm.org/10.1145/1280000/1273614/p935-van_hulse.pdf?ip=162.129.251.87&id=1273614&acc=ACTIVE%20SERVICE&key=7777116298C9657D%2E34B115928DB6308C%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=456042147&CFTOKEN=38356524&__acm__=1417364818_4cb984aae735ff480317aec64c3d0b1d

[109] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: http://www.jstor.org/stable/2346101

[110] D. Cozzetto, A. Kryshtafovych, and A. Tramontano, "Evaluation of casp8 model quality predictions," *Proteins*, vol. 77 Suppl 9, pp. 157–66, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19714774

[111] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learn-

ing applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[112] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nat Rev Cancer*, vol. 4, no. 3, pp. 177–83, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14993899

[113] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proc Natl Acad Sci U S A*, vol. 100, no. 16, pp. 9440–5, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12883005

[114] S. D. Horn, "Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale," *Biometrics*, vol. 33, no. 1, pp. 237–247, 1977. [Online]. Available: http://www.jstor.org/stable/2529319

[115] C. Bernau, M. Riester, A. L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa, "Cross-study validation for the assessment of prediction algorithms," *Bioinformatics*, vol. 30, no. 12, pp. i105–12, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24931973

[116] M. Hollstein, D. Sidransky, B. Vogelstein, and C. C. Harris, "p53 mutations in human cancers," *Science*, vol. 253, no. 5015, pp. 49–53, 1991. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/1905840

[117] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas, "Intogen-mutations identifies cancer drivers across tumor types," *Nat Methods*, vol. 10, no. 11, pp. 1081–2, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24037244

[118] M. Daszykowski and B. Walczak, "Density-based clustering methods," *Comprehensive Chemometrics*, vol. 2, pp. 635–654, 2010.

[119] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–6, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17218491

[120] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1025, 2007.

[121] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, pp. 849–856, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.8100&rep=rep1&type=pdf

[122] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953. [Online]. Available: https://doi.org/10.1007/BF02289263

[123] U. Pieper, N. Eswar, B. M. Webb, D. Eramian, L. Kelly, D. T. Barkan, H. Carter, P. Mankoo, R. Karchin, M. A. Marti-Renom, F. P. Davis, and A. Sali, "Modbase, a database of annotated comparative protein structure models and associated resources," *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D347–54, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18948282

[124] N. Niknafs, D. Kim, R. Kim, M. Diekhans, M. Ryan, P. D. Stenson, D. N. Cooper, and R. Karchin, "Mupit interactive: webserver for mapping variant positions to annotated, interactive 3d structures," *Hum Genet*, vol. 132, no. 11, pp. 1235–43, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23793516

[125] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The universal protein resource (uniprot): an expanding universe of protein information," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D187–91, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16381842

[126] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a

new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–402, 1997. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9254694

[127] W. J. Kent, "Blat–the blast-like alignment tool," *Genome Res*, vol. 12, no. 4, pp. 656–64, 2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11932250

[128] M. J. Davis, B. H. Ha, E. C. Holman, R. Halaban, J. Schlessinger, and T. J. Boggon, "Rac1p29s is a spontaneously activating cancer-associated gtpase," *Proc Natl Acad Sci U S A*, vol. 110, no. 3, pp. 912–7, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23284172

[129] G. S. Winkler, S. J. Araujo, U. Fiedler, W. Vermeulen, F. Coin, J. M. Egly, J. H. Hoeijmakers, R. D. Wood, H. T. Timmers, and G. Weeda, "Tfiih with inactive xpd helicase functions in transcription initiation but is defective in dna repair," *J Biol Chem*, vol. 275, no. 6, pp. 4258–66, 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10660593

[130] J. O. Lee, H. Yang, M. M. Georgescu, A. Di Cristofano, T. Maehama, Y. Shi, J. E. Dixon, P. Pandolfi, and N. P. Pavletich, "Crystal structure of the pten tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association," *Cell*, vol. 99, no. 3,

pp. 323–34, 1999. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10555148

[131] A. M. Rojas, G. Fuentes, A. Rausell, and A. Valencia, "The ras protein superfamily: evolutionary tree and role of conserved amino acids," *J Cell Biol*, vol. 196, no. 2, pp. 189–201, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22270915

[132] B. Zhang, Y. Zhang, Z. Wang, and Y. Zheng, "The role of mg2+ cofactor in the guanine nucleotide exchange and gtp hydrolysis reactions of rho family gtp-binding proteins," *J Biol Chem*, vol. 275, no. 33, pp. 25 299–307, 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10843989

