Spring 5-15-2018

Discerning Drivers of Cancer: Computational Approaches to Somatic Exome Sequencing Data

Runjun Kumar
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Ron Bose, Chair
Donald F. Conrad
Li Ding
Obi L. Griffith
Daniel C. Link
S. Joshua Swamidass

Discerning Drivers of Cancer:
Computational Approaches to Somatic Exome Sequencing Data

by
Runjun D. Kumar

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2018
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I want to thank Dr. Ron Bose, who is a tremendous mentor. I look to him as a role model of a physician scientist, and I admire his ability to balance the responsibilities of so many domains. However, the quality which I hope to emulate most is his infectious enthusiasm. Science is a tough enterprise, but his passion for his patients, the experiments, and possibly even grant writing, is a quiet gift he offers freely to everyone he works with.

Many Bose Lab members have become trusted and reliable friends. It has been a privilege to witness their work and families grow and mature over the past few years. Lab lunches with Kwabena Sarpong, Tim Collier, Ted Keppel, Naveen Jain, Wei Shen, Elisa Murray, Ari Gao, Edward Stites and Vandna Kukshal were always a great way to end a week, but there are three people who warrant particular thanks. John Monsey led me through experiments, sharing his experience generously. Dr. Shyam Kavuri taught me tissue culture and actually inspired this dissertation. Finally, Dr. Adam Searleman gave me much technical and informal guidance as an MD/PhD student, from clinical rotations to committee selection. They are all great friends.

My committee has been a blessing. Drs. Kristen Naegle, Don Conrad and Li Ding have all encouraged and coached me in equal parts. Dr. Dan Link is the best committee chair anyone could ask for. And I want to give Drs. Joshua Swamidass and Obi Griffith special thanks. Besides being committee members, they have become dependable and thoughtful collaborators.

The Canadian Institutes of Health Research support me with a Doctoral Foreign Study Award (DFS 134-967). I am also supported by the Washington University MD/PhD program. Not many programs can fund international students, and I’m grateful that ours does.

Our administrators, like Brian Sullivan, Christy Durbin, Elizabeth Bayer and Linda Perniciaro in the MSTP office, and Jeanne Silvestrini in the DBBS office, make sure that our program runs smoothly. I know I have caused them all a few headaches, and I thank them for their patience.

I also want to thank my previous research mentors, Drs. Peer Bork, Paul de Bakker and William Navarre, all of whom continue to support me. The same is true of Drs. Doug Templeton and Richard Hegele, who were in charge of my undergraduate programs and have provided much career guidance.

I thank my amazing colleagues in the MD/PhD program, who have been a huge source of inspiration. I had no idea that so many important milestones would come and go, and that by working, graduating, marrying and starting families together (in both senses), we would become such an amazing community. But we have, and I'm grateful for that.

Finally, I want to offer thanks to the special people in my life. The people I talk to when I have problems that are really pretty silly, but who love me too much to just tell me so. There are so many of them, but my mom and dad deserve a special mention. I was insanely lucky to be born to such inspiring and supportive people, who are such positive role models in every way. My relationship with them has proven to be an inexhaustible source of comfort and joy for 27 years, and I think it will continue to be so for a very long time to come.

Runjun D. Kumar

Washington University in St. Louis

May 2018

ABSTRACT OF THE DISSERTATION

Discerning Drivers of Cancer:

Computational Approaches to Somatic Exome Sequencing Data

by

Runjun D. Kumar

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2018

Associate Professor Ron Bose, Chair

Paired tumor-normal sequencing of thousands of patient's exomes has revealed millions of somatic mutations, but functional characterization and clinical decision making are stymied because biologically neutral 'passenger' mutations greatly outnumber pathogenic 'driver' mutations. Since most mutations will return negative results if tested, conventional resource-intensive experiments are reserved for mutations which are observed in multiple patients or rarer mutations found in well-established cancer genes. Most mutations are therefore never tested, diminishing the potential to discover new mechanisms of cancer development and treatment opportunities. Computational methods that reliably prioritize mutations for testing would greatly increase the translation of sequencing results to clinical care. The goal of this thesis is to develop new approaches that use datasets of protein-coding somatic mutations to identify putative cancer-causing genes and mutations, and to validate these predictions *in silico* and experimentally. This effort will be split among several inter-related efforts, which taken together will help experimental biologists and clinicians focus on hypotheses that can yield novel insights into cancer biology, development, and treatment.

1. Introduction

The genetic nature of cancer has been appreciated almost since the beginning of the genetics era: Theodore Boveri suggested that chromosomes were the unit of heredity and carried cancer-causing factors at the turn of the 20th century[1]. However, it was not until the 1970s that the first descriptions of transforming proto-oncogenes were published[2]. The first tumor suppressors were described in the following decade, beginning with the gene RB1[3]. With these genetic underpinnings well established, tumors became some of the first biological samples to be whole-genome and whole-exome sequenced.

Ley *et al.* set the template for cancer genome sequencing studies when they published the first complete tumor genome, an acute myeloid leukemia[4]. In the following years, numerous cancer genomes were published, many under the auspices of The Cancer Genome Atlas, a major collaborative effort funded by the NIH that had a goal of sequencing 10,000 cancer genomes. These studies each encompassed dozens or hundreds of patients and focused on cancers including glioblastoma, ovarian carcinoma, breast cancers, colorectal cancers, endometrial carcinomas, squamous cell lung tumors, and many others[5-26]. Beginning in 2010-2011, the arrival of inexpensive exome-capture (in which a DNA sample is enriched for the 1% of the genome that is exonic) allowed independent groups to publish substantial cancer mutation datasets as well, initially focusing on lung and breast cancers[27-30]. For groups interested in analyses that encompass all cancer types, the literature now encompasses thousands of patients and millions of somatic mutations[31-34].

1.1 Cancer Genome Sequencing

We describe the process of sequencing and analyzing cancer genomes briefly, though reviews are available[35-37]. We focus on the process of exome-sequencing and identifying protein-coding variants, since these are the data used throughout this dissertation.

The process of tumor sequencing begins when patients are biopsied at both the tumor site and a matched normal tissue (usually blood or skin). Once samples are gathered and DNA is extracted, libraries are constructed by shearing DNA and extracting short fragments (often a few hundred base-pairs long). Multiple vendors sell exome capture kits, which can be used to limit the library to protein-coding exons. Libraries can then be sequenced on any of a variety of instruments, most of which sequence multiple molecules in parallel and make use of "sequencing-by-synthesis" design. Once sequencing reads have been quality controlled, they can be aligned to the reference human genome using various short-read alignment algorithms.

The next step is critical: by comparing the aligned exomes, one from a patient's healthy normal tissue and one from the tumor, mutations that are present only in the tumor can be identified. These are the somatic mutations that this dissertation depends on. There are many somatic variant callers that can be applied to this problem[38, 39], but these tools frequently make discordant variant calls[35].

In fact, one of the largest caveats to this dissertation is the extent to which we depend on mutations identified and published by multiple sequencing centers. These groups may use different analytic pipelines - including different exon capture protocols, sequencing instruments, quality controls, alignment algorithms and variant callers - all of which can introduce both false positives and false negatives into published sets of observed somatic mutations. Rather than try

to identify these errors *post-hoc*, we strive to produce rigorous analytic techniques that are robust to systematic errors that may be present in published datasets. However, it must be acknowledged that the results of even the most carefully planned analysis are only as reliable as the data used for design and validation.

1.2 Converting Mutations to Treatments

Even with these caveats, the study of these mutations offers multiple areas of discovery, including the description of new diagnostic and prognostic markers, the discovery of new drug targets and improved use of existing drugs, and the discovery of previously undescribed cancer genes. However, realizing these benefits and improving treatments for patients will require that mutations be functionally characterized using targeted experimental approaches. For instance, our group identified several mutations that activate HER2 and drive tumor development in multiple cell and animal models of breast cancer[40], and followed those studies by extending the results into models of colorectal cancer[41]. As of writing, there is mounting evidence that patients with these mutations can be treated successfully with anti-HER2 targeted drugs[42]. HER2 is an exceptionally well-studied oncogene (reviewed in [43]), with rigorously validated model systems and reagents available to researchers who are interested in studying it. Despite this developed research infrastructure, it took our group several years and thousands of man-hours to translate these observed mutations into improved treatments for patients. The time and cost will be considerably greater for mutations that occur in genes that have historically been less studied, or that occur in cancers with fewer available model systems.

Even so, the cost and effort of characterizing observed somatic mutations would likely be worthwhile if most studied mutations led to improved patient care. Unfortunately that is not the case. Only a small proportion of observed mutations are believed to underlie tumorigenesis

("drivers"), although the exact proportion remains unknown[44]. Genome instability is a characteristic feature of most tumors, and cancers vary widely in terms of overall mutation rate - as much as four orders of magnitude, increasing from AML to melanoma[45]. Particularly in tumors with high background mutation rates, it is likely that only a small fraction of mutations act as drivers, with the rest being incidental to disease development ("passengers").

These passenger mutations greatly stimy the necessary functional assessment of genome sequencing results. Given the cost of experiments, and the low proportion of mutations that act as drivers, testing a given mutation poses a high risk, low reward opportunity for most investigators. There are two exceptions. The first is for mutations which occur in multiple patients (e.g. R882C/H/P in newly established cancer gene DNMT3A[46]). These "recurrent" mutations are much more likely to act as drivers, justifying the cost of experiments. The other exception is when mutations are within well-established cancer genes (e.g. V777L in HER2[40]). Because these genes generally have well-developed research infrastructures, the cost of characterizing mutations is considerably less than if the mutations were in a less-familiar gene. Given these considerations, driver mutations which occur in fewer numbers or less studied genes are unlikely to ever be tested.

This situation can be described as a bottleneck at the interface of high-throughput, low cost hypothesis generation and low-throughput, high cost hypothesis testing. The goal of this dissertation is to mitigate the bottleneck by improving the prioritization of genes and mutations for functional characterization. If prioritized mutations and genes can be functionally tested with a high or even moderate success rate, it is much more likely that even rare mutations will be tested. As such, we design our analytic methods with the goals of providing highly precise and biologically relevent predictions. In this dissertation, we develop several new, interrelated

methods that use exomic somatic mutation datasets to identify putative cancer-causing genes and mutations, and we validate these predictions using both *in silico* and *in vitro* methods. Each of these methods can be used as a filter. When assembled together, they can be used to reduce large sets of mutations to only the most promising hypotheses.

The three approaches we explore are 1) the use of the somatic mutations to identify cancer genes that possess non-random sets of mutations and likely experience selection during tumor development, 2) the use of unsupervised approaches for making functional impact predictions for individual mutations, and 3) identifying functional regions of proteins by considering the patterns of somatic mutations in aligned gene families. The remainder of this chapter explores each of these topics briefly, while chapters 2-4 discuss each topic and our results in-depth. The pan-cancer dataset that is used across all three efforts is described in chapter 2.

1.3 Specific Approaches

1.3.1 Identifying Cancer Genes

The term “cancer genes” encompasses both tumor suppressors (TSGs), which exert a pro-tumor effect through a loss-of-function, and oncogenes, which exert a pro-tumor effect through a gain-of-function. Many cancer genes were identified prior to genetic sequencing, and mutations within them can be prioritized without additional information (for instance, the aforementioned activating mutations in HER2). However, new cancer genes can also be identified through observed patterns of somatic mutations[45]. That is, we can use somatic mutation data to identify genes that appear to be under selection, possessing non-random mutation patterns. Then mutations in these putative cancer genes can be prioritized for further testing. Taking a gene-centric approach to the problem of identifying putative drivers is very appealing, because many

elements of experimental design are dictated by the target gene, making this approach critical to translating mutation data to new biological knowledge.

Several tools have been developed to prioritize genes whose mutations are likely nonrandom. The best known methods rely on mutation significance[45, 47], though newer methods also rely on other signals of selection, including functional impact scores[48], intra-gene mutation clustering and recurrence[49], post-translational modifications[50], and DNA lesion likelihood[51]. Earlier work also demonstrated the importance of co-mutation events[52] and patient-specific mutation rates in detecting cancer genes[53]. The goal is to identify the small subset of genes (and mutations) which are crucial to cancer progression.

Three challenges exist for the field of *in silico* cancer gene discovery. First, while individual methods are well-designed, it is clear that combining complementary methods (which rely on detecting different signals of selection) would improve detection of cancer genes overall[54, 55]. For instance, Tamborero *et al*. showed that genes identified by multiple methods are more likely to be found in the Cancer Gene Census, the best available list of predicted and known cancer genes; however, the authors were unable to produce a true combination classifier that models interactions between different tools to make predictions[55]. A second shortcoming of these methods is that they treat cancer genes as a single class and do not attempt to separate oncogenes and TSGs. This is particularly undesirable in the case of oncogenes, which are of great interest since they can be targeted by small-molecule inhibitors. However, recent studies demonstrate how rates of truncating mutations, mutation clustering, and copy-number data can be used to separate oncogenes and TSGs[56]. The final challenge is a lack of external validation. Since no panel of *bona fide* cancer genes exists, studies must rely on simulation or inter-method comparisons to gauge performance. Neither is ideal. Simulation studies are highly dependent on

the assumptions used to produced simulated ata, while comparisons between methods are ambiguous (methods either agree and reinforce one another, or they are complementary and produce novel insights).

Although there is great promise in cancer gene prediction as a way of facilitating functional studies, the shortcomings outlined above limit the overall impact of the field. Fortunately, each of these shortcomings can be addressed though supervised modeling, in which a statistical model is trained to separate experimentally validated “gold panel” cancer genes from other genes (for a description of supervised modeling in general, see [57]). The model can make its predictions based on combinations of existing tools or newly developed gene descriptors, allowing it to make better predictions than individual tools can in isolation. Moreover, such a model could be trained to identify oncogenes and tumor suppressors as separate groups, providing even more context to its predictions, and possibly further improving performance over methods which treat all cancer genes as a single group. Through careful design, the same gold panel can be used to both train and assess the performance of the model, and also to compare its performance to other tools. The ultimate result will be a set of putative oncogenes and a set of putative tumor suppressors. The mutations observed within these genes would then be the focus of further studies. Chapter 2 explores these possibilities in-depth.

1.3.2 Predicting Mutation Functional Impact

Even if cancer genes can be reliably identified, only a subset of the mutations within them likely act as drivers. Predicting the effects of protein-coding mutations is a complex but well-established problem. Several statistical functional impact scores (FIS) are in use, each aiming to predict whether a given amino acid change is neutral or functional with regard to protein function. Methods such as SIFT use a conservation-based approach, in which evolutionarily

conserved residues are assumed to be critical for protein function[58]. Others such as Polyphen2 and VEST use a machine-learning based approach and integrate multiple data types[59, 60], while newer methods like CADD extend these principles to non-protein-coding variants[61]. More recently, methods such as CHASM and CanDrA have focused on somatic mutations in cancer, rather than all mutations occuring over evolutionary time[62, 63].

With the exception of SIFT, these methods use supervised modeling to make predictions. In this approach, a model is trained using mutations that are designated as pathogenic or neutral. An advantage of this strategy is that models can be developed for specific tasks by choosing appropriate training data. For instance, curated training data allows CanDrA and CHASM to detect cancer drivers specifically[62, 63]. However, there are two major weaknesses to this approach when applied to the problem of identifying driver mutations in cancer.

The first major weakness is that there are no “gold standard” datasets of rigorously validated drivers and passengers available for model training. Instead, models must be trained using proxy “silver standard” mutation sets, which introduce biases that can skew models towards known biology and can limit generalizability[64]. Previously, diverse sources including HGMD, dbSNP, UniProt, COSMIC, and simulated mutations have provided training examples[59-63]. These training sets introduce assumptions as to what constitutes a passenger or driver, biasing the model and limiting its generalizability. In general, supervised modeling may not be tractable if available datasets do not adequately represent the sought-after classes of mutations.

The second major shortcoming is that no method utilizes two unique features of somatic mutations in cancer: recurrence and the configuration of mutations within tumors. In general, mutations which appear in multiple tumours are more likely to be causative. However, current

methods are blind to this pattern, as they are generally trained with datasets in which distinct mutations appear only once. These methods are also blind to tumor configuration. For instance, consider a breast tumor with three mutations: two are in olfactory receptors, and one is in HER2. Knowing the context, it is clear that the HER2 mutation should be carefully considered, even if it does not seem likely to be functional on the basis of gene homology or biochemistry. This illustrates how taking tumor configuration into account can yield important new information.

These shortcomings make applying current FISs to cancer somatic mutation data non-intuitive. To address these concerns we will develop an unsupervised, parsimony-guided paradigm for functional impact prediction and apply it to the problem of identifying cancer driver mutations. To do so we will make use of an expectation-maximization (EM) framework. EM is a general approach to fitting statistical models when some data is unobserved. For instance, EM can be used to fit regression models when data is incomplete[65]. It has also been applied to problems such as finding recurrent motifs in unaligned DNA sequences[66].

We will focus on using EM to create a model for predicting mutational functionality, with the functionality labels “missing” for the purposes of training. This is essentially an unsupervised form of training. Since it does not rely on labeling mutations as drivers or passengers at the outset, this approach introduces fewer assumptions and should improve the generalizability of the final model. In lieu of labeled training mutations, the algorithm will assume that drivers should be more equitably distributed among patients than passenger mutations or, equivalently, that the proportion of mutations that are drivers drops as tumor mutation rates increase. This assertion follows from previous observations; studies by Youn *et al* demonstrate that cancer genes (which are enriched in driver mutations) are mutated in relatively *hypo*-mutated tumors more often than chance[53]. More recently, Tomasetti *et al* used mathematical modeling to

suggest that cancers depend on a small but consistent number of driver events, over all mutation rates[67].

This approach elegantly solves the shortcomings we identified above: recurrent mutations will be accurately represented in the training data, leaving this aspect of the data intact; the algorithm will be constantly adjusting scores assigned to mutations such that a few drivers are given to each tumor, taking into account the combinations of mutations within tumors to do so; and finally, the algorithm will not make use of pre-labeled training data, avoiding unnecessary assumptions and biases that can limit generalizability.

The final result of this effort will be a model that assigns mutations a score that represents the likelihood that they act as drivers. While it could be used as a stand-alone FIS, such a model would be particularly useful when used in combination with the other methods we describe, especially for identifying which mutations among a group of otherwise similar variants should be prioritized for testing. Our efforts on this topic are discussed in chapter 3.

1.3.3 Identifying Tumor Drivers in the Kinome

Although the specific effects of many mutations are unknown, many strategies rely on aggregating mutations to draw biological conclusions. For instance, mutations can be drawn from several genes to identify gene networks and pathways that are related to tumor growth[68]. As discussed above, many tools also query mutations at the gene level to identify genes with non-random patterns of mutations that are likely related to cancer development. As the number of mutations increases, even regions within proteins can be assessed[69]. Even though knowledge of specific mutations may be lacking, this approach can guide researchers towards the most promising mutations for further study. However, one limitation of these approaches is that

they operate genome-wide, often without taking into account relevant knowledge of specific gene families or protein types.

For instance, one particularly well studied gene family is the protein kinases. They are an evolutionarily conserved group of phosphotransferases. There are approximately 500 protein kinase domains encoded in the human genome, spread between roughly 485 genes. These signaling molecules have well-known links to a variety of human diseases as well as particular links to cancer due to their widespread functions in regulating cell behaviors (reviewed in [70, 71]). Genome-wide FISs are applicable to protein kinases but do not make use of specialized kinase knowledge, nor do they account for the structural relationships between mutations.

Torkamani and Schork observed that known disease-causing protein kinase mutations are not randomly distributed throughout the protein and developed a machine-learning method for identifying disease-causing mutations[72-74]. When applied to cancer mutations, they observed that predicted functional mutations clustered in hotspots, suggesting that functional mutations may be shared among protein kinases[75]. Recent studies used machine-learning approaches to predict new activating mutations in EGFR[76]. Another approach is to seek common structural effects of functional mutations. Dixit and colleagues demonstrated over several studies that activating mutations shift the active-inactive equilibrium towards the active conformation, and that this is broadly true for several protein kinases[77, 78]. Furthermore, they identified the catalytic and activation loops as particularly prone to gain-of-function events[79, 80].