[133] L. Gossage, T. Eisen, and E. R. Maher, "Vhl, the story of a tumour suppressor gene," *Nat Rev Cancer*, vol. 15, no. 1, pp. 55–64, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25533676

[134] S. Nikolaev, F. Santoni, M. Garieri, P. Makrythanasis, E. Falconnet, M. Guipponi, A. Vannier, I. Radovanovic, F. Bena, F. Forestier, K. Schaller, V. Dutoit, V. Clement-Schatlo, P. Y. Dietrich, and S. E. Antonarakis, "Extrachromosomal driver mutations in glioblastoma and low-grade glioma," *Nat Commun*, vol. 5, p. 5690, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25471132

[135] L. G. Martelotto, C. K. Ng, M. R. De Filippo, Y. Zhang, S. Piscuoglio, R. S. Lim, R. Shen, L. Norton, J. S. Reis-Filho, and B. Weigelt, "Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations," *Genome Biol*, vol. 15, no. 10, p. 484, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25348012

[136] M. A. Molina-Vila, N. Nabau-Moreto, C. Tornador, A. J. Sabnis, R. Rosell, X. Estivill, T. G. Bivona, and C. Marino-Buslje, "Activating mutations cluster in the "molecular brake" regions of protein kinases and do not associate with conserved or catalytic residues," *Hum Mutat*, vol. 35, no. 3, pp. 318–28, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24323975

[137] E. Capriotti and R. B. Altman, "A new disease-specific machine learning approach for the prediction of cancer-causing missense variants," *Genomics*, vol. 98, no. 4, pp. 310–7, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21763417

[138] M. Meyer, J. F. Beltrn, S. Liang, R. Fragoza, A. Rumack, J. Liang, X. Wei, and H. Yu, "Interactome insider: A multi-scale structural interactome browser for genomic studies," *bioRxiv*, 2017. [Online]. Available: http://biorxiv.org/content/early/2017/04/12/126862.abstract

[139] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nat Genet*, vol. 46, no. 3, pp. 310–5, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24487276

[140] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Res*, vol. 39, no. 17, p. e118, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21727090

[141] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nat Methods*, vol. 7, no. 4, pp. 248–9, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20354512

[142] A. Gonzalez-Perez, J. Deu-Pons, and N. Lopez-Bigas, "Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation," *Genome Med*, vol. 4, no. 11, p. 89, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23181723

[143] A. P. Bradley, "The use of the area under the roc curve in the evaluation

of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[144] P. K.-S. Ng, J. Li, K. J. Jeong, S. Shao, H. Chen, Y. H. Tsang, S. Sengupta, Z. Wang, V. H. Bhavana, R. Tran, S. Soewito, D. C. Minussi, D. Moreno, K. Kong, T. Dogruluk, H. Lu, J. Gao, C. Tokheim, D. C. Zhou, J. Zeng, C. K. M. Ip, Z. Ju, M. Wester, S. Yu, Y. Li, C. Vellano, N. Schultz, R. Karchin, L. Ding, Y. Lu, L. W. T. Cheung, K. Chen, K. L. Scott, S. Yi, N. Sahni, H. Liang, and G. B. Mills, "Systematic functional annotation of somatic mutations in cancers," *in review*.

[145] A. H. Berger, A. N. Brooks, X. Wu, Y. Shrestha, C. Chouinard, F. Piccioni, M. Bagul, A. Kamburov, M. Imielinski, L. Hogstrom, C. Zhu, X. Yang, S. Pantel, R. Sakai, J. Watson, N. Kaplan, J. D. Campbell, S. Singh, D. E. Root, R. Narayan, T. Natoli, D. L. Lahr, I. Tirosh, P. Tamayo, G. Getz, B. Wong, J. Doench, A. Subramanian, T. R. Golub, M. Meyerson, and J. S. Boehm, "High-throughput phenotyping of lung cancer somatic mutations," *Cancer Cell*, vol. 30, no. 2, pp. 214–228, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27478040

[146] A. Petitjean, E. Mathe, S. Kato, C. Ishioka, S. V. Tavtigian, P. Hainaut, and M. Olivier, "Impact of mutant p53 functional properties on tp53 mutation patterns and tumor phenotype: lessons from

recent developments in the iarc tp53 database," *Hum Mutat*, vol. 28, no. 6, pp. 622–9, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17311302

[147] E. Kim, N. Ilic, Y. Shrestha, L. Zou, A. Kamburov, C. Zhu, X. Yang, R. Lubonja, N. Tran, C. Nguyen, M. S. Lawrence, F. Piccioni, M. Bagul, J. G. Doench, C. R. Chouinard, X. Wu, L. Hogstrom, T. Natoli, P. Tamayo, H. Horn, S. M. Corsello, K. Lage, D. E. Root, A. Subramanian, T. R. Golub, G. Getz, J. S. Boehm, and W. C. Hahn, "Systematic functional interrogation of rare cancer variants identifies oncogenic alleles," *Cancer Discov*, vol. 6, no. 7, pp. 714–26, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27147599

[148] A. Zehir, R. Benayed, R. H. Shah, A. Syed, S. Middha, H. R. Kim, P. Srinivasan, J. Gao, D. Chakravarty, S. M. Devlin, M. D. Hellmann, D. A. Barron, A. M. Schram, M. Hameed, S. Dogan, D. S. Ross, J. F. Hechtman, D. F. DeLair, J. Yao, D. L. Mandelker, D. T. Cheng, R. Chandramohan, A. S. Mohanty, R. N. Ptashkin, G. Jayakumaran, M. Prasad, M. H. Syed, A. B. Rema, Z. Y. Liu, K. Nafa, L. Borsu, J. Sadowska, J. Casanova, R. Bacares, I. J. Kiecka, A. Razumova, J. B. Son, L. Stewart, T. Baldi, K. A. Mullaney, H. Al-Ahmadie, E. Vakiani, A. A. Abeshouse, A. V. Penson, P. Jonsson, N. Camacho, M. T. Chang,

H. H. Won, B. E. Gross, R. Kundra, Z. J. Heins, H. W. Chen, S. Phillips, H. Zhang, J. Wang, A. Ochoa, J. Wills, M. Eubank, S. B. Thomas, S. M. Gardos, D. N. Reales, J. Galle, R. Durany, R. Cambria, W. Abida, A. Cercek, D. R. Feldman, M. M. Gounder, A. A. Hakimi, J. J. Harding, G. Iyer, Y. Y. Janjigian, E. J. Jordan, C. M. Kelly, M. A. Lowery, L. G. T. Morris, A. M. Omuro, N. Raj, P. Razavi, A. N. Shoushtari, N. Shukla, T. E. Soumerai, A. M. Varghese, R. Yaeger, J. Coleman, B. Bochner, G. J. Riely, L. B. Saltz, H. I. Scher, P. J. Sabbatini, M. E. Robson, D. S. Klimstra, B. S. Taylor, J. Baselga, N. Schultz, D. M. Hyman, M. E. Arcila, D. B. Solit, M. Ladanyi, and M. F. Berger, "Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients," *Nat Med*, vol. 23, no. 6, pp. 703–713, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28481359

[149] D. Chakravarty, J. Gao, S. M. Phillips, R. Kundra, H. Zhang, J. Wang, J. E. Rudolph, R. Yaeger, T. Soumerai, M. H. Nissan, M. T. Chang, S. Chandarlapaty, T. A. Traina, P. K. Paik, A. L. Ho, F. M. Hantash, A. Grupe, S. S. Baxi, M. K. Callahan, A. Snyder, P. Chi, D. Danila, M. Gounder, J. J. Harding, M. D. Hellmann, G. Iyer, Y. Janjigian, T. Kaley, D. A. Levine, M. Lowery, A. Omuro, M. A. Postow, D. Rathkopf, A. N. Shoushtari, N. Shukla, M. Voss, E. Paraiso, A. Zehir, M. F. Berger, B. S. Taylor, L. B. Saltz, G. J. Riely, M. Ladanyi, D. M. Hyman,

J. Baselga, P. Sabbatini, D. B. Solit, and N. Schultz, "Oncokb: A precision oncology knowledge base," *JCO Precis Oncol*, vol. 2017, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28890946

[150] J. Ye, A. Pavlicek, E. A. Lunney, P. A. Rejto, and C. H. Teng, "Statistical method on nonrandom clustering with application to somatic mutations in cancer," *BMC Bioinformatics*, vol. 11, p. 11, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20053295

[151] B. Meszaros, A. Zeke, A. Remenyi, I. Simon, and Z. Dosztanyi, "Systematic analysis of somatic mutations driving cancer: uncovering functional protein regions in disease development," *Biol Direct*, vol. 11, p. 23, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27150584

[152] E. Porta-Pardo and A. Godzik, "e-driver: a novel method to identify protein regions driving cancer," *Bioinformatics*, vol. 30, no. 21, pp. 3109–14, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25064568

[153] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, B. J. Raphael, D. S. Marks, B. F. F. Ouellette, A. Valencia, G. D. Bader, P. C. Boutros, J. M. Stuart, R. Linding, N. Lopez-Bigas, L. D.