It is clear that mutations occuring in one protein kinase can be used to draw inferences in another, and that somatic protein kinase mutations can be analyzed under a variety of regimes.

However, these kinase-specific methods rely on prior structural knowledge, which may have no analog in other gene families, limiting generalizability.

We will attempt to identify and validate functional, driver protein kinase mutations using an approach different from those above. Rather than using protein kinase structural knowledge to find functional mutations, we will first pursue the reverse task: using observed mutations and a kinase alignment to identify homologous kinase positions that experience non-random mutations and presumably host driver mutations. To find these mutations, we will develop a series of statistical tests, very similar to those used for identifying cancer genes. Mutations at these positions will then be our putative driver mutation list.

Unlike our other efforts, the results of this analysis will be mutations that exclusively occur in protein kinases, a group of enzymes which our lab is specifically equipped to study. Therefore, we will take the opportunity to validate putative drivers identified by this approach using basic mammalian cell culture techniques. These experiments are fully described in chapter 4.

1.4 Connectedness of Approaches

Taken together, these methods will form a framework for prioritizing mutations for functional characterization. Although they are interrelated, the analyses themselves make use of different assumptions and paradigms to draw conclusions. As such, they can be used in series, each filtering out large portions of passenger mutations, eventually leaving a small set of mutations that should be markedly enriched for driver events. This can be accomplished without limiting scope to recurrent mutations, or those which occur in known cancer genes. Since we rely on observed somatic mutations exclusively in each analysis, we are not limiting potential hypotheses with prior beliefs. Instead, we are making use of the same somatic mutations from

multiple perspectives, ultimately gleaning biological hypotheses that are more precise and reliable than was previously possible. Ultimately, these improvements could spur experiments on a wider variety of hypotheses discovered during high-throughput sequencing efforts.

1.5 Candidate Contributions

Chapters 2, 3, and 4 are substantially adapted from previously written manuscripts (cited at the beginning of each chapter). These manuscripts are at various points in the process of publication, but will ultimately appear in peer-reviewed scholarly journals. I (Runjun D. Kumar) substantially designed and executed relevant experiments, wrote these manuscripts, and adapted them to this dissertation. This is reflected by the fact that I am the lead author on each of the manuscripts themselves. Portions of this introduction, especially section 1.3, are also adapted from these manuscripts. In addition to being the primary author of the works presented in this dissertation, I am also the creator of all tables and figures.

2. Identifying Cancer Genes

This chapter is adapted from:

Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, Bose R. (2015). Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics.* 31(22). pp 3561-3568.

2.1 Introduction

In this chapter we develop new approaches to the problem of detecting cancer genes from somatic mutation data and use them to identify new putative cancer genes. We use a pan-cancer dataset of 1.7 million mutations and a manually curated set of 99 high confidence (HiConf) cancer genes to develop a panel of five statistical tests. The tests detect different signals of positive selection and are designed to detect putative oncogenes and TSGs (see Figure S2.1 for representative examples). Several of our tests make use of previously described signals of selection, such as functional impact bias, mutation clustering, or rates of truncating events[48, 49]. We also identify patient and cancer type bias as new signals of selection and leverage them to markedly improve detection of oncogenes. We then integrate these tests into a random forest model which can identify oncogenes and tumor suppressors as separate types of cancer genes. We validate by assessing the performance of previous tools and our new methods against several independent panels of known and putative cancer genes. Finally, we explore the performance of these methods in specific cancer types and suggest new putative cancer genes. By producing a model that uses several signals of selection to identify oncogenes and tumor suppressors as separate classes, and by validating performance using objective criteria, we address many of the shortcomings of previous studies that attempted to identify cancer genes from somatic mutation data.

2.2 Materials and Methods

2.2.1 Data Gathering and Quality Control

Mutation Annotation Files (MAFs) were drawn from data repositories for the TCGA, ICGC and COSMIC. Only columns for Cancer Type, Study, Patient Identifier, Chromosome, Start Position, End Position, Reference and Variant Allele were retained. A small number of hg18-based studies (accounting for ~2% of the dataset) were converted to hg19 using the UCSC Genome Browser liftOver utility with default settings[81]. Patient samples were frequently included in more than one dataset, potentially producing duplicate or contradictory mutations. For a given patient and genomic position, only mutations from the most recent dataset were retained. Data were annotated with the ANNOVAR software suite using RefSeq libraries[82]. Mutations were also labelled with functional impact scores to allow the use of Oncodrive-fm[48]. In cases where a gene was related to multiple transcripts and isoforms, the transcript which preserved the most mutations was used first, with mutations from alternate isoforms being annotated as such. Data was gathered July 27th to August 1st, 2013.

2.2.2 HiConf Cancer Gene Panel Construction

As our goal is to use mutation data to find potential cancer genes that can be confidently carried into biological experiments, we sought high confidence (HiConf) cancer genes that are biologically established as a training data set. This HiConf panel was focused towards known cancer genes that have previously been detected through genetic criteria, and which could plausibly be detected with exome sequencing data. The following steps were taken to ensure the HiConf cancer gene panel met these criteria. The Cancer Gene Census (CGC) provided candidates[83]. Genes which have only been observed in translocations (as per the CGC annotations) were immediately eliminated, as our dataset lacks translocation events and it is

often unclear whether translocation partners are active cancer genes individually. This left 204 candidate genes.

A literature search was then performed. A gene qualified for the HiConf panel if a scientific publication could be found which fulfilled one of the following: 1) Demonstrated a cancer-like phenotype in cell lines when the gene was activated or inhibited. 2) Demonstrated a change in disease progression in mouse models of cancer when the gene was activated or inhibited. 3) Demonstrated the gene as a causative agent of a Mendelian human tumor syndrome. Importantly, all means of gene alteration (RNAi, ectopic expression, drug or antibody targeting, null models, etc.) were accepted for animal and cell studies, and any phenotype outlined in the Hallmarks of Cancer was accepted as cancer-like[84]. The sources for the literature search included OMIM and PubMed. The literature search left 99 HiConf cancer genes (Table 2.1). Based on the preponderance of literature recovered, these genes were further categorized as oncogenes (ONCs, gain of function causes pro-cancer phenotype) or TSGs (loss of function causes pro-cancer phenotype).

2.2.3 Comparison Tools

Three existing tools were applied to the dataset: MutSigCV[45], OncodriveCLUST[49], and Oncodrive-fm[48].

MutSigCV identifies likely cancer genes by detecting genes with elevated mutation rates. MutSigCV v1.3 was run on the dataset using scripts downloaded from http://www.broadinstitute.org/cancer/cga/mutsig following the provided instructions and default settings. The MutSigCV p-value was used to assess tool performance. OncodriveCLUST is a cancer gene detection method which uses intra-protein mutation clustering to identify possible cancer genes. Software was downloaded from http://bg.upf.edu/group/projects/oncodrive-

clust.php, and run on the dataset using the included instructions and default settings. Oncodrive-fm detects genes with unusually impactful mutations, as judged by a suite of functional impact scores (SIFT, PolyPhen2 and MutationAssessor)[48]. The Oncodrive-fm method was re-implemented in R, using the dataset itself as an internal null distribution.

2.2.4 Calculation of Individual Tests

We assembled five statistical tests that target several signals of positive selection in cancer genes. *Patient Distribution* and *Cancer Type Distribution* operate similarly and detect genes that are mutated in nonrandom sets of patients or cancer types. *Unaffected Residues* is our method to identify genes with unusual levels of mutation recurrence. *VEST Mean* uses VEST scores[60] to identify functional impact bias among genes. Finally, *Truncation Rate* is our approach to detecting genes that have unusual numbers of truncation events; either an enrichment (as is expected of TSGs) or depletion (as is expected of oncogenes).

Patient Distribution and *Cancer Type Distribution* are calculated similarly. Each mutation occurs within a patient (or cancer type). A randomly mutated gene should be mutated in a random set of patients (or cancer types). This null hypothesis can be tested using the Pearson Chi-Square Goodness-of-Fit test, with the entire dataset providing the null expectations. For each gene *g*, a chi-square statistic was calculated:

$$X_g^2 = \sum_{p=1}^{P} \frac{(O_p - E_p)^2}{E_p} \qquad E_p = \frac{N_p N_g}{N}$$

Where *O* is the observed count of mutations for a given patient (or cancer type), *E* is the expected number of mutations for the same patient (or cancer type), and *P* is the number of

unique patients (or cancer types). The expected count for a given patient (or cancer type) and gene is the product of the total number of mutations in the patient (N_p) and the total number of mutations in the gene (N_g) divided by the number of mutations in the dataset (N). The p-value is calculated by simulation since the low expectations would violate normality assumptions required to use the theoretical chi-square distribution. Given the number of mutations in a gene, the test statistic is calculated for 10,000 random draws from the full list of patient (or cancer type) labels with replacement, and the upper tail probability of a higher test statistic under the null distribution is reported. All mutations, including synonymous mutations, are used when calculating *Patient Distribution* and *Cancer Type Distribution*.

Unaffected Residues detects high levels of recurrence by considering the number of un-mutated residues in a gene. First, given the number of mutations and the protein length, the probability of a residue being un-mutated is calculated based on the Poisson distribution. Because the mutation count is zero, the estimated probability of an unaffected residue simplifies to:

$$\hat{P}_{zero} = e^{-n/l}$$

Where n is the number of mutations in the protein, l is protein length, and $\hat{P}_{zero}$ is the estimated probability of a given residue being un-mutated. Once $\hat{P}_{zero}$ is calculated, the binomial distribution is used to calculate the probability of a gene having at least the observed number of unaffected residues:

$$P\ (X \geq x) = \sum_{i=x}^{l-1} \binom{l}{i} \hat{P}_{zero}{}^{i}\ (1 - \hat{P}_{zero})^{l-i}$$

Where *x* is the observed number of unaffected residues and *l* is the protein length. *Unaffected Residues*, represents the probability of a gene having as many or more unaffected residues as observed if mutation location is entirely random. Only nonsynonymous protein-coding mutations are used to calculate this test, as recurrent synonymous mutations can suggest alignment errors and may produce false positives.

VEST Mean is calculated in a very similar manner as the individual sub-scores used within Oncodrive-fm[48], but uses the Variant Effect Scoring Tool as the base functional impact score[60]. It is the upper tail probability of a gene having a mean VEST score greater than that observed, given the number of mutations, based on 10,000 random draws with replacement from all observed VEST scores. VEST scores are limited to missense mutations, so imputation was required for other mutations. We used the same rationale as was used in Oncodrive-fm. Synonymous and non-coding mutations were assigned a value of 0, the lowest functionality score under VEST, while in-frame and frameshift indels, premature stop, nonstop, and splice site mutations were assigned the highest value of 1. Synonymous and nonsynonymous mutations are used in this calculation.

To use *Truncation Rate*, a gene's mutations are categorized as truncating (i.e. Splice Site, Frameshift Insertion/Deletion, Premature Stop/Nonsense) or non-truncating. Then the upper tail binomial probability of at least the observed number of truncation events (using the truncation rate across the dataset for the null distribution) is calculated as:

$$P\ (T \geq t) = \sum_{i=t}^{n} \binom{n}{i} \hat{P}_{Trunc}{}^{i}\ (1 - \hat{P}_{Trunc})^{n-i}$$

Where $\hat{p}_{Trunc}$ = 182,030 truncation events / 1,703,709 mutations = 0.107, *t* is the observed number of truncating events for a given gene, and *n* is the total number of mutations in the gene. Synonymous and nonsynonymous mutations are used in this calculation.

2.2.5 Imputation of Missing Data

Our tests rely on very basic annotations (e.g. Sample ID, Cancer Type, Mutation Type, *etc.*) and consequently we had very low rates of missingness. Two exceptions warrant note, and both are related to *Unaffected Residues.* This test requires a valid protein length to be calculated; however, after integrating datasets, ~4% of genes had protein lengths smaller than the most downstream mutations. In these cases, the test uses the most downstream mutation position as a conservative proxy of protein length. The other exception is in model training. Most of our tests are calculable for virtually all genes. The exception is *Unaffected Residues*, which cannot be calculated for the ~10% of genes with no coding nonsynonymous mutations. The data matrix was filled in by mean imputation prior to model training. Missing values were excluded from the calculation or assessment of individual tests.

2.2.6 Generation of Ensemble Model

We compared Random Forests, SVMs and Naïve Bayes classifiers in separating the three gene classes (Unknown Function, HiConf Oncogenes, HiConf TSGs) using the individual tests of our panel. Random Forests and SVMs both performed well. Random Forests were chosen because they have been used in previous tools such as OncodriveROLE[56] and worked well with default settings (*mtry*=2, trees=500).

To generate the scores and predictions used in the study, we trained a random forest (RF5) on the five individual tests (*Patient Distribution, Cancer Type Distribution, Unaffected Residues, Truncation Rate, VEST Mean*) and labels generated from the HiConf panel (22,801 unknown

genes, 48 TSGs, 51 Oncogenes). TSGs and ONCs were up-sampled during training to better calibrate the model (trees were trained on 300 unknowns, 30 TSGs and 30 ONCs). 5-fold cross validation was used to generate predictions, repeated 50 times. The repeated cross validation runs were averaged to generate the stable predictions presented in the final results.

2.2.7 Assembly of Validation Gene Panels

In addition to the manually curated HiConf gene panel, we also sought out additional panels of established cancer genes. These panels are necessary to validate the performance of our random forest model, since even with cross validation its performance on the HiConf panel could be over-optimistic.

We gathered the High Confidence Driver (HCD), Cancer5000 and TSGene lists as presented in Schroeder *et al.* (2014)[56]. These were originally generated by Tamborero *et al.* (2013), Lawrence *et al.* (2014) and Zhao *et al.* (2013) respectively[31, 55, 85]. In addition to the filters applied by Schroeder and colleagues, we ensured independence by depleting these lists for any members of the HiConf panel, leaving 149, 96 and 55 genes in the respective panels. Note that while they are independent of the HiConf list, they do overlap with one another.

While the HCD and Cancer5000 lists may contain both oncogenes and TSGs, the TSGene list is composed of TSGs exclusively. To generate an oncogene-only list, we defined the Kinase list as any kinase bearing a known activating cancer mutation in Kin-Driver[86]. This panel consists of 29 genes after being depleted of HiConf genes.

2.2.8 Cancer Subset Analysis

Cancers with at least 500 patients or 200,000 mutations were considered in the cancer type analysis. Tests and RF5 models were applied using identical procedures to the pan-cancer analysis.

2.2.9 Statistics and Software

All comparisons of AUROCs were performed as two-sided DeLong Tests[87] with adjustment for ties. All analyses were performed in R v2.15. Modeling was performed using methods available through the randomForest[88] and e1071 (Support Vector Machines, Naïve Bayes) packages. AUROCs, ROC plots and DeLong Tests were performed using the pROC package.

2.2.10 Data Availability

The dataset is available along with a script, instructions and sample data to be used to train RF5 models on any dataset. Please see www.github.com/Bose-Lab/Improved-Detection-of-Cancer-Genes.

2.3 Results

2.3.1 Description of Data

The analytic flow follows the schema in Supplementary Figure S2.2. The final dataset includes 1,703,709 mutations across 10,239 patients (Supplementary Figure S2.3). 22,902 genes appear at least once in the dataset, with a median of 49 mutations per gene. This is one of the largest assembled pan-cancer data sets and is publicly accessible (See section 2.2.1 for more details).

2.3.2 Developing a Panel of Known Cancer Genes

Based on a criteria-driven literature review, 99 genes were collected into a high-confidence cancer gene panel (HiConf, Table 2.1, see section 2.2.2 for details). The HiConf gene panel was further divided into 48 TSGs (with 15,698 mutations) and 51 ONCs (with 11,243 mutations).

Rather than defining a separate set of presumptively neutral genes, the remaining 22,801 genes were labelled as “unknown”. Most unknown genes are neutral with regards to cancer progression, and the set as a whole is treated as neutral for the purposes of training and assessment.

2.3.3 Assessing Individual Tests

Individual methods of cancer gene prediction must separate the distinct mutation patterns of ONCs, TSGs and neutral genes. In particular, TSGs tend to be enriched in truncation events, while oncogenes are depleted; in addition, oncogenes tend to have clustered mutations (Supplementary Figure S2.1). We performed statistical tests for each of five signals of positive selection, and refer to them as follows: *Truncation Rate* (rate of truncating events), *Unaffected Residues* (intra-gene mutation clustering/recurrence), *VEST Mean* (functional impact bias), *Patient Distribution* (bias in patient labels), and *Cancer Type Distribution* (cancer type bias). OncodriveCLUST, Oncodrive-fm, and MutSigCV were also applied to the dataset[45, 48, 49].

We use the Area Under Receiver Operator Characteristic (AUROC) to gauge performance as it is threshold independent and testable[87]. In particular, we consider the following classification tasks: separation of the HiConf oncogenes (ONCs) and tumor suppressors (TSGs) from other genes of unknown function (UK) as separate and pooled classes, and separation of ONCs and TSGs from one another.

Patient Distribution is notable because it relies on a novel cancer gene signal which we call patient bias. The contribution of tumors to the pan-cancer dataset is highly unequal because tumor mutation rates vary by up to four orders of magnitude (Supplementary Figure S2.3). However, mutations within HiConf TSGs and oncogenes are much more evenly distributed between patients (Figure 2.1A). *Patient Distribution* makes use of a chi-square statistic to detect

genes which are frequently mutated in relatively hypo-mutated tumors. Unlike many of the other statistics and tools we assessed, *Patient Distribution* detects HiConf oncogenes and HiConf TSGs equally well (Figure 2.1B, Table 2.2). In fact, it is the single best test for detecting oncogenes and the HiConf panel as a whole, with AUROCs of 0.894 and 0.900, respectively. HiConf oncogenes including TRAF7 and ALK are missed by previously published tools at the $p<0.05$ cutoff, but are easily detected by *Patient Distribution*.

Cancer Type Distribution is very similar to *Patient Distribution*, but relies on cancer type bias to identify cancer genes. For instance, it easily highlights VHL, a HiConf tumor suppressor which is frequently truncated in renal clear cell carinomas. It also identifies the HiConf tumor suppressor PTCH1, which is not identified by existing tools.

Unaffected Residues is our test of mutation clustering and recurrence. Rather than testing for clustering directly, as was the approach taken by OncodriveCLUST[49], we instead examine the number of unmutated residues. *Unaffected Residues* is a one-tailed binomial test for the number of unmutated residues, assuming the number of mutations per residue is poisson distributed. It is the second best method for detecting HiConf ONC (AUROC=0.855) and third best for the whole HiConf panel (AUROC=0.861). It is superior to Oncodrive-CLUST in these tasks (Table 2.2). The top four genes according to *Unaffected Residues* are KRAS, PIK3CA, BRAF and TP53, all of which are HiConf cancer genes with well known mutation clusters.

VEST Mean tests for functional impact bias. It reports the probability of a randomly mutated gene having a higher average functional impact, very similar in concept to the method used in Oncodrive-fm[48], but with better oncogene detection (AUROC=0.796 vs 0.710, Table 2.2). It is also the best method for detecting TSGs (AUROC=0.938).

Truncation Rate is the final test and has unique properties. It is well appreciated that TSGs are enriched in protein-truncating mutations (Figure 2.1C). This pattern lead Vogelstein and colleagues to suggest that genes with greater than 20% truncating events be considered putative TSGs[44]. *Truncation Rate* formalizes this concept using a one-sided binomial test, which reports the probability of a randomly mutated gene bearing an equal or greater number of truncation events (i.e. splice site, premature stop and frameshift indels), given a fixed number of mutations. However, as a one-sided test, *Truncation Rate* is also sensitive to the relative depletion of truncation events in oncogenes (Figure 2.1D). It is by far the best method for separating oncogenes and TSGs (AUROC=0.922, Table 2.2, ROC curves in Supplementary Figure S2.4).