Stein, C. Mutation, and C. Pathway Analysis Working Group of the International Cancer Genome, "Pathway and network analysis of cancer genomes," *Nat Methods*, vol. 12, no. 7, pp. 615–621, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26125594

[154] A. Gonzalez-Perez, V. Mustonen, B. Reva, G. R. Ritchie, P. Creixell, R. Karchin, M. Vazquez, J. L. Fink, K. S. Kassahn, J. V. Pearson, G. D. Bader, P. C. Boutros, L. Muthuswamy, B. F. Ouellette, J. Reimand, R. Linding, T. Shibata, A. Valencia, A. Butler, S. Dronov, P. Flicek, N. B. Shannon, H. Carter, L. Ding, C. Sander, J. M. Stuart, L. D. Stein, N. Lopez-Bigas, P. International Cancer Genome Consortium Mutation, and G. Consequences Subgroup of the Bioinformatics Analyses Working, "Computational approaches to identify functional genetic variants in cancer genomes," *Nat Methods*, vol. 10, no. 8, pp. 723–9, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23900255

[155] K. Ellrott, M. H. Bailey, G. Saksena, K. R. Covington, C. Kandoth, C. Stewart, M. McLellan, H. J. Sofia, C. Hutter, G. Getz, D. Wheeler, L. Ding, T. M. W. Group, and T. C. G. A. R. Network, "Automating somatic mutation calling for ten thousand tumor exomes," *in review*.

[156] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, T. Fennell, A. O'Donnell-Luria, J. Ware, A. Hill, and B. Cummings, "Analysis of

protein-coding genetic variation in 60,706 humans," *BioRxiv*, p. 030338, 2016.

[157] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer," *Nature reviews Genetics*, vol. 14, no. 10, pp. 703–718, 2013.

[158] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: an information aesthetic for comparative genomics," *Genome research*, vol. 19, no. 9, pp. 1639–1645, 2009.

[159] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, and D. C. Wedge, "Landscape of somatic mutations in 560 breast cancer whole-genome sequences," *Nature*, vol. 534, no. 7605, pp. 47–54, 2016.

[160] K. Schulze, S. Imbeaud, E. Letouz, L. B. Alexandrov, J. Calderaro, S. Rebouissou, G. Couchy, C. Meiller, J. Shinde, and F. Soysouvanh, "Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets," *Nature genetics*, vol. 47, no. 5, pp. 505–511, 2015.

[161] C. E. Barbieri, S. C. Baca, M. S. Lawrence, F. Demichelis, M. Blattner, J.-P. Theurillat, T. A. White, P. Stojanov, E. Van Allen, and N. Stransky,

"Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer," *Nature genetics*, vol. 44, no. 6, pp. 685–689, 2012.

[162] A. V. Biankin, N. Waddell, K. S. Kassahn, M.-C. Gingras, L. B. Muthuswamy, A. L. Johns, D. K. Miller, P. J. Wilson, A.-M. Patch, and J. Wu, "Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes," *Nature*, vol. 491, no. 7424, pp. 399–405, 2012.

[163] P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, and G. R. Bignell, "The landscape of cancer genes and mutational processes in breast cancer," *Nature*, vol. 486, no. 7403, pp. 400–404, 2012.

[164] C. C. Pritchard, S. J. Salipante, K. Koehler, C. Smith, S. Scroggins, B. Wood, D. Wu, M. K. Lee, S. Dintzis, and A. Adey, "Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens," *The Journal of Molecular Diagnostics*, vol. 16, no. 1, pp. 56–67, 2014.

[165] G. M. Frampton, A. Fichtenholtz, G. A. Otto, K. Wang, S. R. Downing, J. He, M. Schnall-Levin, J. White, E. M. Sanford, and P. An, "Development and validation of a clinical cancer genomic profiling test based on

massively parallel dna sequencing," *Nature biotechnology*, vol. 31, no. 11, pp. 1023–1031, 2013.

[166] I. Ionita-Laza, K. McCallum, B. Xu, and J. BUXBAUM, "A spectral approach integrating functional genomic annotations for coding and noncoding variants," *Nature genetics*, vol. 48, no. 2, p. 214, 2016.

[167] A. Torkamani and N. J. Schork, "Prediction of cancer driver mutations in protein kinases," *Cancer Res*, vol. 68, no. 6, pp. 1675–82, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18339846

[168] N. Mirkovic, M. A. Marti-Renom, B. L. Weber, A. Sali, and A. N. Monteiro, "Structure-based assessment of missense mutations in human brca1," *Cancer research*, vol. 64, no. 11, pp. 3790–3797, 2004.

[169] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature medicine*, vol. 10, no. 8, p. 789, 2004.

[170] J. Gagan and E. M. Van Allen, "Next-generation sequencing to guide cancer therapy," *Genome medicine*, vol. 7, no. 1, p. 80, 2015.

[171] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhsng, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim, "Pan-cancer patterns of somatic copy

number alteration," *Nat Genet*, vol. 45, no. 10, pp. 1134–40, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24071852

[172] K. Yoshihara, Q. Wang, W. Torres-Garcia, S. Zheng, R. Vegesna, H. Kim, and R. G. Verhaak, "The landscape and therapeutic relevance of cancer-associated transcript fusions," *Oncogene*, vol. 34, no. 37, pp. 4845–54, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25500544

[173] D. D. De Carvalho, S. Sharma, J. S. You, S. F. Su, P. C. Taberlay, T. K. Kelly, X. Yang, G. Liang, and P. A. Jones, "Dna methylation screening identifies driver epigenetic events of cancer cell survival," *Cancer Cell*, vol. 21, no. 5, pp. 655–67, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22624715

[174] L. Brenan, A. Andreev, O. Cohen, S. Pantel, A. Kamburov, D. Cacchiarelli, N. S. Persky, C. Zhu, M. Bagul, E. M. Goetz, A. B. Burgin, L. A. Garraway, G. Getz, T. S. Mikkelsen, F. Piccioni, D. E. Root, and C. M. Johannessen, "Phenotypic characterization of a comprehensive set of mapk1/erk2 missense mutants," *Cell Rep*, vol. 17, no. 4, pp. 1171–1183, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27760319

[175] H. Carter, R. Marty, M. Hofree, A. M. Gross, J. Jensen, K. M. Fisch,

X. Wu, C. DeBoever, E. L. Van Nostrand, Y. Song, E. Wheeler, J. F. Kreisberg, S. M. Lippman, G. W. Yeo, J. S. Gutkind, and T. Ideker, "Interaction landscape of inherited polymorphisms with somatic events in cancer," *Cancer Discov*, vol. 7, no. 4, pp. 410–423, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28188128

[176] D. Marbach, J. C. Costello, R. Kuffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, D. Consortium, M. Kellis, J. J. Collins, and G. Stolovitzky, "Wisdom of crowds for robust gene network inference," *Nat Methods*, vol. 9, no. 8, pp. 796–804, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22796662

[177] J. M. Murphy, D. M. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, "Quantification of modelling uncertainties in a large ensemble of climate change simulations," *Nature*, vol. 430, no. 7001, pp. 768–72, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15306806

[178] E. Rheinbay, M. M. Nielsen, F. Abascal, G. Tiao, H. Hornshj, J. M. Hess, R. I. I. Pedersen, L. Feuerbach, R. Sabarinathan, H. T. Madsen, J. Kim, L. Mularoni, S. Shuai, A. A. Lanzs Camaioni, C. Herrmann, Y. E. Maruvka, C. Shen, S. B. Amin, J. Bertl, P. Dhingra, K. Diamanti, A. Gonzalez-Perez, Q. Guo, N. J. Haradhvala, K. Isaev, M. Juul,