2.3.4 Integration into a Single Model

As Table 2.2 illustrates, the tests we have identified are complementary, each having different performances in our classification tasks. We hypothesized that a model integrating the full panel would be able to separate all three gene classes (HiConf Oncogenes, HiConf TSGs, all other genes of unknown function) from one another. To test this hypothesis, we trained a Random Forest model on the five individual test values as predictor variables. Gene labels were generated from the HiConf panel, resulting in 51 oncogenes (ONC), 48 tumor suppressors (TSG), and 22,801 unknown genes (UK). Most of the UK genes are passenger genes, so this large class serves as a neutral class for training. Training was performed in 5-fold cross validation, with results averaged over 50 repetitions.

The five-test model, which we refer to as RF5, produces a score for probability of membership in each class. These scores summate to 1, allowing genes to be visualized in a ternary plot (Figure

2.2). UK genes which are placed near the ONC and TSG regions are putative cancer genes, while HiConf ONCs and TSGs which are assigned to the Unknown region are false negatives.

As Figure 2.2 shows, RF5 is able to delineate most HiConf ONCs and TSGs from the bulk of UK genes. It also suggests a large number of UK genes which appear similar to ONCs and TSGs. Simultaneously, RF5 separates the ONCs and TSGs from one another. When assessed for performance at each task separately, RF5 is significantly better or not significantly different from the best individual tests. It performs markedly better than *VEST Mean* in detecting TSGs (AUROC=0.980 vs 0.938), the same as *Patient Distribution* in detection ONCs (AUROC=0.891 vs 0.894), and the same as *Truncation Rate* in separating ONCs and TSGs (AUROC=0.924 vs 0.922, Table 2.2). HiConf genes which are identified by few of the individual tests can often be identified confidently by RF5, demonstrating the importance of integrating multiple approaches (Supplementary Figure S2.5A).

2.3.5 Detection of Validation Gene Panels

The HiConf panel serves as our primary method of assessment for pre-existing tools and our new methods. However, RF5 is trained to detect the HiConf panel, and it is possible the RF5 performance estimates are optimistic even with cross-validation. Therefore, we retrieved four validation panels and depleted them of the HiConf panel members to ensure independence (see section 2.2.7). We then assessed the ability of our methods to prioritize the validation panels over other genes of unknown function (Table 2.3).

The High Confidence Driver (HCD) panel was defined by Tamborero *et al. (*2013) using a variety of existing tools including Oncodrive-fm[55]. After excluding HiConf cancer genes, it consists of 149 members. We find that RF5 has the best performance (AUROC=0.884) on this set, but that *VEST Mean* and Oncodrive-fm are not significantly different. This is expected, as

the list was defined in part using Oncodrive-fm. We also examined the Cancer5000 gene panel, which has 96 members after depletion of HiConf genes[31]. It overlaps by roughly 50% with the HCD panel, and RF5 still has the strongest performance (AUROC=0.943). This panel was defined using MutSigCV, which performs well as expected (AUROC=0.882).

Unlike the HCD and Cancer5000 panels, the TSGene and Kinase panels are largely composed of TSGs and oncogenes, respectively (see section 2.2.7). The TSGene panel consists of 55 manually curated tumor suppressors[85]. *VEST Mean* has the highest performance (AUROC=0.876), but RF5 is not significantly worse. The Kinase panel consists of 29 manually curated kinases that are known to harbor activating mutations in cancer[86]. RF5 again has the strongest performance (AUROC=0.801).

2.3.6 Predicted Cancer Genes

For brevity and clarity we will focus on the top 100 predictions made by RF5 in the pan-cancer setting. They include many potentially new cancer genes, of which we will highlight a few (Supplementary Figure S2.5B, Supplementary Figure S2.6). Several genes are related to chromatin structural and epigenetic regulation. GPS2 and HDAC2 are members of the NCOR-HDAC3 complex[89] and are predicted TSGs. HIST1H1E, a linker protein in nucleosomes, is predicted to be an oncogene. Other novel predicted cancer genes are drawn from a range of biological classes: CACNG3 (predicted oncogene) is a voltage-dependent calcium channel subunit; NXF1 (predicted TSG) is a nuclear RNA export factor; and HLA-DRB1 (predicted TSG) is a subunit of MHC Class II. Additionally, several experimentally known cancer genes are linked to human tumors through somatic mutation data for the first time. Among these are the oncogenes SGK1[90] and TMEM30A[91] as well as the TSGs RBM5[92], CHD4[93] and CHD2[94]. While these are not new cancer genes, their identification by patterns of somatic

mutations supports their relevance in human disease. None of these genes are listed in the Cancer Gene Census.

Many top predicted cancer genes are potentially druggable. A query of the Drug Gene Interaction Database reveals that 26 of the top 100 predicted cancer genes have known interactions with drugs, and an additional 43 belong to a potentially druggable gene category[95]. With the majority of top predictions being potentially druggable, the prioritized gene list presents opportunities for both new discoveries in cancer biology and more immediate pharmacologic interventions.

RF5 also makes high quality predictions. For instance, very few of the top 100 predicted cancer genes are biologically implausible. Among these genes, there is one olfactory receptor (OR4C5) and one collagen (COL2A1)[45]. However, technical artifacts remain a concern. For instance, the highly ranked genes IL32 and PLAC4 have multiple recurrent frameshift and synonymous events. An examination of alignment files from several of the affected patients suggests these genes are prone to alignment errors (data not shown). These examples illustrate the need for human expertise in scrutinizing prioritized gene lists, and the quality of data and associated mutation calls in particular.

2.3.7 Application to Specific Cancer Types

Of our tests, only *Cancer Type Distribution* relies on multiple cancer types; the others may perform differently in individual cancers. To address this possibility, the tests from Table 2.2 as well as RF5 models were generated for each cancer with at least 500 patients or 200,000 mutations (breast, colorectal, lung, melanoma and endometrial cancers). This analysis demonstrates that the relative performance of these tests is quite consistent across cancer types

and that our new methods outperform previous tools in a variety of settings (Supplementary Figure S2.7).

Analyzing the pan-cancer set may allow us to detect additional cancer genes due to increased power, but it may also mask cancer-specific cancer genes. To explore this possibility, we examined the detection of the HiConf panel in the pan-cancer and cancer-specific datasets. We found that 11 HiConf cancer genes were detected in the pan-cancer dataset, but not in the individual cancers, while 10 were detected in at least one of the specific cancers, but not in the pan-cancer set (Supplementary Figure S2.8A). This suggests that we are likely to make some predictions only in the pan-cancer set, and others only in specific cancers. In fact, we found that 30 of our top 100 pan-cancer predicted cancer genes could only be detected in the pan-cancer set (Supplementary Figure S2.8B). These included many promising potential cancer genes such as HDAC2, NXF1 and TMEM30A, illustrating the value of pooling cancers.

We then sought cancer genes that are cancer-specific and compared their detection across cancer types. We gathered the top 100 predictions for each of breast, colorectal, lung, melanoma and endometrial cancers. Roughly half of the top predictions were cancer-specific (Supplementary Figure S2.9). A few examples include: MED23 and MYB as putative TSGs in breast cancer; TGIF1 and B3GNT6 as putative TSGs in colorectal cancer; CDK14, IRF2BPL and NTRK2 as putative oncogenes in lung adenocarcinoma; CCDC28B and ATAD2 as potential TSGs in melanoma; and EIF3C as a potential oncogene in endometrial cancer. We conclude that large numbers of cancer genes may be cancer-specific. Taken together, these results suggest the importance of searching for cancer-genes in the pan-cancer and specific-cancer settings.

2.4 Discussion

One use of cancer genome sequencing results is the identification of novel cancer genes. This problem has two stages: first, cancer genes must be separated from genes bearing only passenger mutations; second, cancer genes must be sorted into likely tumor suppressors and oncogenes. Both stages are crucial because mechanism-specific predictions are needed to guide downstream analyses and experiments. In this chapter, we gathered a pan-cancer dataset of 1.7 million variants and a manually curated set of 99 known cancer genes (HiConf panel). Using these data, we designed and assessed a panel of statistical tests which identify cancer genes using several signals of selection, as well as separate cancer genes by mechanism of action. We also compared the performance of these tests to previous tools in accomplishing these tasks.

In general, we found that HiConf TSGs were easier to detect than HiConf oncogenes. Several methods had AUROCs of 0.9 or higher, including the published tool Oncodrive-fm and our tests of patient and cancer type bias (*Patient Distribution*, *Cancer Type Distribution*). However, the best single method for detecting TSGs was our test of functional impact bias, *VEST Mean*, with an AUROC of 0.938.

In contrast, HiConf oncogenes were less easily identified. This is concerning because oncogenes provide more direct targets for drug development. The best performing existing tool for detecting the HiConf oncogenes was OncodriveCLUST with an AUROC of 0.808. With the exception of *Truncation Rate*, all of the tests in our panel had AUROCs of 0.80 or greater when detecting HiConf oncogenes. Particular improvement was observed with *Unaffected Residues*, which tests for mutation clustering/recurrence and had an AUROC of 0.855, and *Patient Distribution*, which was the best performer with an AUROC of 0.894.

Two of our tests warrant emphasis. *Patient Distribution* uses a novel signal of positive selection. It identifies genes with mutations that occur in nonrandom sets of patients, particularly genes with mutations that occur in relatively *hypo*-mutated tumors, as would be anticipated of genes bearing driver mutations. For identifying the HiConf panel as a whole (TSG + ONC), *Patient Distribution* is the strongest performer with AUROC of 0.900. Another important member of our statistical panel is *Truncation Rate*. This test is a formalized version of the 20/20 rule for TSGs put forward by Vogelstein *et al*.[44], and uses the binomial distribution to model the expected number of truncation events per gene. *Truncation Rate* can be used to separate TSGs and oncogenes with an AUROC of 0.922. It is the only method that usefully accomplished this task.

Since the individual tests of our panel offered complementary strengths, we also integrated them into a single model. We found that a random forest built on our five tests (RF5) was effective at separating HiConf oncogenes and TSGs from passenger genes, and from one another. Moreover, this integration did not require any loss in performance: RF5 is as good as or better than the individual methods at every classification tasks we assessed. We also confirmed these results in several independent validation gene panels.

RF5 identifies many potential pan-cancer cancer genes. These include the predicted oncogenes CACNG3 and HIST1H1E, and the predicted TSGs HDAC2, GPS2, NXF1 and HLA-DRB1. It also identifies several known cancer genes through genome sequencing for the first time, including SGK1, TMEM30A, CHD2, CHD4 and RBM5. Many RF5 predictions are potentially druggable. Furthermore, additional cancer genes can be identified when focusing on single cancer types. In fact, we found that half of RF5 predictions within tumor types were cancer-specific. These results illustrate the importance of searching for cancer genes in both the pan-cancer and specific-cancer settings and suggest many new potential avenues of research.

However, there remains room for improvement. As Figure 2.2 illustrates, some oncogenes and TSGs could not be detected by RF5, and some were not detectable by any individual test or pre-existing tool (Supplementary Figure S2.5A). There are two major explanations. Foremost, cancer genes will be undetectable if they are primarily altered through means other than somatic mutations in the exome. Additionally, our cancer gene panels may include genes that are involved in later stages of disease progression, such as metastasis and drug resistance. These are true cancer genes, but may be undetectable in the available data as tumor samples largely come from newly diagnosed patients[32]. Fortunately, our methods are highly expandable, and multiple strategies could improve performance, such as: 1) Introduction of additional, heterogeneous data types. 2) Improved tests. 3) Improved model design and training. 4) Expansion of the HiConf cancer gene panel.

In conclusion, our results demonstrate that the detection of putative cancer genes requires a mix of complementary methods. We have developed a panel of five statistical tests that outperform previous methods. In particular, *Patient Distribution* detects oncogenes especially well. We have also integrated these tests into a single classifier, and demonstrated that it performs as well or better than previous tools in both training and validation cancer gene panels. We rely on a manually curated set of high-confidence cancer genes to objectively measure the performance of our new methods and existing strategies. The innovations presented in this chapter address many of the shortcomings present in previous works that attempted to identify cancer genes from somatic mutation data.

Oncogenes						Tumor Suppressors				
ABL1	ERBB2	IDH1	MET	PDGFRA	TSHR	AMER1	CEBPA	MEN1	PRKAR1A	SUFU
AKT1	EZH2	JUN	MITF	PIK3CA		APC	CREBBP	MLH1	PTCH1	TET2
AKT2	FAS	KDR	MLL	REL		ATM	CYLD	MSH2	PTEN	TNFAIP3
ALK	FGFR2	KIT	MYC	RET		AXIN1	DICER1	MSH6	RB1	TP53
BCL6	FGFR3	KRAS	MYCL1	RNF43		BAP1	EP300	NF1	SETD2	TSC1
BRAF	FLT3	MAP2K1	MYCN	SMO		BRCA1	FBXW7	NF2	SMAD4	TSC2
CARD11	GNA11	MAP2K2	MYD88	SOX2		BRCA2	GATA3	NOTCH1	SMARCA4	VHL
CCNE1	GNAQ	MAP2K4	NFE2L2	STAT3		CDH1	HNF1A	NOTCH2	SMARCB1	WT1
CTNNB1	GNAS	MDM2	NKX2-1	TERT		CDKN2A	KDM6A	PAX5	SOCS1	
EGFR	HRAS	MDM4	NRAS	TRAF7		CDKN2C	MAX	PIK3R1	STK11	

Table 2.1. HiConf cancer gene panel members. The genes included in the HiConf panel are listed, according to their status as an oncogene or tumor suppressor (see section 2.2.2).

	ONC+TSG vs. UK	TSG vs. UK+ONC	ONC vs. UK+TSG	ONC vs. TSG	In RF5	Description
RF5	**0.935**	**0.980**	**0.891**	**0.924**[b]		Cross-validation predictions of a Random Forest trained on the five indicated features and the HiConf panel.
Patient Distribution	0.900	0.905	0.894*	0.556	Yes	Detects deviation from expected patient distribution with a chi-square statistic. P values by resampling.
Truncation Rate	0.788[a]	0.904	0.694	0.922*	Yes	Detects enrichment or depletion of Frameshift Indels, Nonsense and Splice Site events using binomial distribution.
Unaffected Residues	0.861	0.865	0.855*	0.479	Yes	Detects clustering using poisson and binomial distributions to calculate probability of unaltered residues.
VEST Mean	0.866	0.938	0.796	0.710	Yes	Detects high functional impact (based on VEST3). P-values from resampling.
Cancer Type Distribution	0.853	0.905	0.803	0.612	Yes	Detects deviation from expected cancer distribution with a chi-square statistic. P values by resampling.
MutSigCV	0.760	0.896	0.632	0.723	No	P-value retrieved from MutSigCV. Detects high rates of mutation based on gene-specific background mutation rate.
OncodriveCLUST	0.776	0.741	0.808	0.597	No	P-value retrieved from OncodriveCLUST summary report. Detects high rates of clustering.
Oncodrive-fm	0.818	0.932	0.710	0.725	No	P-value retrieved from Oncodrive-fm. Detects high rates of functional events using several functional impact scores.

Table 2.2. AUROCs of individual tests and RF5 model with HiConf panel. AUROC = Area Under Receiver Operator Characteristic, for the separation of the indicated gene classes. ONC=HiConf Oncogenes, TSG=HiConf Tumor Suppressor Genes, UK=genes of unknown relevance to cancer growth. (*) These performances are *not* significantly different from RF5 performance at $p<0.05$. [a]*Truncation Rate* is converted to a two-tail test when identifying the combined HiConf panel. [b]The ratio of the RF5 TSG and ONC scores is used to separate these classes.

	HCD	Cancer5000	TSGene	Kinases
RF5	**0.884**	**0.943**	**0.843**	**0.801**
Patient Distribution	0.713	0.768	0.644	0.745*
Truncation Rate	0.751[a]	0.836[a]	0.711	0.515
Unaffected Residues	0.825	0.849	0.792*	0.761*
VEST Mean	0.876*	0.921*	0.876*	0.731*
Cancer Type Distribution	0.811	0.861	0.760	0.612
MutSigCV	0.801	0.882	0.771*	0.525
OncodriveCLUST	0.686	0.742	0.699	0.657
Oncodrive-fm	0.876*	0.923*	0.835*	0.627

Table 2.3. AUROCs of individual tests and RF5 with validation gene panels. (*) These performances are *not* significantly different from RF5 at $p<0.05$. [a]*Truncation Rate* is calculated as a two-tail test for the HCD and Cancer5000 panels, as they combine TSGs and oncogenes.

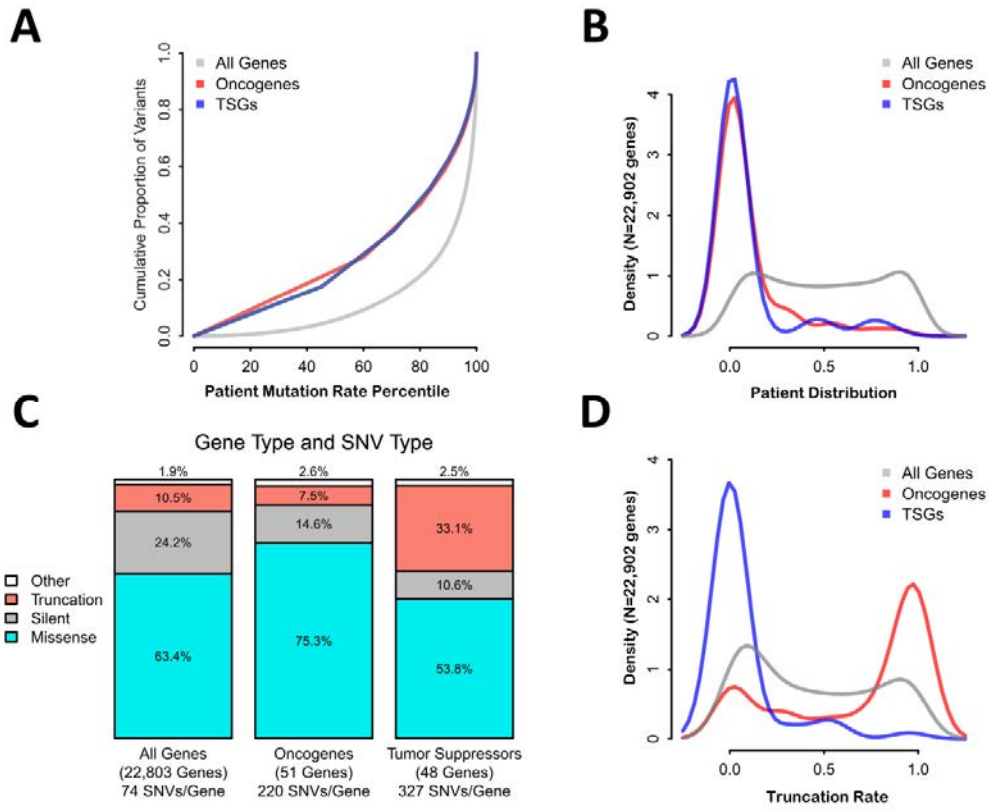


Figure 2.1. Tests of patient distribution and truncation rate. A) Patients have unequal mutation rates, but this effect is less pronounced when considering only HiConf ONCs and TSGs. **B)** *Patient Distribution* is the p-value from a chi-square goodness-of-fit test for the distribution of patients a gene is mutated in, versus the distribution of patients generally. *Patient Distribution* can separate ONCs and TSGs from most other genes, but not from one another**. C)** Distribution of mutation types for each of three gene types. TSGs are relatively enriched for truncating events (nonsense, frameshift and splice site) while ONCs are depleted. **D)** *Truncation Rate* is the binomial upper tail probability of a gene having an equal or higher percentage of truncating mutations. *Truncation Rate* can separate HiConf TSGs and ONCs from one another.

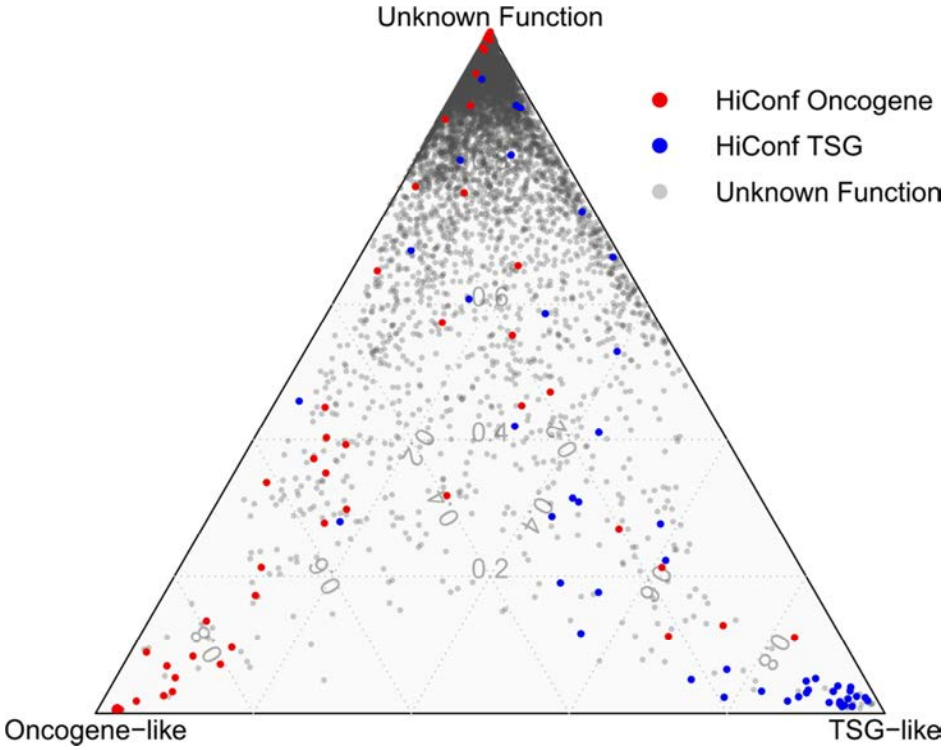


Figure 2.2. Predictions from the RF5 ensemble model. Cross-validated predictions of the random forest model. N=22,902 genes. Using the three class-specific scores generated by RF5, genes can be stratified as oncogene or TSG-like. Genes which are judged as oncogene or TSG-like, but are not on the HiConf panel, are putatively related to cancer.

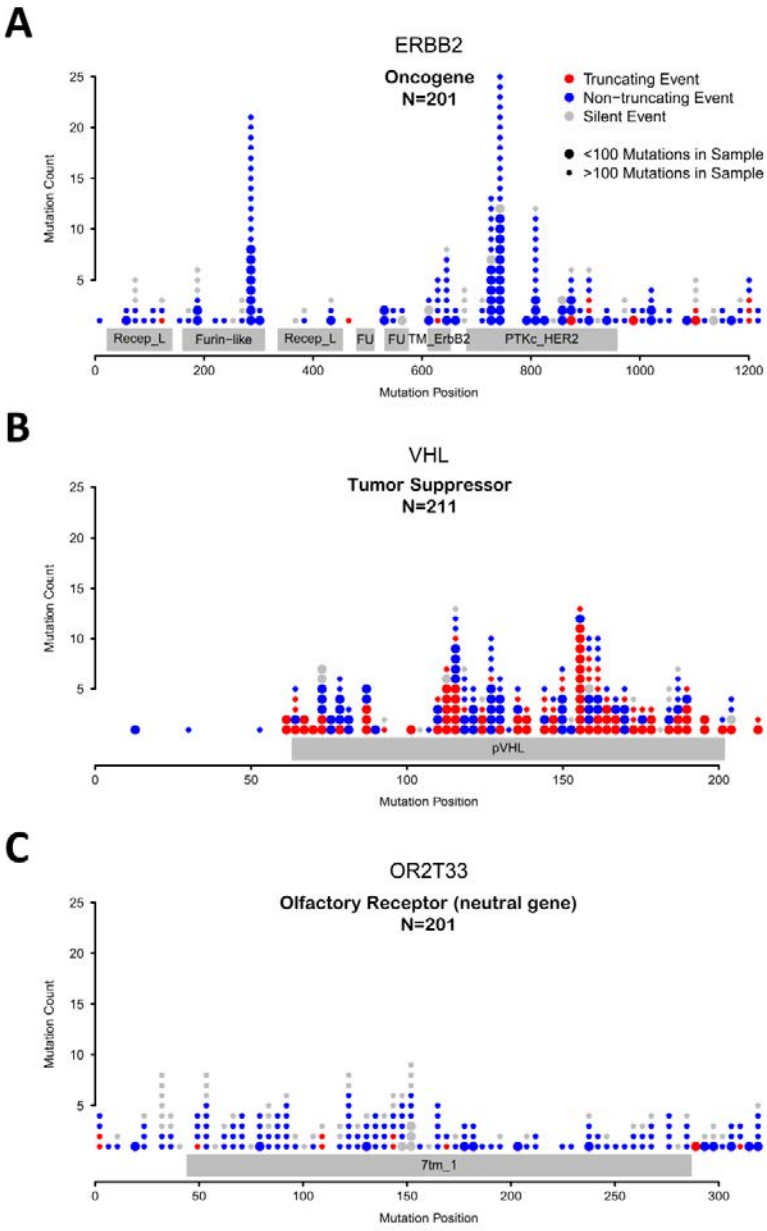


Figure S2.1. Mutation profiles of typical gene-class members. A) ERBB2 (HER2, 201 mutations) is a well-known oncogene and has regions of high mutation density which correspond to known activating mutations. B) VHL (Von Hippel-Lindau factor, 211 mutations) is a known tumor suppressor, and is enriched for truncating (nonsense, frameshift and splice site) events in this dataset. C) OR2T33 is an olfactory receptor (201 mutations), with mutations that are presumed to be neutral in tumor development. Compared to ERBB2 and VHL, OR2T33 has few mutations from tumors with low mutation rates (patients with fewer than 100 mutations total). Proteins are broken into 75 equal windows for plotting.

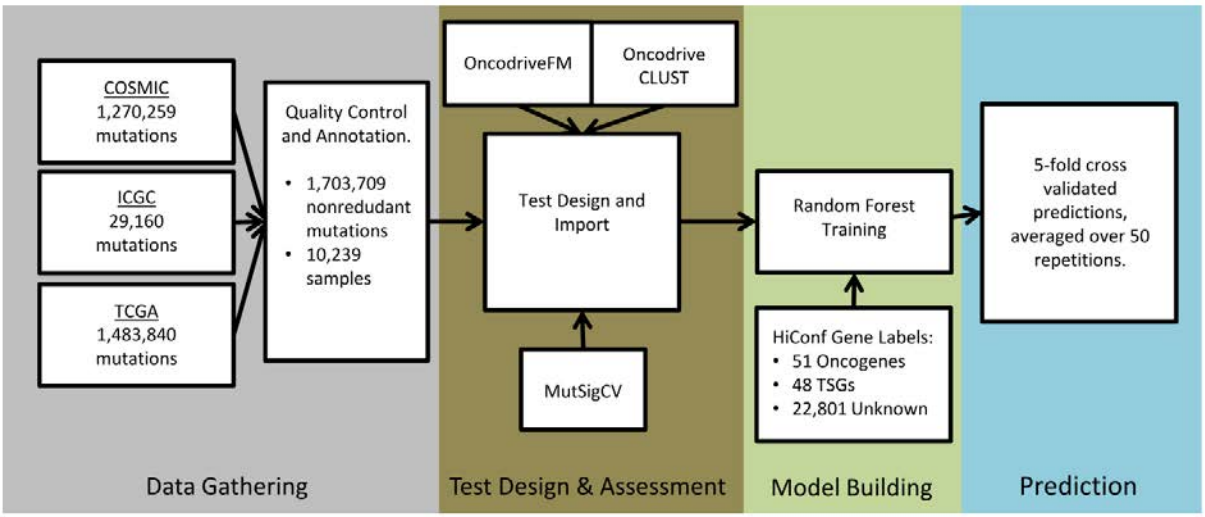


Figure S2.2. Cancer gene analysis overview. See section 2.2 for details.

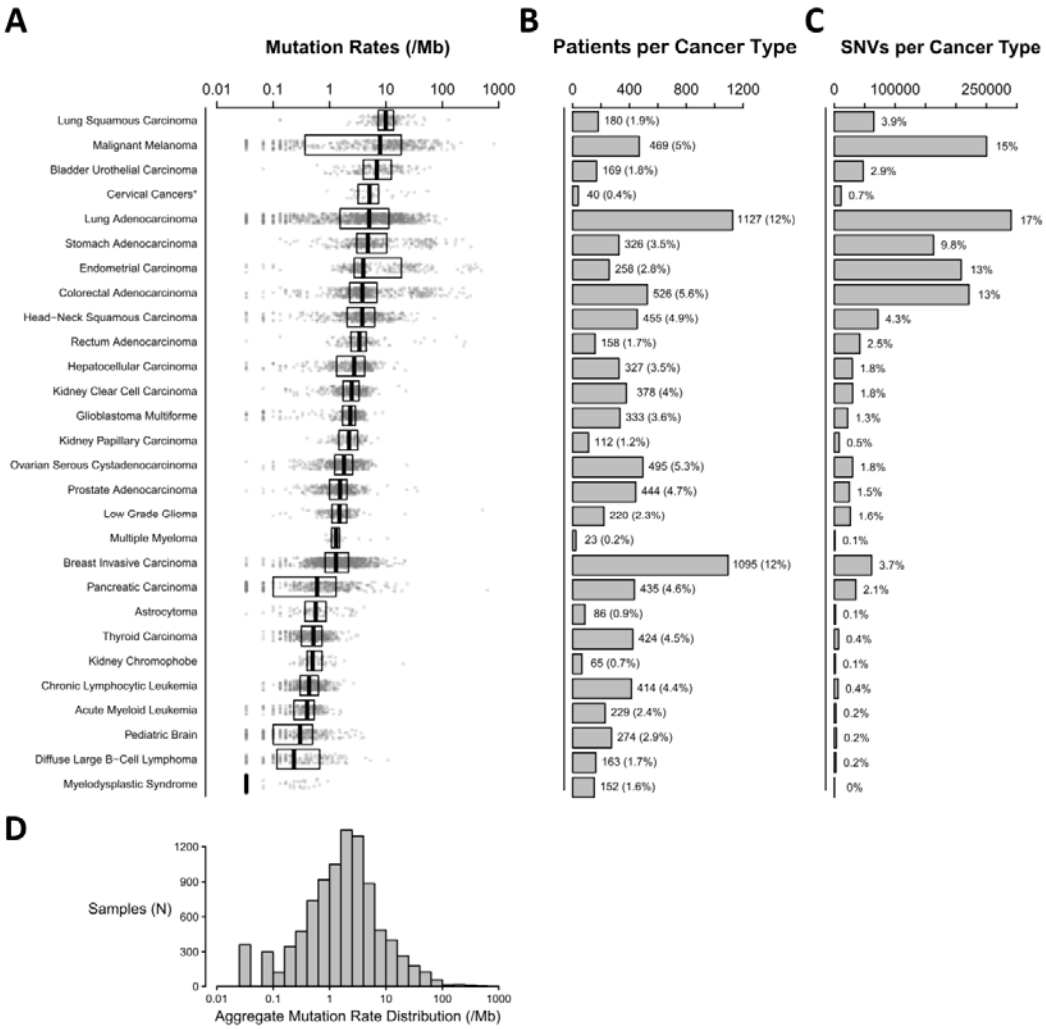


Figure S2.3. Pan-cancer dataset composition. 28 major cancer types accounting for 9,377 samples and 1,670,925 mutations are considered in the figure. A) Exome mutation rates (per Mb) are plotted against cancer type for each patient. Boxes enclose interquartile range and highlight the median. B) The number of patients per cancer type and percent contribution. C) The number of mutations per cancer type and percent contribution. D) Mutation rate (per Mb) distribution for all patients with the 28 listed cancer types. *Cervical Adenocarcinoma and Cervical Squamous Cell Carcinomas are included together as Cervical Cancers.

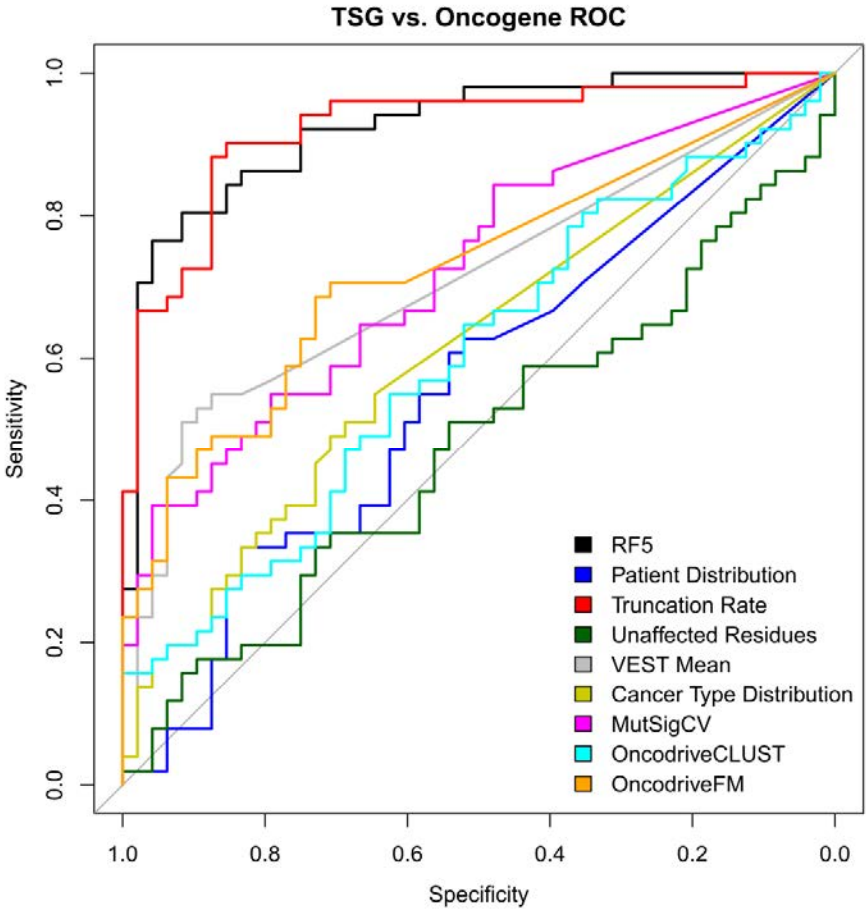


Figure S2.4. ROC curve for separation of HiConf Oncogenes and TSGs.

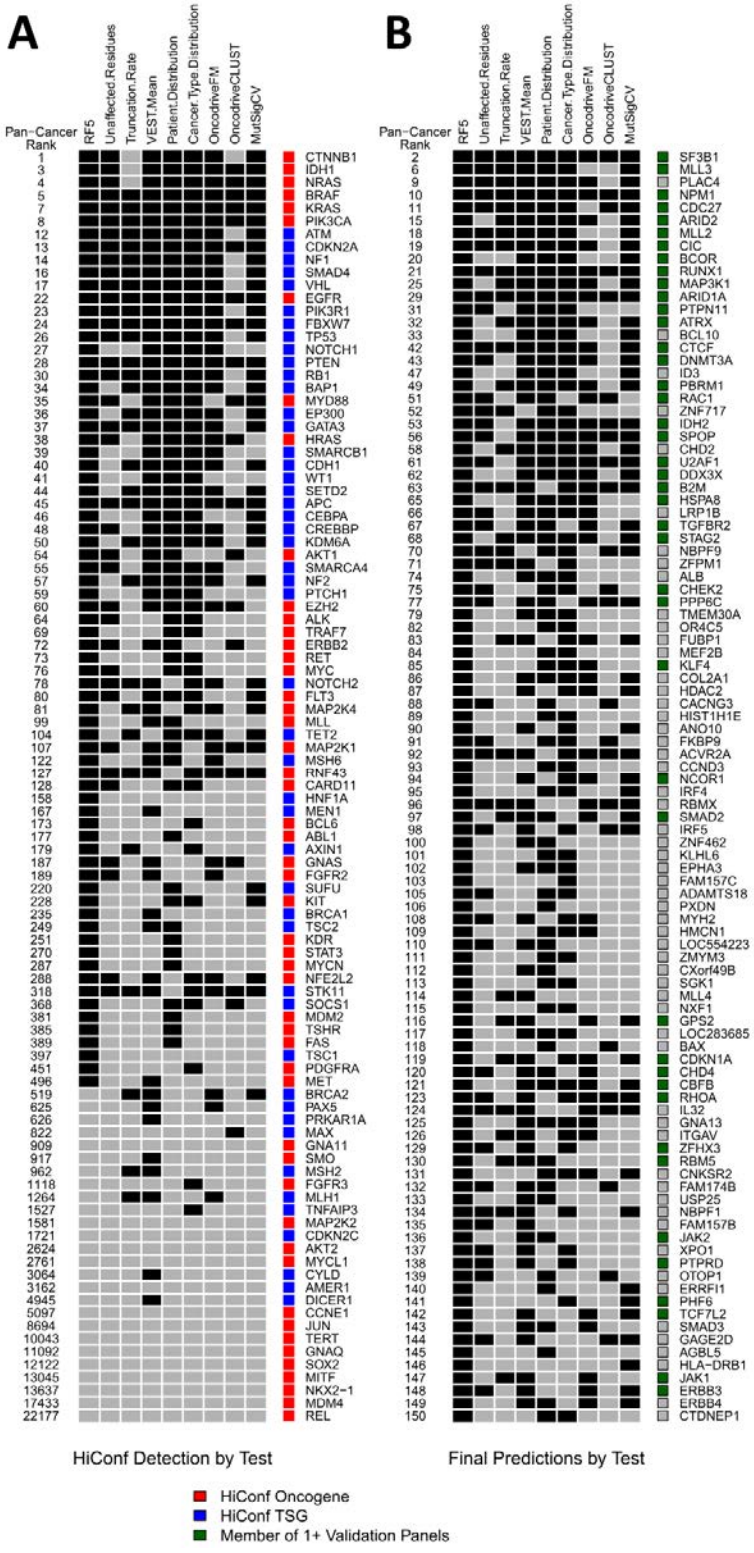


Figure S2.5. Pan-cancer ranking of genes by RF5. A) The detection of HiConf genes by RF5 and individual tests is indicated. Genes are considered detected if they are within the top 500 genes according to the indicated test. Genes are ordered according to the sum of their RF5 ONC and TSG scores. Overall RF5 rankings are shown to the left, and panel membership is indicated on the right. B) The top 100 predicted cancer genes from RF5 are shown, along with their detection by individual tests. Formatting is the same as in Panel A. *Note, *Truncation Rate* is calculated as a two sided test here.