J. Komorowski, s. kumar, D. Lee, L. Lochovsky, E. M. M. Liu, O. Pich, d. tamborero, H. M. Umer, L. Uuskla-Reimand, C. Wadelius, L. Wadi, J. Zhang, K. A. Boroevich, J. Carlevaro-Fita, D. Chakravarty, C. W. Y. Y. W. Chan, N. A. Fonseca, M. P. Hamilton, C. Hong, A. Kahles, Y. Kim, K.-V. Lehmann, T. A. A. Johnson, A. Kahraman, K. Park, G. Saksena, L. Sieverling, N. A. Sinnott-Armstrong, P. J. Campbell, A. Hobolth, M. Kellis, M. S. Lawrence, B. Raphael, M. A. Rubin, C. Sander, L. Stein, J. Stuart, T. Tsunoda, D. A. Wheeler, R. Johnson, J. Reimand, M. B. Gerstein, E. Khurana, N. Lopez-Bigas, I. Martincorena, J. S. S. Pedersen, and G. Getz, "Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes," *bioRxiv*, 2017. [Online]. Available: http://biorxiv.org/content/early/2017/12/23/237313.abstract

[179] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos, "Systematic localization of common disease-associated variation in regulatory dna," *Science*, vol. 337, no. 6099, pp. 1190–5, 2012. [Online].

Available: https://www.ncbi.nlm.nih.gov/pubmed/22955828

[180] A. Gusev, S. H. Lee, G. Trynka, H. Finucane, B. J. Vilhjalmsson, H. Xu, C. Zang, S. Ripke, B. Bulik-Sullivan, E. Stahl, C. Schizophrenia Working Group of the Psychiatric Genomics, S.-S. Consortium, A. K. Kahler, C. M. Hultman, S. M. Purcell, S. A. McCarroll, M. Daly, B. Pasaniuc, P. F. Sullivan, B. M. Neale, N. R. Wray, S. Raychaudhuri, A. L. Price, C. Schizophrenia Working Group of the Psychiatric Genomics, and S.-S. Consortium, "Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases," *Am J Hum Genet*, vol. 95, no. 5, pp. 535–52, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25439723

[181] F. Skoulidis, L. A. Byers, L. Diao, V. A. Papadimitrakopoulou, P. Tong, J. Izzo, C. Behrens, H. Kadara, E. R. Parra, J. R. Canales, J. Zhang, U. Giri, J. Gudikote, M. A. Cortez, C. Yang, Y. Fan, M. Peyton, L. Girard, K. R. Coombes, C. Toniatti, T. P. Heffernan, M. Choi, G. M. Frampton, V. Miller, J. N. Weinstein, R. S. Herbst, K. K. Wong, J. Zhang, P. Sharma, G. B. Mills, W. K. Hong, J. D. Minna, J. P. Allison, A. Futreal, J. Wang, I. Wistuba, and J. V. Heymach, "Co-occurring genomic alterations define major subsets of kras-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities,"

Cancer Discov, vol. 5, no. 8, pp. 860–77, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26069186

Vita

Collin Tokheim
116 W University Pkwy, APT 1226
Baltimore, MD, 21210
ctokheim@jhu.edu

EDUCATION

Johns Hopkins University
Ph.D., Biomedical Engineering, September 2018
GPA 4.0/4.0

University of Iowa
BSE, Biomedical Engineering, 2011
GPA 4.0/4.0

RESEARCH EXPERIENCE

Graduate research assistant, Johns Hopkins University (2014-2018)
Advisor: Dr. Rachel Karchin, Biomedical Engineering and Oncology
Topic: Statistical identification of mutational drivers in human cancers through development of cutting-edge computational methods

Staff research associate, University of California, Los Angeles (1/2013-6/2013)
Advisor: Dr. Yi Xing, Microbiology, Immunology, & Molecular Genetics
Topic: Genomic analysis of the transcriptome with emphasis on alternative splicing

Research Assistant, University of Iowa (4/2011-12/2012)

Advisor: Dr. Yi Xing, Microbiology, Immunology, & Molecular Genetics
Topic: Genomic analysis of the transcriptome with emphasis on alternative splicing

Undergraduate Research Assistant, University of Iowa (4/2010-8/2011)
Advisor: Dr. Xiaodong Wu, Electrical and Computer Engineering and Radiation Oncology
Topic: Computational analysis of computed tomography (CT) scans of tumors.