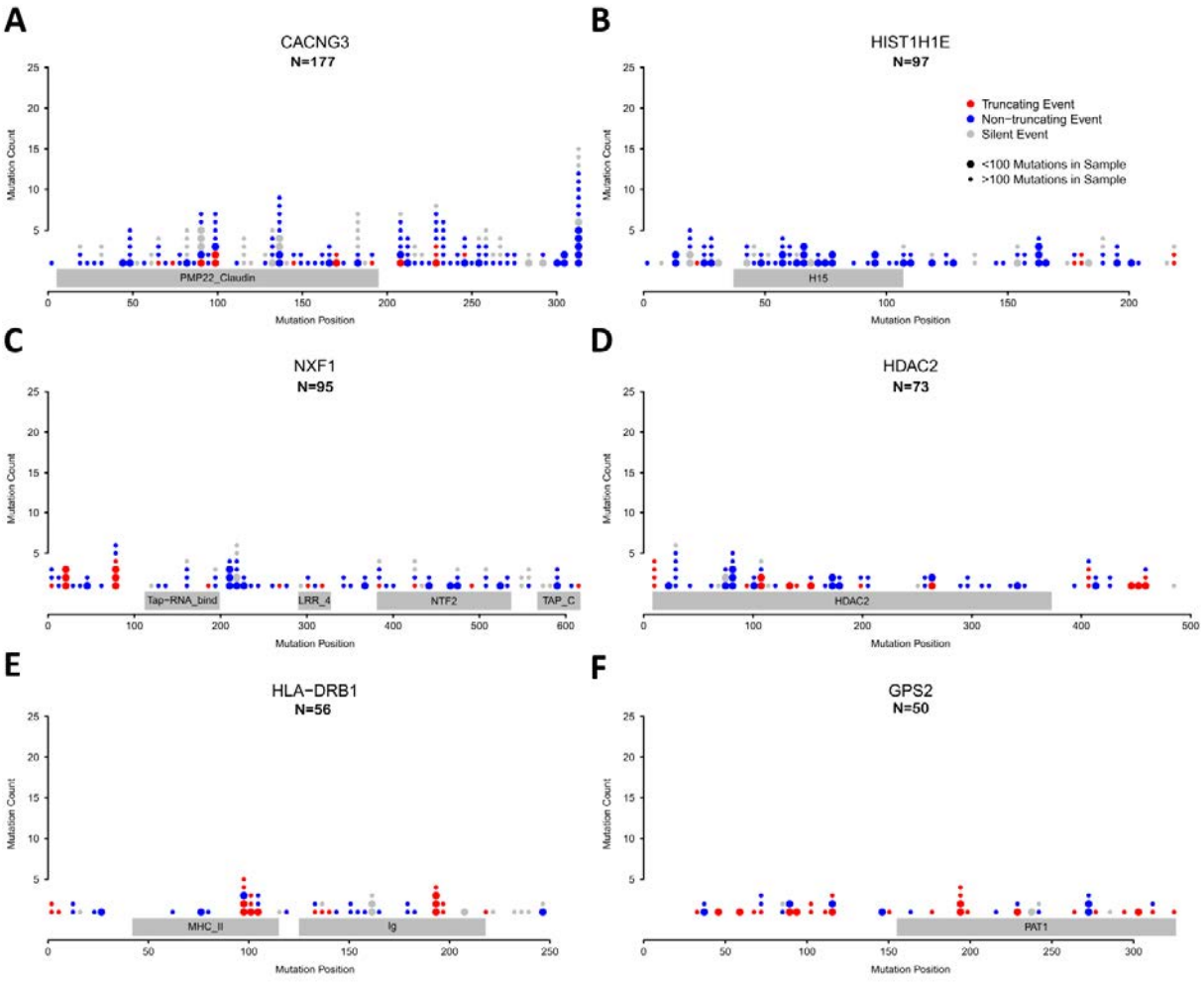


Figure S2.6. Mutation profiles of novel putative cancer genes.

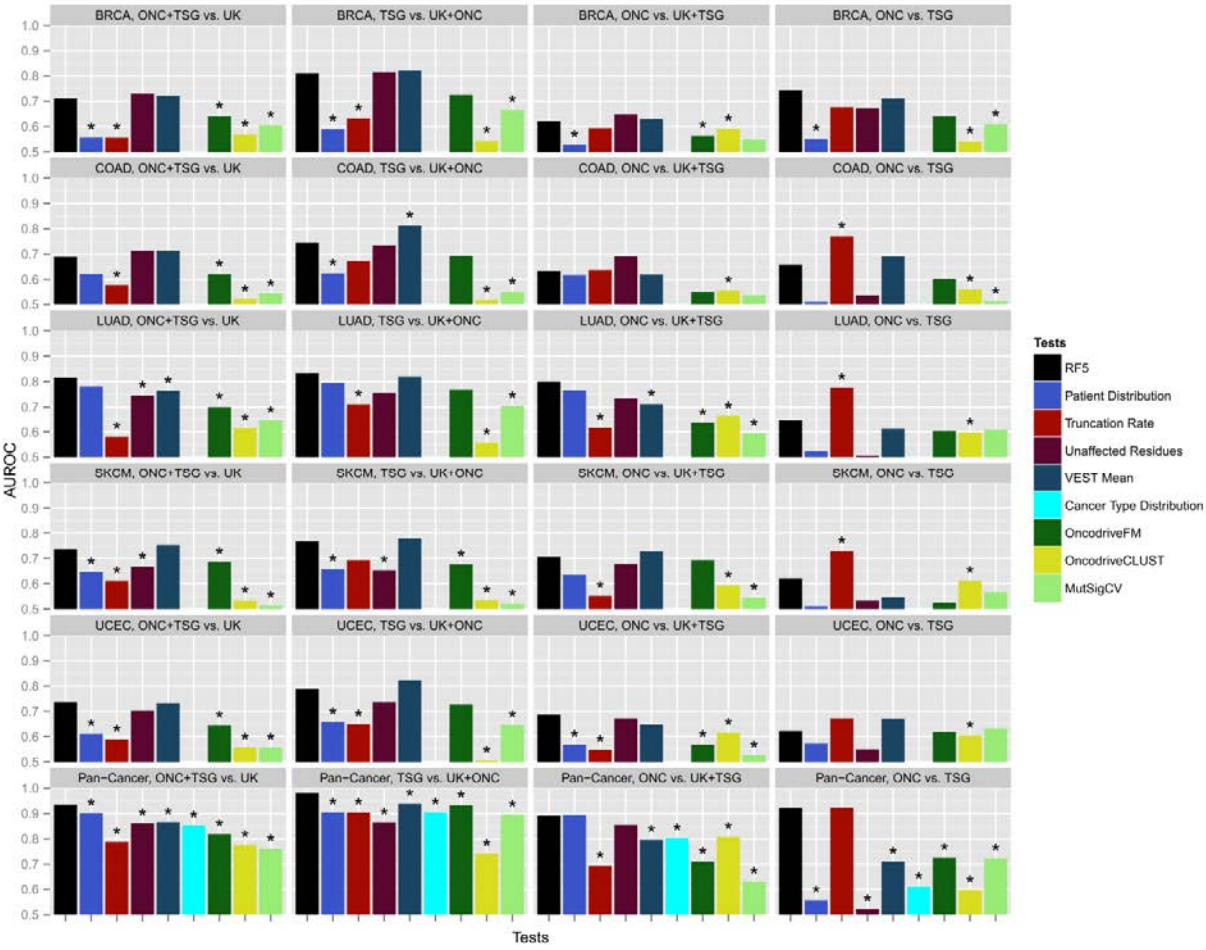


Figure S2.7. Performance in specific cancer types. The tests and performance statistics of Table 2.2 were calculated for each of 5 cancer types (BRCA = breast, COAD = colorectal, LUAD = lung adenocarcinoma, SKCM = melanoma, UCEC = endometrial carcinoma). Cancer types required at least 200,000 mutations or 500 patients to be included. RF5 models were trained for each cancer type. Tests which are significantly different from RF5 performance (DeLong Test $p<0.05$) are starred (*).

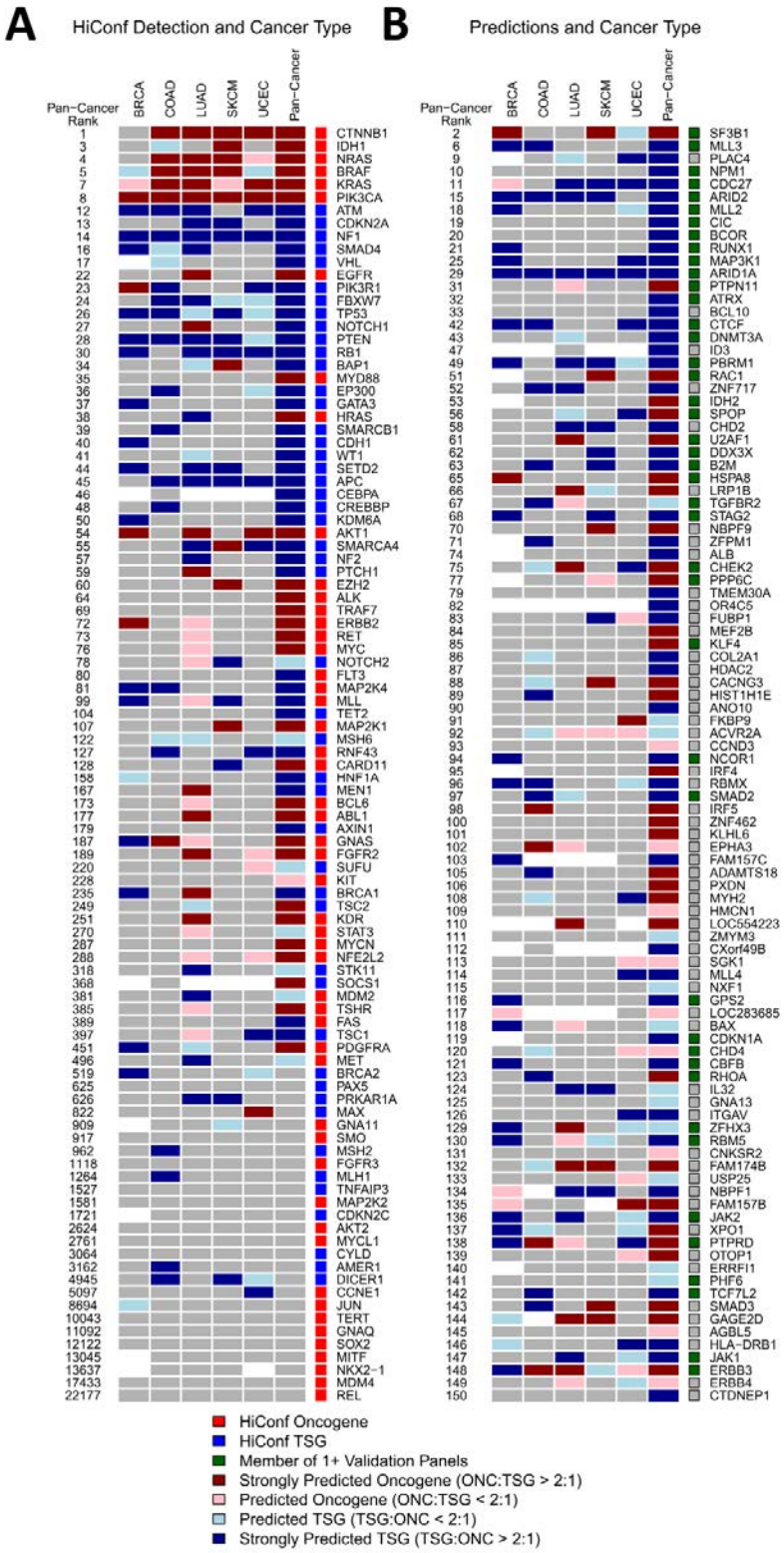
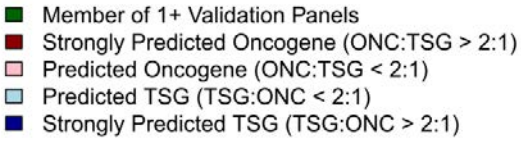
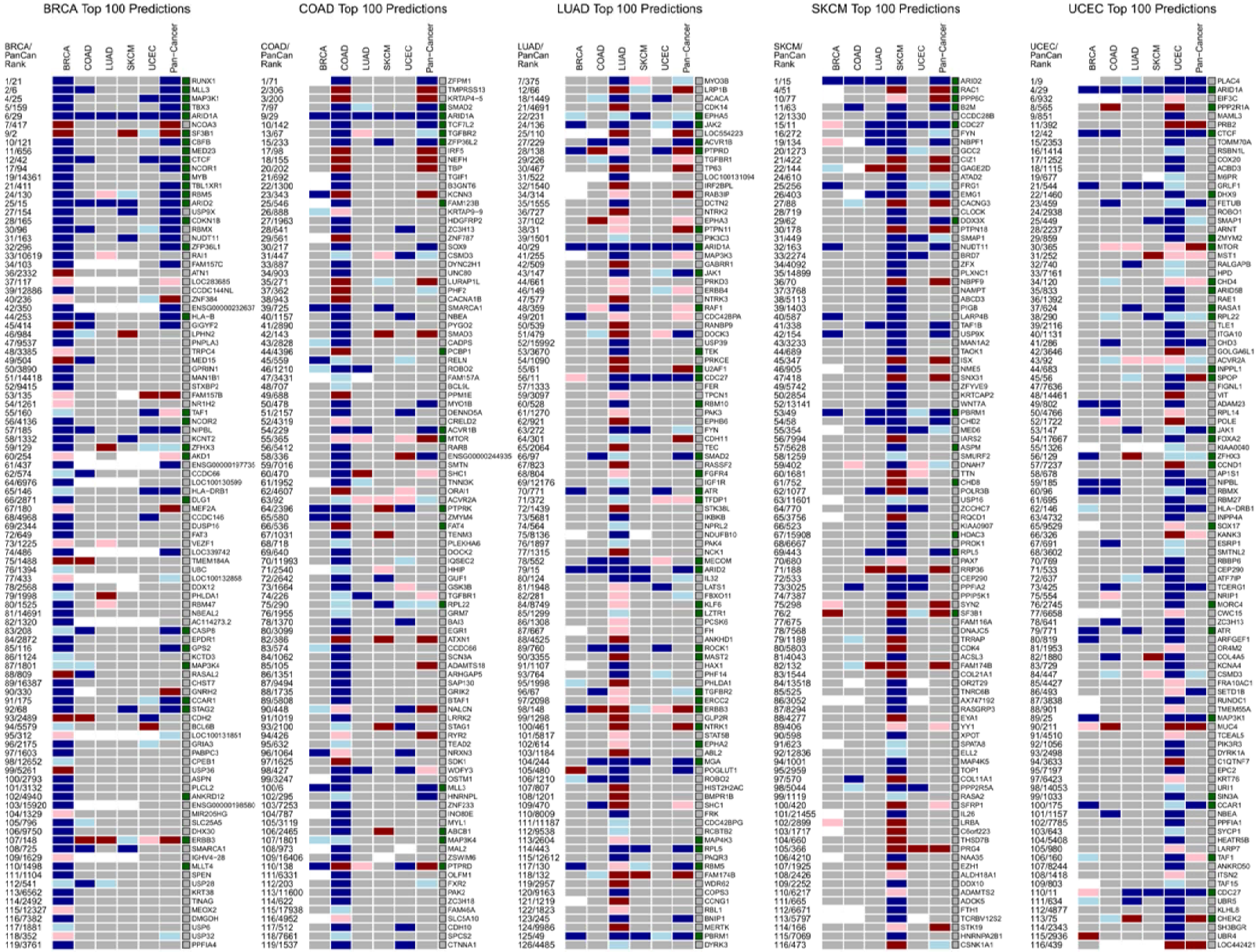


Figure S2.8. Detection of HiConf and top 100 pan-cancer predictions in specific cancers. A) The HiConf cancer genes are listed, and their detection by RF5 models trained in specific cancer types is indicated. Genes are detected if they are within the top 500 predictions in the indicated cancer type. Genes are ordered according to the sum of their Pan-Cancer RF5 ONC and TSG scores. Overall rankings are shown to the left, and panel membership is indicated on the right. Whitespace indicates that the gene was not mutated in the specified cancer type. The ratio of the RF5 ONC and TSG scores for the gene and cancer type define whether the gene was strongly identified (ratio > 2:1) as an oncogene or TSG. B) The top 100 Pan-Cancer RF5 predictions (excluding HiConf panel members) are listed and their detection in individual cancer types is indicated. Formatting is the same as in Panel A.

Figure S2.9. Cancer-specific predicted cancer genes. RF5 models were trained within 5 specific cancer types. The top 100 predictions (excluding HiConf panel members) from each cancer type are listed (from left to right, BRCA = breast, COAD = colorectal, LUAD = lung adenocarcinoma, SKCM = melanoma, UCEC = endometrial carcinoma), and their detection in other cancer types is indicated. Genes within the top 500 predictions for the cancer type are considered detected. Genes are ordered according to the sum of their RF5 ONC and TSG scores. Overall rankings are shown to the left for the specified cancer type and pan-cancer datasets, and panel membership is indicated on the right. Whitespace indicates that the gene was not mutated in the specified cancer type. The ratio of the RF5 ONC and TSG scores for the gene and cancer type define whether the gene was strongly identified (ratio > 2:1) as an oncogene or TSG.




BRCA Top 100 Predictions
COAD Top 100 Predictions
LUAD Top 100 Predictions
SKCM Top 100 Predictions
UCEC Top 100 Predictions
BRCA/ PanCan Rank
COAD/ PanCan Rank
LUAD/ PanCan Rank
SKCM/ PanCan Rank
UCEC/ PanCan Rank
BRCA
COAD
LUAD
SKCM
UCEC
Pan−Cancer

3. Identifying Drivers with Parsimony

This chapter is adapted from:

Kumar RD, Swamidass SJ, Bose R. (2016). Parsimony-guided prioritization of driver mutations within an expectation-maximization framework. Submitted.

3.1 Introduction

In this chapter, we present a new parsimony-guided paradigm for functional impact prediction and apply it to the problem of identifying cancer driver mutations. To do so, we introduce one new assumption: drivers are more equitably distributed among samples than passengers. Our goal is to use this knowledge to train a parsimonious model that predicts a few drivers in each patient. We first adapt an expectation-maximization (EM) framework to identify a parsimonious set of simple nucleotide polymorphisms that broadly explains cancer incidence in a training set of unlabeled pan-cancer mutations (ParsSNP). We then train a model to identify these putative drivers and detect similar mutations prospectively. This approach should be more generalizable than existing methods since it uses relatively simple assumptions and avoids the need for pre-labeled training data. Additionally, unlike most previous methods, our approach is applicable to all single nucleotide substitutions (including synonymous, nonsynonymous and premature stop mutations) and small frameshift and in-frame insertions/deletions (indels). We first characterize the process of training ParsSNP. We then use four classification tasks to assess the ability of ParsSNP and other independent tools to detect likely or known driver mutations in pan-cancer and other datasets. We also compare the predictions these tools produce in an independent cancer exome sequencing dataset, representing a typical usage scenario. Finally, we explore the specific mutations and genes that ParsSNP prioritizes in the pan-cancer dataset.

3.2 Materials and Methods

3.2.1 Data Gathering and Quality Control

We constructed the pan-cancer dataset in chapter 2 and a full description can be found there[96]. Mutation data was drawn from the TCGA, ICGC and COSMIC. Data was updated to build hg19 and duplicate data was deleted[81]. Mutations were annotated with ANNOVAR using RefSeq gene and ljb26 libraries[82]. The dataset contains 1,703,709 mutations drawn from 10,239 tumors representing 28 cancer types.

In keeping with established practice, we first removed potentially biologically distinct hypermutated samples[32, 45]. Since there is no universal cutoff for defining hypermutation[32, 56], we used the median mutation burden (715 mutations) to generate two equally sized segments that differ only by mutation rate: 435 samples with 851,996 mutations, and 9,804 samples with 851,713 mutations. The 9,804 non-hypermutated tumors were randomly split 2:1 to generate a 6,536 tumor (566,223 mutation) training dataset and a 3,268 tumor (285,490 mutation) test dataset.

We also apply our models to external data. We drew 2,314 mutations from the IARC R17 systematic P53 yeast screen collection as a benchmarking set[97]. Like Reva *et al*, we averaged the normalized scores of all eight downstream targets to reduce technical variation[98]. We also constructed the "driver-dbSNP" benchmarking dataset from several sources, consisting of: 49,880 common SNPs (minor allele frequency > 1% in human populations) from dbSNP build 142 as presumably non-functional germline mutations; 289 known activating kinase mutations from Kin-Driver[86]; and 849 known non-neutral mutations from Martelotto *et al*'s recent benchmarking study[99]. We also drew exome sequencing results from Kakiuchi *et al*'s study of 30 diffuse-type gastric carcinomas[100]. Once intergenic and intronic mutations were removed,

2,988 mutations remained in this dataset. All external data was re-annotated and treated the same as our pan-cancer datasets except where noted. Mutations that could not be annotated are excluded.

At several points in the analysis, we make use of the Cancer Gene Census, a curated list of mutations that are associated with cancer[83]. We further narrow this list with the approach used by Schroeder *et al*[56]. Specifically, we remove genes that have only been associated with translocation events, since our dataset does not contain similar events and many of these genes may not be directly associated with cancer. Similarly, we remove genes that have no recorded somatic mutations according to CGC annotations. This leaves 208 genes in the dataset. For the remainder of this chapter, when we refer to the CGC, we are referring to this reduced set of genes. Where appropriate, we further divide this set into putative oncogenes or tumor suppressors based on their annotated genetic profile (dominant or recessive, respectively)[56]. Genes with ambiguous profiles are excluded.

3.2.2 Mutation Level Descriptors

ParsSNP uses 20 mutation-level descriptors. Rather than directly train on functional, structural or evolutionary descriptors, ParsSNP incorporates such data indirectly by including 16 previous functional impact scores (FIS) from the ANNOVAR ljb26 libraries[82]. Details are available through ANNOVAR, but they include established tools such as SIFT, Polyphen2, MutationAssessor, FATHMM, VEST, and CADD. To these we added three additional mutation-level descriptors. Normalized Position is equal to the mutation position divided by the protein length. The Blossum62 score was assigned for amino acid substitutions. The final variable is Mutation Type, which is encoded as two descriptors (VarClassS, VarClassT) that indicate if the

mutation is silent (including synonymous, intronic and untranslated mutations) or truncating (including frameshift, splice site, nonstop and nonsense mutations).

3.2.3 Gene Level Descriptors

Four descriptors provide gene-level data. The first is protein length. The rest are drawn from the work presented in chapter 2[96]. *Unaffected Residues* tests for nonrandom mutation recurrence within the gene. *Truncation Rate* tests for enrichment or depletion of truncation events within a gene. Finally, *Cancer Type Distribution* tests for genes that are mutated in nonrandom subsets of cancers. We chose these three tests because they are non-redundant with the other information sources available to ParsSNP. These tests were calculated as outlined previously using the training dataset, and applied to additional datasets as annotations.

3.2.4 Imputation and Data Scaling

As our calculations are based on the configuration of mutations within tumors, we cannot simply remove mutations with missing data without removing whole samples and quickly depleting the dataset. Therefore we make use of data imputation at several levels.

Most important is the handling of non-missense mutations, to which many FISs do not apply. We adapted the strategy of OncodriveFM to impute these values[48]. We consider “truncations” (encompassing nonsense, nonstop, splice site, frameshift and inframe indels) as more likely to be drivers, while we consider “silent” mutations (including synonymous, intronic and untranslated mutations) as less likely. For 9/16 FISs, a classification as functional or neutral is made based on thresholds provided by the original authors[82]. For each of these impact scores, truncation events with missing values were assigned the average value given to predicted functional mutations. Similarly, silent mutations with missing values were assigned the average value of neutral missense mutations. For tools that had them, intermediate classes were deemed

functional. For the seven scores without classification schemes, we used the 95^{th} and 5^{th} percentiles as imputation values for the truncation and silent mutation classes, respectively. ParsSNP results are robust to reasonable changes in the percentiles used.

Remaining missing values are then replaced by mean imputation. Only one of the descriptors had more than 5% missingness in the training set, and none were greater than 10%. Finally, the training set is scaled so that each descriptor is in the range of[0,1], in keeping with best practice for neural network models[101]. Wherever applicable, descriptors, imputation and scaling values were calculated using the training set and applied to other datasets.

3.2.5 Adapting the Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm can fit statistical models with missing or latent data[65]. However, it requires that constraints be placed on the possible solutions. For instance, Zaretzki *et al* used the EM algorithm to predict atomic sites of P450 metabolism using region-level data, but constrained the number of metabolic sites per region[102]. In ParsSNP's learning phase, the missing data is the status of mutations as drivers or passengers, which is constrained so that 1) drivers are relatively equitably distributed among samples, and 2) they are consistent with the descriptors. The E-step will use the first constraint, while the M-step uses the second.

3.2.6 Learning Initialization

ParsSNP begins with descriptors for an unlabeled training set of mutations (X matrix), with each mutation belonging to a biologic sample. ParsSNP uses EM to find a set of labels for the training mutations; more precisely, as a probabilistic model, ParsSNP finds a set of probabilities that describe the unseen, binary driver/passenger labels (we refer to these probabilities as 'ParsSNP labels' in the main text). We initialize ParsSNP with a random uniform vector of probabilities. Samples are assigned into equally sized folds that will be used during the M-step throughout the

training process, such that mutations never inform their own updates. The probabilities are then iteratively refined by the E and M-steps until they stabilize.

3.2.7 The E-step

The E-step updates probabilities based on the combination of mutations within samples. The probabilities for a given sample (*Y*) are updated under the belief that the unseen total number of driver mutations in the sample (*t)* is between a lower (*l*) and upper (*u*) bound. Each probability is updated based on the following question: of all the possible configurations of driver/passenger binary labels for the sample in which *t* is between *l* and *u*, what proportion require the mutation be a driver (weighted by probability)?

This can be formalized using Bayes' Law and some additional definitions. *M* is the unseen vector of binary mutation labels which sum to *t* in each sample, while *m* is the unseen label of the given mutation. *Y* is the current vector of probabilities, while *y* is the probability of the given mutation. We denote values that exclude the mutation using prime notation. Given these definitions, the value of *y* can be updated with the following equation:

$$y_{new} = \mathrm{E}[m \mid l \leq t \leq u] = y * \mathrm{P}(l\text{-}1 \leq t' \leq u\text{-}1) \,/\, \mathrm{P}(l \leq t \leq u)$$

While we cannot calculate *t* and *t'* directly, our beliefs regarding the mutation labels *M* and *M'* are described by *Y* and *Y'*. Therefore *t* and *t'* can be treated as poisson binomial random variables parameterized by *Y* and *Y'*[103]. For each sample in the dataset, Bayes' Law is applied to each mutation. The poisson binomial cumulative density function is calculated exactly in samples with fewer than 30 mutations and with a refined normal approximation for samples with more[103].

The most important parameters in the E-step are the lower and upper bounds. The lower bound is the simpler of the two and is fixed at one driver per sample. A higher value lacks biological justification in cancer, since tumors have been observed with no exomic mutations[11]. However, setting the lower bound to zero leads the algorithm to converge on a vector of very low values; this is consistent with the constraint, but non-informative (Supplementary Figure S3.1).

The upper bound is more complex as the algorithm makes use of two versions. The 'fixed' upper bound is defined as $\log_2$[mutation burden]. Therefore, a sample with 8 mutations is believed to have between 1 and 3 drivers, while a sample with 1024 is believed to have between 1 and 10 drivers; these ranges illustrate how drivers are presumed to be more equitably distributed than passengers. Using $\log_2$ as a function yields reasonable upper bounds over the range of mutation burdens; however, its stringency can cause underflow errors even with double precision arithmetic. The problem most often occurs in very mutated samples in early iterations. Therefore, we define a second, less stringent 'sliding' upper bound that is often used initially and is set to some proportion (p) of the total current belief for the sample (default p=0.9). The E-step uses whichever is greater of the sliding and fixed upper bounds. For instance, a sample with 1024 mutations is initialized with a random vector of probabilities, and is therefore currently expected to have ~500 functional mutations on average. The sliding upper bound is p*500=450, while the fixed upper bound is $\log_2(1024)$=10. Given the current probabilities, the probability of the total number of drivers being less than 10 is essentially zero, requiring the use of the sliding upper bound. The sliding upper bound applies a consistent, downward pressure on probabilities until the fixed upper bound can be used without risking underflow errors. The algorithm is robust to reasonable alternatives for defining both upper bounds (Supplementary Figure S3.1).

These concepts are clearer when viewing the E-step code (see section 3.2.13). An important point should be made here however: *the upper and lower bounds are "soft" bounds. In the final solution, many samples will be assigned a number of drivers beyond the suggested bounds.* While these bounds are important to understanding the behavior of the E-step, in practice any reasonable function for defining these values leads to similar results.

3.2.8 The M-step

The M-step follows a standard machine-learning approach. Samples are assigned to one of five cross-validation folds, which are fixed so that samples never inform their own updates. ParsSNP was robust to changes in the number of folds (Supplementary Figure S3.1). We fit a single layer neural network to the data, which is capable of producing well-scaled outputs that can be interpreted as probabilities by the E-step and has been used for this reason previously[102]. The neural network parameters are set by grid search (weight decay[0.1, 0.01, 0.001], size[6, 12, 18]) through bootstrap selection (10 samples of 10,000 mutations each). The cross-validated predicted probabilities are then returned and passed to the E-step if the stop criteria are not satisfied.

3.2.9 Algorithm Stop and Model Training

The algorithm ends once the vector of mutation probabilities stabilizes. We define stabilization as a mean-square difference (MSD) between iterations of less than 1e-5, or a change in MSD of less than 5% between iterations. In practice, the second relative cut-off was invoked more often than the absolute cut-off, but the algorithm consistently converged by 30 iterations even under highly stringent cut-offs.

Once the algorithm ends, we have a vector of probabilities for the training data that optimally meets our constraints, and are interpreted by ParsSNP as the probability of each mutation acting as a driver. These probabilities are the 'ParsSNP labels' which we refer to in the main text. As

the cross-validation approach we use is non-deterministic, we run the algorithm 50 times and average the final outputs to generate a final result (ParsLR was trained with 50 runs, ParsFIS and ParsNGene were trained with 10 runs, and cancer-specific models were trained with 5 runs, see section 3.2.10 for details). The final ParsSNP neural network model is trained with the labels and descriptors using identical settings as the M-step. Note that we could use a different machine learner or even different descriptors at this stage: the EM component of ParsSNP has generated a set of probabilistic labels for the training data, and now the question is one of modeling. The final model is then applied to the various datasets to produce ParsSNP scores.

3.2.10 Methodological Controls

Since ParsSNP consists of several components, we include several variations on ParsSNP for comparison. ParsLR uses logistic regression rather than a neural network. ParsFIS uses only the 16 descriptor FISs, while ParsNGene only excludes the three gene-level descriptors described in chapter 2 (*Truncation Rate, Unaffected Residues, Cancer Type Distribution*)[96]. We also include supervised versions of ParsSNP, which use neural networks and all descriptors but are trained to perform each task directly. The recurrence-trained and CGC-trained models are trained in the pan-cancer training set, and assessed in the test set, with labels defined as they were for the primary ParsSNP model. The driver-dbSNP- and P53-trained models were trained and tested using 10-fold cross validation directly in the corresponding datasets, since we had insufficient data for separate training/test sets in these cases. The model type and tuning procedure is identical to that used in the M-step of ParsSNP's training. Since ParsSNP incorporates several gene-level descriptors, we ran gene-level tools which are designed to detect cancer genes on the pan-cancer training set and then assessed their performance in the test set (MutSigCV,

OncodriveFM, and OncodriveCLUST[45, 48, 49]. We also consider CGC membership as a simple approach for defining drivers.

3.2.11 AUROCs for Measuring Performance

For comparing ParsSNP with alternate strategies, we follow the strategy of Carter *et al* and use the area-under-receiver-operator-characteristic (AUROC)[63]. The receiver-operator-characteristic is a curve constructed from the sensitivity and specificity of a classifier at each possible threshold. The area under this curve summarizes classifier performance: a classifier which can achieve perfect sensitivity and specificity simultaneously has AUROC=1, while random guesses should produce AUROC=0.5. AUROCs are advantageous because they encompasses all thresholds simultaneously, while remaining statistically testable[87]. This maximizes comparability between ParsSNP and the independent tools, which often do not have fixed thresholds (e.g. CHASM), or have recommended thresholds that are meant to optimize performance in datasets that are different from ours. It is important to note that we have not used thresholded predictions from the independent tools: unless otherwise noted, all tools are assessed based on their raw scores, even when the original authors provide thresholds. In practice of course, users may want to apply thresholds to ParsSNP to improve interpretability. We recommend that thresholds be set in a context specific manner, taking into account the relative need for sensitivity and specificity, and the relative costs of false positives or negatives.

3.2.12 Statistics and Software

ROC curves were compared using Delong tests for correlated or paired data[87]. Gene-to-gene comparisons were made using Wilcoxon one- or two-sample tests. All tests were two-sided unless otherwise noted. Multiple comparisons were Bonferroni corrected unless otherwise noted.

All analyses and calculations were performed in 64-bit R version 3.1 using double precision arithmetic. Poisson binomial distributions were calculated with the 'poibin' R package[103], while neural networks were fitted and tuned using functions from the 'nnet' and 'e1071' packages[104]. ROC analysis were performed with the 'pROC' package[105].

CanDrA, CHASM, FATHMM Cancer, TransFIC and Condel, MutSigCV and OncodriveCLUST were applied to our datasets using the software made available through the original publications. Oncodrive-fm was re-implemented in R according to the protocol in the original publication.

3.2.13 Code Availability & URLs

The software and datasets required to replicate this analysis are available at github.com/Bose-Lab/ParsSNP. Annovar, http://annovar.openbioinformatics.org; Cancer Gene Census, http://cancer.sanger.ac.uk/census/; CanDrA, http://bioinformatics.mdanderson.org/main/CanDrA; CHASM, http://www.cravat.us/; FATHMM Cancer, http://fathmm.biocompute.org.uk/cancer.html; TransFIC, http://bg.upf.edu/transfic/home; Condel, http://bg.upf.edu/fannsdb/; MutSigCV, https://www.broadinstitute.org/cancer/cga/mutsig; OncodriveCLUST, https://bitbucket.org/bbglab/oncodriveclust/get/0.4.1.tar.gz.

3.3 Results

3.3.1 ParsSNP overview

ParsSNP identifies likely drivers using a training set of unlabeled mutations from a collection of biological samples and two constraints. First, predicted drivers should be few in number and distributed relatively equitably among samples. Second, predicted drivers must be identifiable using the descriptors. Figure 3.1A provides an overview of ParsSNP, with details in section 3.2). There is a learning and application phase.

In the learning phase, ParsSNP generates probabilistic driver labels for the training mutations. The labels are initialized with random values from 0 to 1; they are then iteratively refined by expectation-maximization (EM), each step representing a constraint. In the expectation (E) step, each label is updated using Bayes Law and the belief that in a sample with *N* mutations, between 1 and $\log_2(N)$ mutations drive tumor growth . Since this range scales logarithmically, the E-step ensures that predicted drivers are uncommon and relatively equitably distributed among samples. The maximization (M) step builds a probabilistic model and updates the labels using the descriptors in cross-validation, ensuring that predicted drivers can always be defined in terms of the descriptors. We use a neural network, since this model produces well scaled probabilities[102]. The E-step and M-step iterate until convergence.

In the application phase, the refined labels are used to train a final ParsSNP neural network. For clarity, we differentiate between the probabilistic "ParsSNP labels" produced by EM for the training data, and the "ParsSNP scores", which are defined by the model in all datasets (Figure 3.1A).

3.3.2 Datasets & Analysis Design

Our pan-cancer dataset consists of 1,703,709 protein-coding somatic mutations from 10,239 samples[96], broken into three partitions: a 435 sample (851,996 mutation) "hypermutator" set; a 6,536 sample (566,223 mutation) "training" set; and a 3,268 sample (285,490 mutation) "test" set (see section 3.2.1). We also use experimental and germline data. The "driver-dbSNP" dataset has 49,880 common variants from dbSNP plus 1,138 experimentally validated drivers from Kin-Driver[86] and Martelotto *et al*[99], while the "P53" dataset has 2,314 mutations from the IARC R17 P53 yeast screen[97]. Finally, we assess ParsSNP in a typical usage case with the Kakiuchi *et al* dataset of 30 diffuse-type gastric carcinomas[100]. We also make use of the Cancer Gene

Census, excluding genes that are only involved in translocations (see section 3.2.1)[56, 83]. We use 23 descriptors, including 16 published FISs, plus three mutation-level and four gene-level annotations (see sections 3.2.2 and 3.2.3).

Since no "gold standard" for cancer drivers exists, we use the above datasets to assess ParsSNP in four classification tasks, each designed to represent qualities of driver mutations (see Table 3.1). The first is detecting recurrent mutations that occur in two or more samples of the test set. This is an important task, since recurrent mutations are often considered as driver proxies[62]. The second task is identifying test set mutations that are within CGC members, since all else equal, mutations in known cancer gene are more likely drivers. Since cancer genes are enriched in recurrent mutations, we limit this task to only non-recurrent mutations to avoid redundancy. The third task is identifying experimentally defined drivers among common (and presumably neutral) germline variants. The driver-dbSNP dataset is meant to resemble real world data, where mutations are drawn from many genes and only a few mutations are drivers. The final task is identifying disruptive events (mutant activity is <25% of wild type activity) over non-disruptive events in P53. This classification task is crucial, since many applications will focus on experimental data in one or a few genes.

Our primary performance measure is the area-under-receiver-operator-characteristic (AUROC). AUROCs are useful because they summarize performance across all prediction thresholds and are statistically testable, and have been used for these reasons previously[63]. These values are equivalent to the accuracy of a tool when sorting random pairs consisting of one driver and one passenger, as defined in each task (1-AUROC is the corresponding error rate of the tool). They can range from 0.5 (equivalent to classifying mutations by guessing) to 1.0 (perfect classification accuracy with no errors; see section 3.2.11 for details).

For each task, we compare ParsSNP to competing methods that are designed to detect cancer drivers but were not used as descriptors. These "independent tools" include CanDrA, a supervised ensemble method trained to detect recurrent mutations in pan-cancer data[62]; CHASM, a supervised model that is trained using curated cancer mutations[63]; FATHMM Cancer, a Hidden-Markov-Model based approach[106]; TransFIC, a method of recalibrating FISs for cancer data (the base score is MutationAssessor, which had the best performance in the original study)[107]; and Condel, an ensemble method that was not designed specifically for cancer, but was shown by its authors to be useful for detecting drivers[108]. Each tool was applied to the datasets using the published software (see section 3.2.13). Matching or improving upon the performance of these tools will demonstrate the value of ParsSNP as a method for detecting driver mutations.

3.3.3 ParsSNP Training, Robustness and Performance

An EM approach requires careful empirical testing to ensure it returns an appropriate result. ParsSNP consistently converged within 15-20 iterations (Supplementary Figure S3.2A). It was highly reproducible, with an average pairwise correlation of 0.99 over 50 runs. Though small, the variations lead us to average the 50 runs for the final labels. They are right-skewed as expected, suggesting a minority of mutations as drivers (Figure 3.1B). After training the final ParsSNP model using these labels, we assessed the contribution of descriptors to the neural network using Garson's algorithm (as described by Olden *et al*[109], Figure 3.1C). We find that all descriptors make at least moderate contributions to the model.

ParsSNP is trained such that putative driver mutations should be distributed relatively equitably among samples (*i.e.* enriched in the least mutated samples, and depleted in hypermutators on a per-mutation basis). We tested to ensure ParsSNP scored mutations in this way (Figure 3.1D).

Mutations from less mutated tumors were more likely to be identified as drivers regardless of ParsSNP threshold; this pattern continues into the hypermutator set, which was not used for training. On average, hypermutators have 23-times more mutations than nonhypermutators (1,958:86.6), but only 5.5 times more mutations with ParsSNP scores over 0.1 (7.17:1.3). At a very stringent cutoff of 0.5, the ratio is only 3.3 (0.23:0.07). Therefore, ParsSNP assigns putative drivers relatively evenly in both the training and held-out hypermutator sets, suggesting that the parsimony-guided training worked as intended.

Both the training data and several tunable parameters may affect algorithm behavior. To test consistency across datasets, we split the training data into two equal halves and found that ParsSNP produced highly correlated scores (r=0.96, Supplementary Figure S3.2B). We also compared the results of fifteen alternative parameters settings (Supplementary Figure S3.1), definitions in sections 3.2.5-3.2.8). The algorithm consistently converged and usually produced labels that were highly correlated with the reference; even when results differed, the overall ordering of mutations was highly consistent, suggesting that ParsSNP will identify a consistent set of putative drivers over reasonable parameters and data.

To help understand ParsSNP's performance, we applied several methodological controls to the classification tasks in addition to the primary ParsSNP model. These controls include versions of ParsSNP that use a simplified model (logistic regression), versions that lack gene-level descriptors, stand-alone gene-level tools, and models that are explicitly trained to perform the tasks through supervised learning rather than the unsupervised EM training (Supplementary Figures S3.3-S3.6, see section 3.2.10 for model descriptions). We found that using logistic regression rather than a neural network slightly degraded ParsSNP's performance in most tasks. We observed that gene-level tools generally do not perform well in the tasks when used in

isolation; however, removing the gene-level descriptors from ParsSNP does markedly degrade performance in most tasks. We also found that supervised learning is not as effective as the unsupervised EM training in 12 of 16 comparisons. As expected, supervised models often performed well at the tasks they were trained to perform, but unlike ParsSNP their performance was inconsistent in other tasks. We conclude that the most important source of ParsSNP's performance is the combination of the novel feature set (particularly the inclusion of gene-level features) with the unsupervised EM training. The following sections explore ParsSNP's performance in the classification tasks in-depth.