DISTINCTIONS, AWARDS, & HONORS

- Martin and Carol Macht Award (2018)
- First author PNAS paper featured in Baltimore Sun newspaper (Jan 2017)
- National Cancer Institute (NCI) NRSA Fellowship F31CA200266 (2015-present)
- Alpha Eta Mu Beta Biomedical Engineering Honor Society (2016-present)
- Top performing team in the Personal Genome Project challenge in Critical Assessment of Genome Interpretation (CAGI) competition (Mar 2016)
- Graduated with Highest Distinction (May 2011)
- Tau Beta Pi Engineering Honor Society (2010-present)
- William C. Blackburn Engineering Scholarship (2009-2011)
- University of Iowa President's List (2008-2011)
- E. Lester and Frances M. Williams Fund Scholarship (2008-2009)
- College of Engineering Dean's List (2007-2011)
- Engineering Excellence Scholarship (2007-2008)

PEER-REVIEWED PUBLICATIONS

(Google Scholar H-index = 7, total citations = 295)
* co-first author

Matthew H. Bailey*, Collin Tokheim*, Eduard Porta-Pardo*, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C. Wendl, Jaegil Kim, Brendan Reardon, 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 Cortes-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 Van Allen, Han Liang, The MC

Working Group, The Cancer Genome Atlas Research Network, Michael Lawrence, Adam Godzik, Nuria Lopez-Bigas, Josh Stuart, David Wheeler, Gad Getz, Ken Chen, Alexander J. Lazar, Gordon B Mills, Rachel Karchin, Li Ding. Comprehensive Discovery and Characterization of Driver Genes and Mutations in Cancer. Cell. In Press.

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A 113, 14330-14335, doi:10.1073/pnas.1616440113 (2016).

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D. M., Kim, R., Ryan, M., Masica, D. L. & Karchin, R. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res 76, 3719-3731, doi:10.1158/0008-5472.CAN-15-3190 (2016).

Bertotti, A., Papp, E., Jones, S., Adleff, V., Anagnostou, V., Lupo, B., Sausen, M., Phallen, J., Hruban, C. A., Tokheim, C., Niknafs, N., Nesselbush, M., Lytle, K., Sassi, F., Cottino, F., Migliardi, G., Zanella, E. R., Ribero, D., Russolillo, N., Mellano, A., Muratore, A., Paraluppi, G., Salizzoni, M., Marsoni, S., Kragh, M., Lantto, J., Cassingena, A., Li, Q. K., Karchin, R., Scharpf, R., Sartore-Bianchi, A., Siena, S., Diaz, L. A., Jr., Trusolino, L. & Velculescu, V. E. The genomic landscape of response to EGFR blockade in colorectal cancer. Nature 526, 263-267, doi:10.1038/nature14969 (2015).

Patrick Kwok-Shing Ng, J. L., Kang Jin Jeong, Shan Shao, Hu Chen, Yiu Huen Tsang, Sohini Sengupta, Zixing Wang, Venkata Hemanjani Bhavana, Richard Tran, Stephanie Soewito, Darlan Conterno Minussi, Daniela Moreno, Kathleen Kong, Turgut Dogruluk, Hengyu Lu, Jianjiong Gao, Collin Tokheim, Daniel Cui Zhou, Jia Zeng, Carman Ka Man Ip, Zhenlin Ju, Matthew Wester, Shuangxing Yu, Yongsheng Li, Christopher Vellano, Nikolaus Schultz, Rachel Karchin, Li Ding, Yiling Lu, Lydia Wai Ting Cheung, Ken Chen, Kenna R. Shaw, Funda Meric-Bernstam, Kenneth L. Scott, Song Yi, Nidhi Sahni, Han Liang, and Gordon B. Mills. Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell. In press.

Chung, C. H., Guthrie, V. B., Masica, D. L., Tokheim, C., Kang, H., Richmon, J., Agrawal, N., Fakhry, C., Quon, H., Subramaniam, R. M., Zuo, Z., Seiwert, T., Chalmers, Z. R., Frampton, G. M., Ali, S. M., Yelensky, R., Stephens, P. J., Miller, V. A., Karchin, R. & Bishop, J. A. Genomic alterations in head and neck squamous cell carcinoma determined by cancer gene-targeted sequencing. Ann Oncol 26, 1216-1223, doi:10.1093/annonc/mdv109 (2015).