3.3.4 Testing ParsSNP with Pan-Cancer Data

ParsSNP was applied to the withheld test dataset of 3,268 pan-cancer tumors (285,490 mutations). Because the independent tools do not apply to synonymous and truncating mutations, the analysis was limited to 182,483 missense mutations, except where noted.

The first classification task we considered was identifying recurrent mutations, since they are often treated as drivers[62]. ParsSNP scores are positively correlated with mutation recurrence in the test set (Figure 3.2A), and overall are more highly associated with recurrence than any of the independent tools (Figure 3.2B). Overall, ParsSNP identified 9,434 recurrent missense mutations with AUROC=0.656 (95% CI 0.650-0.663), better than any independent tool (Figure 3.2C, Table 3.1, all Delong tests $p<2.2e-16$). CanDrA was the next best (AUROC=0.608, 95% CI 0.601-0.615), which is not surprising considering it was trained with recurrent mutations[62]. Recurrence also provides an opportunity to assess ParsSNP performance within genes, which is crucial since ParsSNP uses gene-level descriptors and 75% of its variation is between genes (based on sums-of-squares). We found that ParsSNP generally performed as well in single genes as it did in the entire dataset (Figure 3.2D).

The second classification task was identifying mutations in one of 208 CGC members, which should be enriched in driver events. We limited scope to non-recurrent missense mutations to avoid confounding with the prior analysis. ParsSNP identifies the 3,760 non-recurrent cancer gene mutations with AUROC=0.833 (95% CI 0.825-0.841), better than the independent tools (Figure 3.2E, Table 3.1, all Delong tests $p<2.2e-16$).

Unlike the independent tools, ParsSNP generates scores for non-missense mutations. It is biologically intuitive that there will be an interaction between mutation type and gene type in predicting drivers: we expect that truncations (frameshifts, premature stops) are less likely to be drivers than missense mutations when present in oncogenes, while the opposite is true in tumor suppressor genes (TSG), and silent mutations are unlikely to be drivers regardless of gene type[44]. We split CGC members into putative oncogenes and TSGs (see section 3.2.1) and found that ParsSNP was able to identify precisely the expected pattern (Figure 3.2F). We questioned which descriptors were responsible, since ParsSNP is not directly aware of gene type. *Truncation Rate*, which separates oncogenes and TSGs based on rates of truncation events [96], showed a marked interaction with mutation type (Supplementary Figure S3.7A). The ability to detect interactions between the descriptors illustrates the value of using a neural network rather than a simpler model (Supplementary Figure S3.7B).

3.3.5 Testing ParsSNP with Experimental Data

In our third classification task, we assessed ParsSNP's performance in the driver-dbSNP dataset, which combines a large number of (presumably neutral) common germline variants with relatively few experimentally validated drivers. 13,738 genes are mutated at least once in the dataset, and 49 have at least one functional mutation. ParsSNP detected the drivers with AUROC=0.975 (95% CI 0.970-0.981, Figure 3.3A), slightly better than FATHMM Cancer

(Table 3.1, Delong test p=0.205) and significantly better than the other independent tools (all Delong tests p<1e-4). We also performed a precision-recall analysis, which suggested that CanDrA, ParsSNP and FATHMM Cancer were the best performers, with area-under-precision-recalls (AUPRs) of 0.84, 0.83 and 0.83 respectively (Figure 3.3B). We conclude that ParsSNP prioritizes rare drivers over common germline variants better than existing tools.

The fourth classification task focuses on the IARC P53 dataset, which consists of P53 transactivation activity against downstream targets for 2,314 missense mutations[97]. Like many FISs, ParsSNP ascribes higher scores to mutations that abrogate P53 activity (Figure 3.3C). However, ParsSNP is more strongly associated with the P53 fold activity change than any independent tool (Figure 3.3D). ParsSNP is a strong performer when identifying the 475 mutations that reduce P53 activity to 25% or less of wild type activity, though it was statistically tied with CHASM (Delong test p=0.39) and Condel (Delong test p=0.59, Table 3.1, Figure 3.3E).

3.3.6 Summary of ParsSNP Performance

As Table 3.1 shows, ParsSNP outperforms most or all of the independent tools in every classification task. Out of the 20 comparisons, ParsSNP outperformed existing methods 19 times, 17 of which were statistically significant. The three statistical ties are: ParsSNP and FATHMM Cancer in the driver-dbSNP task; ParsSNP and CHASM in the P53 task; and ParsSNP and Condel in the P53 task. Importantly, we note that there is no single tool that can act as an alternative to ParsSNP across all tasks. We anticipate that ParsSNP will be applied to diverse data that will resemble the classification tasks to varying degrees. Therefore, the fact that ParsSNP performs very well against existing methods in all tasks is an extremely important

finding, since it suggests that ParsSNP's performance relative to other tools will be consistent in novel datasets.

Moreover, as a summary of accuracy, seemingly modest differences in AUROCs can imply large performance gains under particular conditions. For instance, several tools perform well in the driver-dbSNP task, often with AUROCs over 0.90. However, since AUROCs of 1.0 represent perfect accuracy with no errors, even small gains represent large drops in the AUROC error rate: ParsSNP's performance in this task (AUROC=0.975) represents more than a two-fold reduction in errors when compared to CHASM (AUROC=0.948), which is the most cited of the independent tools. Another valuable consideration is the precision (positive-predictive-value) if only a few predictions can be tested. For instance, ParsSNP and CanDrA had the top overall performance in the Recurrence task (AUROCs=0.656 and 0.608, respectively). The difference between these tools is emphasized when considering just the top 100 candidate drivers identified by each. At this threshold, ParsSNP has a precision of 98% (2 mutations of 100 are false positives), while CanDrA has a precision of only 84% (16 false positives), an 8-fold increase in false positives. These examples illustrate how dramatically ParsSNP reduces errors compared to other methods under typical conditions.

3.3.7 Application of ParsSNP to an Independent Dataset

We illustrate the advantages of ParsSNP in a typical usage scenario by applying it to recent data from Kakiuchi *et al*. This dataset contains 2,988 protein-coding mutations from 30 exome-sequenced diffuse-type gastric carcinoma patients[100]. For comparison, we also apply CanDrA and CHASM, the most recently published and the most cited of the independent tools, respectively. We define candidate drivers as those with scores in the top one percent for each tool (30 candidates per tool).

We compared the candidate drivers generated by these tools (Figure 3.4). We found that approximately one third (11 out of 30) of ParsSNP's candidate drivers overlap with other methods. These candidate drivers include missense mutations in well-established cancer genes including PIK3CA, FGFR2 and P53 (gene symbol: TP53). Two thirds of ParsSNP's candidate drivers (19 out of 30) were not identified by other tools at this cut-off. These mutations include truncations in known TSGs (ARID1A, CDKN2A, SMAD4 and P53)[96, 110], recurrent mutations (CDC27 Y173S), and confirmed drivers (RHOA Y42C)[100]. For comparison, mutations uniquely identified by CanDrA included biologically implausible drivers such as mutations in titin and dystrophin (TTN and DMD), which are very large skeletal muscle proteins[45], and mutations which have been experimentally confirmed as functionally neutral (ERBB2 R678Q)[40]. Mutations uniquely identified by CHASM were frequently in genes with no known or suspected role in cancer development, illustrated by the fact that only 2 out of these 27 mutations are present in members of the CGC[83]. These results show that ParsSNP identifies many likely drivers that other tools do not detect; furthermore, many of the mutations identified exclusively by other tools are unlikely to act as cancer drivers.

We also explored which mutations would be identified as drivers on a per-patient basis, since ParsSNP and similar tools will frequently be used in this fashion as it becomes common to exome-sequence clinical cases. Table 3.2 shows the top five candidate drivers as identified by ParsSNP, CanDrA and CHASM in patients 313T, 319T and 361T from the Kakiuchi dataset. In patient 313T, ParsSNP correctly identifies RHOA Y42C as a driver, and also suggests PIK3CA H1047L (H1047R is a confirmed driver[111]) and a truncation in the tumor suppressor ARID1A. Of these three plausible drivers, CanDrA and CHASM only identify the PIK3CA mutation. In patient 319T, ParsSNP identifies several truncations in known tumor suppressors SMAD4,

ARID1A and P53, which the other tools miss since they do not apply to truncations. ParsSNP also informatively suggests that these truncations are more likely to be drivers than the R56C mutation in BAP, which is itself a known cancer gene[83]. Finally, ParsSNP identifies missense mutations in the known cancer genes P53, FGFR2 and NOTCH2 in patient 361T, as well as a truncation in the tumor suppressor CDKN2A[83]. Of these, CanDrA and CHASM only identify the P53 and FGFR2 mutations. However, they also identify an implausible candidate driver in dystrophin (DMD R137Q)[45]. Based on these observations, we conclude that ParsSNP identifies candidate drivers that are more biologically plausible than those produced by competing methods, both across whole datasets and within individual patients.

3.3.8 ParsSNP and Novel Driver Identification

Another use of ParsSNP is to identify biological hypotheses. We pooled ParsSNP scores for the hypermutator, training and test sets. To narrow focus, we considered only the 75 unique mutations with ParsSNP scores over 0.5 (Table 3.3). They include recurrent driver mutations in BRAF (V600E), IDH1 (R132C/L) and NRAS (Q61R), but 54/75 mutations are not recurrent, including the top three: CTNNB1 P687L (ParsSNP=0.795), NRAS E153A (0.789), and CTNNB1 F777S (0.787). Most of the mutations are within CGC members, but thirteen are not: two are in TATA-binding protein (TBP A191T and R168Q) and three are in a calcium-dependent potassium channel (KCNN3 R435C, L413Q and S517Y). Moreover, TBP and KCNN3 have generally elevated ParsSNP scores by one-sample Wilcoxon test (Supplementary Figure S3.8, Bonferroni $p<0.05$). Taken together, ParsSNP suggests these genes as putative cancer genes, with TBP A191T and R168Q, and KCNN3 R435C, L413Q and S517Y as the most promising driver mutations.

We also examined the differences in ParsSNP scores between hypermutators and non-hypermutators (training and test). While many genes have elevated scores exclusively in the nonhypermutators, none could be detected in only the hypermutators (Supplementary Figure S3.9A). However, a differential functionality analysis on a per-gene basis (Supplementary Figure S3.9B) highlighted two genes: RNF43 (a ubiquitin ligase[112]) and UPF3A (involved in nonsense mediated decay[113]) have modestly but significantly elevated ParsSNP scores in the hypermutated samples, suggesting that they may play a role in hypermutator biology.

3.3.9 Avenues for Model Improvement

Two possible approaches for improving ParsSNP are to add additional data or focus the model on particular cancer types. Testing ParsSNP on subsets of the training data shows that performance is roughly constant for each classification task until the dataset drops to less than ~250-500 samples (~5-10% of the training data, Supplementary Figure S3.10). Since adding pan-cancer data appears unlikely to improve importance, we next considered how narrowing scope to a single cancer type would affect ParsSNP. Versions of ParsSNP were trained and tested in breast, lung adenocarcinoma, melanoma, colorectal adenocarcinoma or head and neck squamous cell carcinoma (cancers with at least 150 patients and 25,000 mutations in the training set, Table 3.4). The pan-cancer version of ParsSNP generally outperformed these more targeted models. However, predictions made by cancer-specific and full ParsSNP models were not very correlated, and aggregate performance may mask important differences in predicted drivers (Table 3.4). Additional data for these cancer types will clarify if these results are a consequence of noise or true biological differences.

We also explored the use of thresholding to optimize ParsSNP predictions. Since ParsSNP is trained using unlabeled mutations, there is no single objective criterion for setting a ParsSNP

threshold. One option is to set thresholds so as to optimize the percentage of samples assigned a number of drivers meeting the E-step boundaries. This approach suggests a cutoff of 0.07 for nonhypermutators, and 0.12 for hypermutators (Supplementary Figure S3.11A). Alternatively, a threshold could be selected to optimize accuracy in the classification tasks, suggesting a range of 0.08 to 0.16 (Supplementary Figure S3.11B). While a ParsSNP cutoff of 0.1 may be reasonable in many situations, the observed variations suggest that thresholds be set in a context specific manner, taking into account the relative importance of sensitivity and specificity for the task at hand.

3.4 Discussion

Cancer genome sequencing studies identify large numbers of mutations, and it is likely that only a small fraction are drivers[44, 67]. The presence of many passenger mutations can make it difficult to direct experimental and clinical decision-making. FISs are designed to filter out passengers, but shortcomings include limited generalizability due to biases introduced through pre-labeled training data.

ParsSNP avoids the use of pre-labeled training data by using parsimony to generate its own labels. This requires two constraints: 1) Putative drivers should be relatively equitably distributed among samples, which is the basis of the E-step. 2) Putative drivers should be definable in terms of the descriptors, which is enforced by the M-step. Using these constraints, we found a single set of labels in the training pan-cancer set, and trained a model to generate ParsSNP scores.

In the four classification tasks, we found that ParsSNP outperformed existing methods in 19 out of 20 comparisons. Moreover, no single existing method could consistently rival ParsSNP in these tasks, an important consideration since ParsSNP and other tools will likely be applied to

diverse datasets in practice. To illustrate a typical usage case, we applied ParsSNP to an independent set of thirty diffuse-type gastric cancer genomes, and found that it identified known and likely candidate drivers that other methods did not detect. When combined with the aggregate results in the classification tasks, this analysis leads us to conclude that ParsSNP is superior to existing methods for quickly identifying likely cancer drivers in somatic cancer mutation data.

Many avenues can be explored to improve ParsSNP performance and broaden its applications. We showed that simply adding pan-cancer data to the training set is unlikely to accomplish this goal. However, expanding the set of descriptors is one promising possibility: whereas ParsSNP uses 23 descriptors, CHASM had access to 49[63], and CanDrA had 95[62]. The ParsSNP method can also be adapted beyond protein-coding somatic mutations in cancer. For instance, none of the assumptions that underpin ParsSNP are cancer-specific. One can envision a version of ParsSNP that is trained using germline mutations from patients with other polygenic diseases, although the set of descriptors would need modifications.

Methods are also needed for identifying regulatory drivers of cancer, which will become more prevalent as datasets involve greater proportions of whole-genome-sequenced samples[114]. ParsSNP could be well suited to this task, but several challenges will need to be overcome. ParsSNP's current descriptors are largely applicable only to protein-coding mutations. Fortunately, frameworks for defining informative descriptors for regulatory variants already exist[61, 115]. It seems plausible that combining descriptors as defined by these studies with a ParsSNP training approach could produce models that effectively identify regulatory and protein-coding drivers.

The identification of pathogenic mutations can guide experimental and clinical decisions. We believe that ParsSNP can aid in this task by leveraging the configuration of mutations within samples to generate more biologically relevant predictions. We demonstrated the strength and generalizability of ParsSNP when detecting driver mutations in cancer using a variety of datasets; moreover, beyond the direct applications we have demonstrated, ParsSNP represents a novel paradigm for the problem of functional impact prediction.

AUROCs for Each Classification Task					
	Recurrence	**CGC**	**driver-dbSNP**	**P53**	
	Detection of recurrent mutations (present in >1 samples) in the pan-cancer test set. (9,434/173,049)	Detection of mutations in CGC members in the pan-cancer test set (no recurrent mutations). (3,760/169,289)	Detection of experimentally validated drivers against dbSNP common variants. (1,138/49,880)	Detection of disruptive mutants (activity <25% of wild type) in IARC P53 dataset. (475/1,839)	Description (Cases/Controls)
ParsSNP	**0.656**	**0.833**	**0.975**	**0.843**	**NN trained with parsimony and 23 descriptors**
Independent Tools					
CanDrA	0.608*****	0.764*****	0.959*****	0.707*****	Reference 6
CHASM	0.584*****	0.769*****	0.948*****	0.853	Reference 7
FATHMM Cancer	0.578*****	0.751*****	0.971	0.821*****	Reference 19
TransFIC	0.543*****	0.559*****	0.854*****	0.823*****	Reference 20
Condel	0.543*****	0.608*****	0.918*****	0.839	Reference 21

Table 3.1. Performance summary of ParsSNP and independent tools. Area-under-receiver-operator-characteristics (AUROC) are shown for ParsSNP and five independent tools which were not used as descriptors. Starred values (*) are significantly worse than the performance achieved by ParsSNP ($p<0.05$, Delong Test). Unstarred values are not statistically significantly different from ParsSNP performance ($p>0.05$, Delong Test).

Patient	Total Mutations	Rank	Top ParsSNP Drivers		Top CanDrA Drivers		Top CHASM Drivers	
313T	**170**	**1**	RHOA	Y42C	PIK3CA	H1047L	PIK3CA	H1047L
		2	PIK3CA	H1047L	NPFFR2	F150V	PCDH17	T426N
		3	ARID1A	E1860*	MTMR8	S307A	RPAP1	G993R
		4	TMPRSS13	R384W	SLITRK4	K110T	CDC27	N129S
		5	PXDN	R1409W	FBLN2	C1036S	VCAN	G360R
319T	**75**	**1**	SMAD4	L414fs	BAP1	R56C	NRXN3	R103C
		2	ARID1A	R2116*	TTLL5	S1071N	DHX36	A819T
		3	TP53	R64*	KIF4B	K1097N	EPHB1	A669V
		4	BAP1	R56C	KIF4B	D1127H	RELN	R730H
		5	FMN2	R1450*	KIF4B	D1061N	ATP8B4	A882E
361T	**58**	**1**	TP53	G202V	TP53	G202V	FGFR2	D538H
		2	FGFR2	D538H	FGFR2	D538H	DMD	R137Q
		3	CDKN2A	V115fs	DMD	R137Q	MAN1A1	R284H
		4	NCOA3	Q1268insPE	RIMS2	R208H	TP53	G202V
		5	NOTCH2	H1390P	PHACTR4	R367C	IMPG1	Y180C

Legend: Confirmed as driver in original study; Truncation in known TSG; Other mutations in CGC members; Mutations large muscle protein

Table 3.2. Driver mutations suggested by ParsSNP and other tools in specific patients. For patients 313T, 319T and 361T from the Kakiuchi *et al* study, the top five predicted drivers are shown as determined by ParsSNP, CanDrA and CHASM.