Li Ding, M. H. B., Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L. Gibbs, Amila Weerasinghe, Kuan-lin Huang, Matthew A Wyczalkowski, Sam Q. Sun, Collin Tokheim, Reyka Jayasinghe, Feng Chen, Lihua Yu, Alison M Taylor, Andrew D Cherniack, Jaegil Kim, Rehan Akbani, Chayaporn Suphavilai, Niranjan Nagarajan, Joshua M. Stuart, Gordon B Mills, Benjamin Vincent, Carolyn M Hutter, Jean Claude Zenklusen, Katherine A Hoadley, Michael C. Wendl, llya Shmulevich, Alexander J. Lazar, David Wheeler, Gad Getz, The Cancer Genome Atlas Research Network. Perspective on Central Oncogenic Processes at the End of the Beginning of Cancer Genomics. [Submitted]

Tokheim, C., Park, J. W. & Xing, Y. PrimerSeq: Design and visualization of RT-PCR primers for alternative splicing using RNA-seq data. Genomics Proteomics Bioinformatics 12, 105-109, doi:10.1016/j.gpb.2014.04.001 (2014).

Masica, D. L., Douville, C., Tokheim, C., Bhattacharya, R., Kim, R., Moad, K., Ryan, M. C. & Karchin, R. CRAVAT 4: Cancer-Related Analysis of Variants Toolkit. Cancer Res 77, e35-e38, doi:10.1158/0008-5472.CAN-17-0338 (2017).

Park, J. W., Tokheim, C., Shen, S. & Xing, Y. Identifying differential alternative splicing events from RNA sequencing data using RNASeq-MATS. Methods Mol Biol 1038, 171-179, doi:10.1007/978-1-62703-514-9_10 (2013).

Wang, J., Lu, Z. X., Tokheim, C. J., Miller, S. E. & Xing, Y. Species-specific exon loss in human transcriptomes. Mol Biol Evol 32, 481-494, doi:10.1093/molbev/msu317 (2015).

Lu, Z. X., Huang, Q., Park, J. W., Shen, S., Lin, L., Tokheim, C. J., Henry, M. D. & Xing, Y. Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization. Mol Cancer Res 13, 305-318, doi:10.1158/1541-7786.MCR-14-0366 (2015).

Rettig, E. M., Talbot, C. C., Jr., Sausen, M., Jones, S., Bishop, J. A., Wood, L. D., Tokheim, C., Niknafs, N., Karchin, R., Fertig, E. J., Wheelan, S. J., Marchionni, L., Considine, M., Fakhry, C., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Ha, P. K. & Agrawal, N. Whole-Genome Sequencing of Salivary Gland Adenoid Cystic Carcinoma. Cancer Prev Res (Phila) 9, 265-274, doi:10.1158/1940-6207.CAPR-15-0316 (2016).

Cai, B., Li, B., Kiga, N., Thusberg, J., Bergquist, T., Chen, Y. C., Niknafs, N., Carter, H., Tokheim, C., Beleva-Guthrie, V., Douville, C., Bhattacharya, R., Yeo, H. T. G., Fan, J., Sengupta, S., Kim, D., Cline, M., Turner, T., Diekhans, M.,

Zaucha, J., Pal, L. R., Cao, C., Yu, C. H., Yin, Y., Carraro, M., Giollo, M., Ferrari, C., Leonardi, E., Tosatto, S. C. E., Bobe, J., Ball, M., Hoskins, R. A., Repo, S., Church, G., Brenner, S. E., Moult, J., Gough, J., Stanke, M., Karchin, R. & Mooney, S. D. Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges. Hum Mutat 38, 1266-1276, doi:10.1002/humu.23265 (2017).

RESEARCH PRESENTATIONS & CONFERENCES

Posters

- American Society of Human Genetics Annual Meeting (10/2016). Title: "CRAVAT 4.3: informatics tools for high-throughput analysis of exome variants"
- GRC SNPs & Disease (6/2016). Title: "HotMAPS & MuPiT: Somatic missense mutation clustering in protein structures aids cancer driver discovery and interpretation"
- Critical Assessment of Genome Interpretation conference (3/2016). Title: "Genome Interpretation using Bayesian Inference: Application to Matching Anonymized Genomes to Phenotypic Profiles"

Conferences & meetings

- TCGA PanCanAtlas meeting, Houston (11/2016)
- American Society of Human Genetics Annual Meeting, Vancouver (10/2016)
- Critical Assessment of Genome Interpretation conference, San Francisco (3/2016)
- TCGA PanCanAtlas meeting, Santa Cruz (2/2016)
- American Association of Cancer Research Annual Meeting, Philadelphia (4/2015)
- Biology of Genomes Conference, Cold Spring Harbor (5/2014)

TEACHING EXPERIENCE

Teaching assistant, Foundations of Computational Biology and Bioinformatics II (Spring 2017)
Teaching assistant, Systems Bioengineering 3 (Fall 2016)