Rank	Gene	AA Change	ParsSNP	Recurrence	CGC	Rank	Gene	AA Change	ParsSNP	Recurrence	CGC
1	CTNNB1	P687L	0.795	1	X	**39**	LOR	*313W	0.586	1	
2	NRAS	E153A	0.789	1	X	**40**	CHEK2	Y447D	0.585	1	
3	CTNNB1	F777S	0.787	1	X	**41**	CTNNB1	K170M	0.582	1	X
4	BRAF	V600G	0.786	3	X	**42**	BRAF	L597R	0.582	3	X
5	IDH1	R132L	0.777	10	X	**43**	BRAF	G466V	0.576	6	X
6	SF3B1	I1241T	0.775	1	X	**44**	EEF1B2	W149C	0.571	1	
7	IDH1	R49C	0.761	2	X	**45**	SLC36A2	L345P	0.570	1	
8	IDH1	I102T	0.759	2	X	**46**	PIK3CA	M1004I	0.563	3	X
9	TBP	A191T	0.751	1		**47**	CTNNB1	R661L	0.560	1	X
10	PPP2R1A	I397N	0.706	1	X	**48**	PPP2R1A	R258C	0.560	2	X
11	RPL8	D176G	0.706	1		**49**	U2AF1	Q157P	0.559	2	X
12	BRAF	Q609H	0.706	1	X	**50**	KRAS	C118S	0.559	1	X
13	BRAF	L505H	0.694	1	X	**51**	PPP2R1A	R46S	0.558	1	X
14	BRAF	V600E	0.681	519	X	**52**	RAC1	P106L	0.554	1	X
15	PPP2R1A	C329F	0.680	1	X	**53**	NFE2L2	R486C	0.549	1	X
16	BRAF	D594G	0.677	3	X	**54**	MYD88	R301C	0.549	1	X
17	BRAF	L514P	0.677	1	X	**55**	KRAS	Q61R	0.543	9	X
18	TBP	R168Q	0.674	1		**56**	IDH1	R132C	0.541	56	X
19	NRAS	Q61P	0.663	2	X	**57**	BRAF	S467L	0.541	3	X
20	PIK3CA	M1043V	0.657	11	X	**58**	BRAF	L537S	0.538	1	X
21	PIK3CA	D1045V	0.649	1	X	**59**	PIK3CA	E542G	0.538	1	X
22	SF3B1	R1245T	0.649	1	X	**60**	LATS2	V729D	0.534	1	
23	BRAF	L485S	0.649	1	X	**61**	OPRD1	C273Y	0.532	1	
24	NRAS	T50I	0.645	2	X	**62**	KCNN3	S517Y	0.532	1	
25	BRAF	M53T	0.629	1	X	**63**	EZH2	C590S	0.531	1	X
26	AKT1	L153P	0.625	1	X	**64**	SF3B1	R451L	0.530	1	X
27	SF3B1	I1268M	0.624	1	X	**65**	NRAS	Q61R	0.526	96	X
28	BRAF	K601T	0.605	1	X	**66**	NFE2L2	W8R	0.525	2	X
29	RAC1	P159L	0.605	1	X	**67**	CTNNB1	I303M	0.524	2	X
30	CTNNB1	C429G	0.604	1	X	**68**	KRAS	T158A	0.521	2	X
31	MYD88	W299C	0.603	1	X	**69**	KRAS	R135T	0.520	1	X
32	BRAF	L597Q	0.602	1	X	**70**	GNAS	D839G	0.508	1	X
33	BRAF	H539P	0.601	1	X	**71**	SF3B1	R736C	0.507	1	X
34	SF3B1	M1195V	0.601	1	X	**72**	BCL2	Y108S	0.506	1	
35	KCNN3	R435C	0.598	1		**73**	NRAS	D154G	0.503	1	X
36	KCNN3	L413Q	0.594	1		**74**	IDH1	D375Y	0.502	1	X
37	BRAF	G563C	0.590	1	X	**75**	BRAF	G596D	0.501	1	X
38	U2AF1	R53C	0.588	1	X						

Table 3.3. Exceptional mutations by ParsSNP score. The top 75 distinct mutations (ParsSNP > 0.5) from the full pan-cancer dataset (hypermutators, training and test sets) are listed. The gene, amino acid change, ParsSNP score, recurrence in the full dataset (*i.e.* the number of samples the mutation was observed in), and the CGC status of the gene are indicated.

	Data Volume				AUROCs in Classification Tasks								EM Correlation	
	Training		Test		Recurrence~^		CGC^		Drive-dbSNP*		P53*			
Cancer Type	Samples	Mutations	Samples	Mutations	Cancer targeted	ParsSNP	Cancer targeted	ParsSNP	Cancer targeted	ParsSNP	Cancer targeted	ParsSNP	Pairwise	with ParsSNP
SKCM	250	48374	127	26486	0.658	0.680	0.726	0.820	0.932	0.975	0.820	0.843	0.997	0.867
LUAD	720	131660	322	58542	0.722	0.732	0.770	0.830	0.935	0.975	0.838	0.843	0.995	0.797
COAD	288	34062	150	18199	0.650	0.660	0.813	0.875	0.894	0.975	0.669	0.843	0.922	0.342
HNSC	300	43031	149	21929	0.737	0.676	0.736	0.849	0.946	0.975	0.694	0.843	0.982	0.446
BRCA	716	38776	379	22562	0.791	0.830	0.814	0.839	0.972	0.975	0.814	0.843	0.875	0.543

Table 3.4. Cancer-specific training and testing of ParsSNP models. Models for each of the indicated cancer types were developed using the same methods as were used to develop the pan-cancer ParsSNP model, using the indicated training data (drawn from the pan-cancer training dataset). These models were then assessed by their ability to detect recurrent events and mutations in the CGC using the indicated test data (drawn from the pan-cancer test dataset), and by their performance in the driver-dbSNP and P53 assessments (using the full assessment datasets). The pan-cancer ParsSNP model was also run in these test sets for comparison. The reproducibility of cancer-specific models is indicated using the mean Pearson pairwise correlation of five training runs for each. The correlation with ParsSNP labels for the same cancer type is also indicated. Abbreviations: SKCM=melanoma, LUAD=lung adenocarcinoma, COAD=colorectal adenocarcinoma, HNSC=head and neck squamous cell carcinoma, BRCA=breast adenocarcinoma. ˜Recurrence is re-defined in each cancer-specific subset of mutations. ^Recurrence and CGC are only assessed in missense mutations of the test set belonging to the corresponding cancer type. *The driver-dbSNP and P53 datasets have no cancer types, so models trained in specific cancer types are applied to the full dataset. Cyan - the cancer-specific model outperformed the pan-cancer ParsSNP model (Delong test $p<0.05$); pink – the pan-cancer ParsSNP model significantly outperformed the cancer specific model (Delong test $p<0.05$); white – the pan-cancer and cancer-specific models were not significantly different (Delong test $p>0.05$).

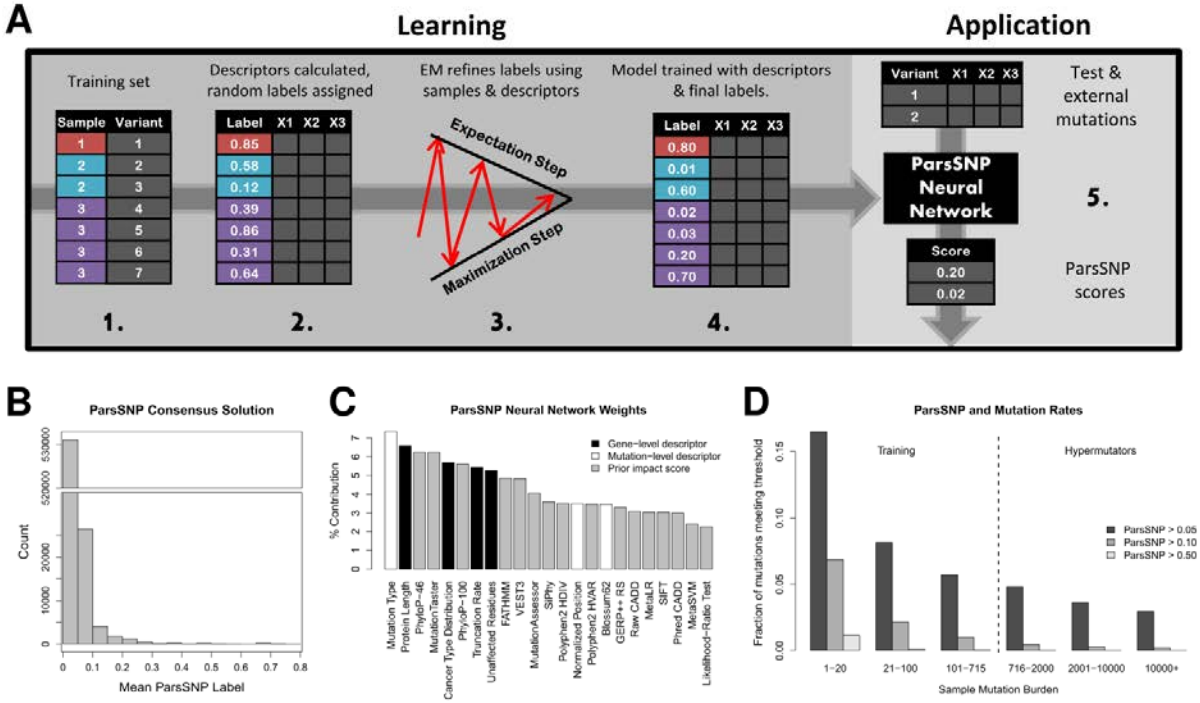


Figure 3.1. Overview of ParsSNP and label learning. A) 1. Label learning begins with a training set of mutations, each belonging to a sample. 2. Descriptors are assigned, and random labels generated (portrayed numbers are illustrative). 3. EM updates labels iteratively such that putative drivers are distributed among samples (E-step) and defined in terms of descriptors (M-step). 4. The final labels and descriptors are used to train a neural network model. 5. The ParsSNP model produces ParsSNP scores when applied to new mutations. B) Distribution of ParsSNP labels after averaging 50 runs (N=566,223). C) Percent contribution of descriptors to ParsSNP scores, using Garson's algorithm for neural network weights (see section 3.3.3). D) The ParsSNP model was applied to the training and hypermutator pan-cancer sets to produce ParsSNP scores. The fraction of mutation identified as drivers is displayed at various sample mutation burdens and ParsSNP thresholds.

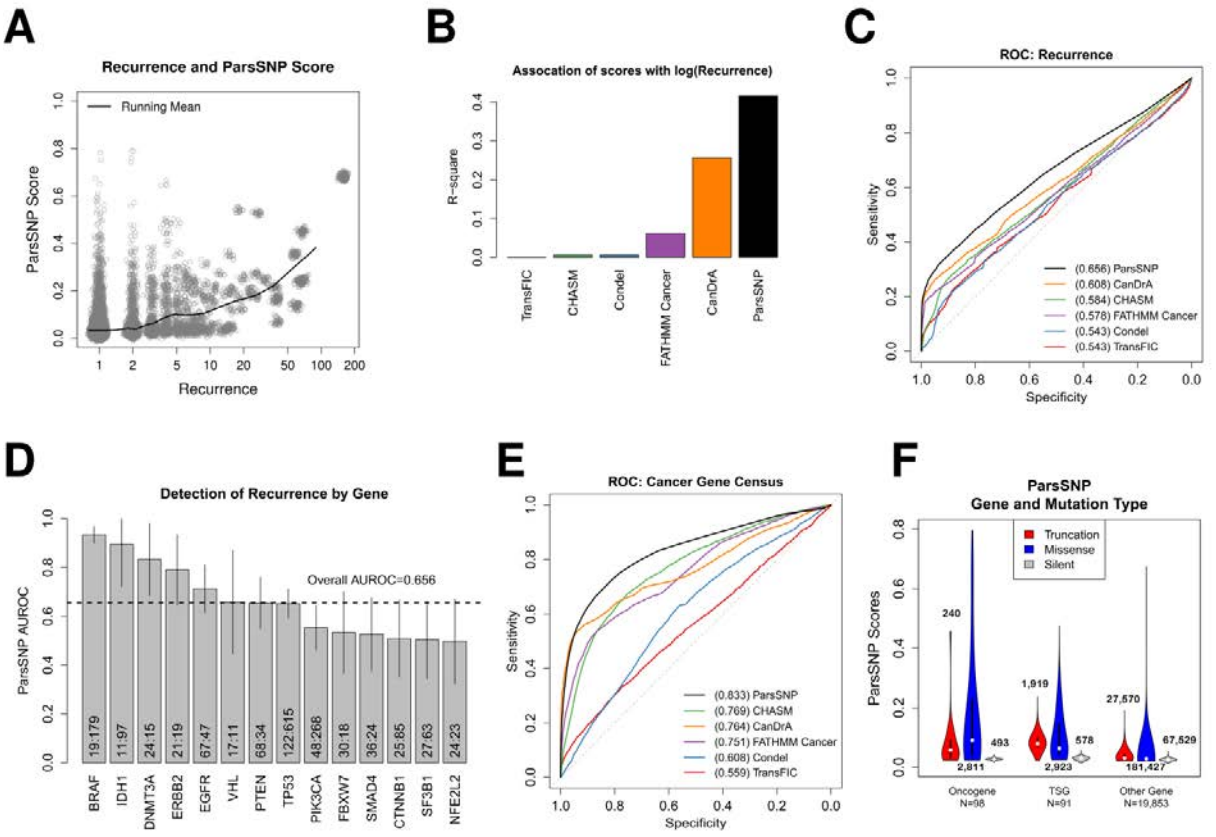


Figure 3.2. Detection of recurrent mutations and mutations in known cancer genes. A) ParsSNP scores plotted against mutation recurrence for missense mutations (N=182,483). Points are jittered to aid visualization. B) The association of ParsSNP and the independent tools to log(mutation recurrence) for missense mutations, measured by R-square. C) ParsSNP identifies 9,434 recurrent missense mutations better than the independent tools (all Delong tests $p<2.2e-16$, AUROCs are depicted). D) The ability of ParsSNP to detect recurrent missense mutations in the test set is assessed on a gene-by-gene basis. Portrayed genes must be members of the CGC; have at least 25 missense mutations; and must have at least 10 mutations in each class. Mutation counts (non-recurrent:recurrent) and 95% confidence intervals are included for each gene. E) Out of 173,049 non-recurrent missense mutations, ParsSNP identifies the 3,760 which occur in the CGC significantly better than the independent tools (all Delong tests $p<2.2e-16$, AUROCs are depicted). F) CGC genes were divided into putative oncogenes and putative tumor suppressor genes (TSG) based on the molecular genetic annotation from the CGC dataset (dominant or recessive, respectively). The distribution of ParsSNP scores in the test set is displayed by mutation and gene type, with the number of genes and mutations in each category displayed. 'Truncation' events include frameshift, premature stop, nonstop and splice-site changes. 'Missense' mutations include missense substitutions as well as inframe insertions/deletions. 'Silent' changes include synonymous nucleotide substitutions as well as non-coding variants.

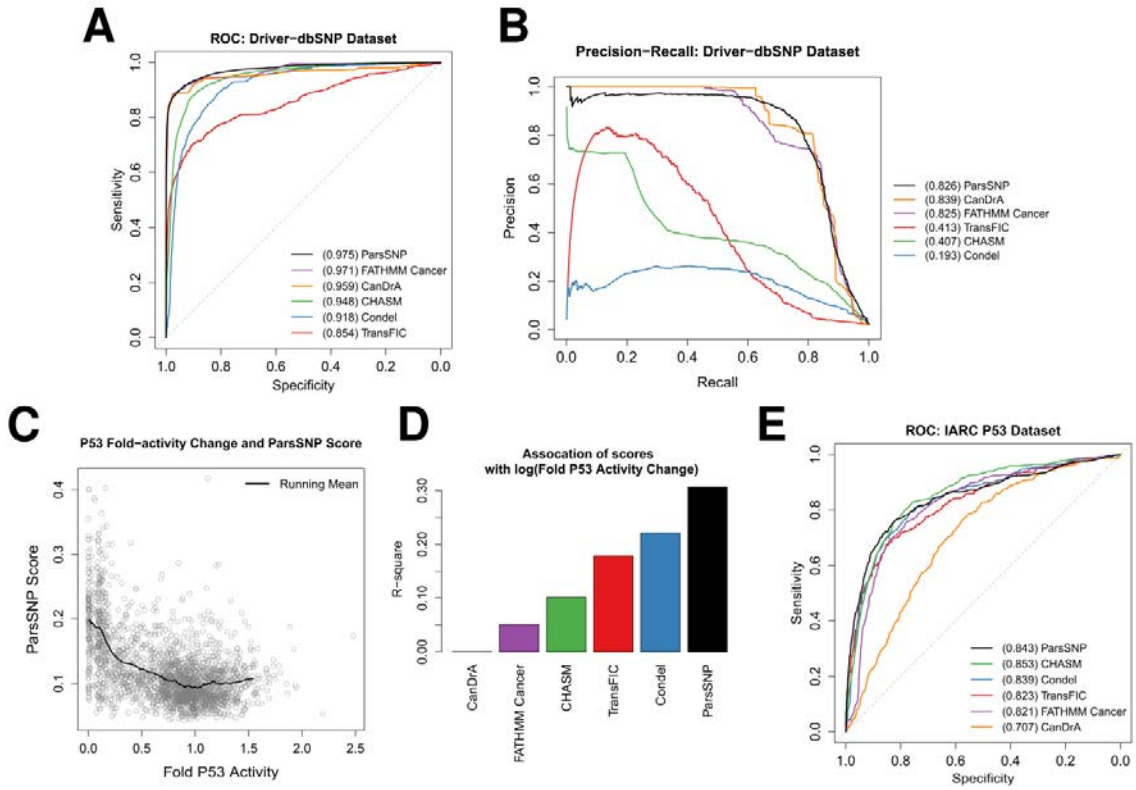
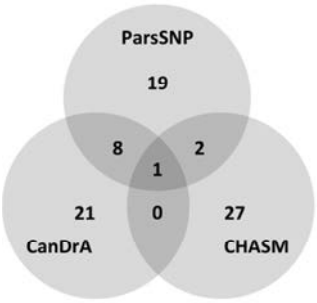


Figure 3.3. Detection of experimentally characterized mutations. A) ParsSNP separates 1,138 driver mutations from 49,880 common SNPs in the driver-dbSNP dataset slightly better than FATHMM Cancer (Delong test p=0.205) and significantly better than the other independent tools (all Delong tests p<1e-4, AUROCs are depicted). B) ParsSNP, CanDrA and FATHMM Cancer have similar performances under a precision-recall analysis of the driver-dbSNP dataset (AUPRs are depicted). C) Plot of ParsSNP scores and P53 transactivation activity change for 2,314 mutations in the IARC dataset. D) The association of ParsSNP and the independent tools with log(P53 activity change) is displayed as measured by R-square values. E) ParsSNP identifies 475 disruptive P53 mutations (mutation P53 activity < 25% of wild type) among 2,314 mutations with similar performance to CHASM (Delong test p=0.39) and Condel (Delong test p=0.59), while CanDrA, FATHMM Cancer and TransFIC perform worse (all Delong tests p<0.05, AUROCs are depicted).

Candidate Drivers in diffuse-type gastric carcinoma



ParsSNP + CanDrA + CHASM

Gene	Mutation	Patient
PIK3CA	H1047L	313T

ParsSNP + CanDrA

Gene	Mutation	Patient
FGFR2	D538H	361T
TP53	V41M	343T
TP53	R43G	337T
TP53	R43H	353T
TP53	V65L	350T
TP53	S88F	315T
TP53	R141H	299T1
TP53	G202V	361T

ParsSNP + CHASM

Gene	Mutation	Patient
CDC27	I174T	315T
SMARCA4	A1186V	325T

Unique to ParsSNP

Gene	Mutation	Patient
ARID1A	E1860*	313T
CDC27	Y173S	315T
CDC27	Y173S	342T
CDKN2A	V115fs	361T
CSMD3	C1597F	312T
CUZD1	L439R	353T
FMN2	Q1400*	299T1
NCOA3	Q1268insPE	361T
RHOA	L22R	302T
RHOA	Y42C	299T1
RHOA	Y42C	313T
RHOA	Y42C	315T
RHOA	Y42C	332T
RHOA	Y74D	356T
SMAD4	L414fs	319T
SPTA1	Y1821C	353T
TMPRSS13	R384W	313T
TP53	E66*	325T
TP53	Splice	351T

Unique to CanDrA

Gene	Mutation	Patient
ATRX	H1978R	343T
BAP1	R56C	319T
CDH1	D254N	307T
CDH1	D288V	314T
CDH1	P593S	337T
CIC	V230I	357T
DMD	L88R	350T
DMD	R137Q	361T
DMXL1	S1402F	357T
ERBB2	R678Q	307T
KDM5B	N460D	357T
KEAP1	D587G	353T
KMT2C	S321N	312T
KMT2C	E765G	383T
NF1	D1091G	359T
NPFFR2	F150V	313T
ROS1	F2046L	315T
TTN	R8231C	299T1
TTN	F19585L	307T
VPS13C	R2439H	307T
XPC	D729V	342T

Unique to CHASM

Gene	Mutation	Patient
AKAP3	R46H	356T
ANAPC5	I308V	359T
AP1G1	R363L	370T
ARFGEF1	R799L	343T
ARHGAP28	L277R	314T
AXIN2	Y549C	312T
DAAM2	R66Q	307T
DDR2	S667P	343T
DUSP5	S301L	297T
KANSL3	H39Y	325T
MAPK10	L248R	350T
MICAL2	R446L	370T
NAT10	R702Q	359T
NRXN3	R103C	319T
P2RY12	R122L	343T
PPP1R12A	N453Y	356T
PTPRD	R716H	359T
RPAP1	G993R	313T
SEMA6D	R517H	383T
SLC6A7	T276I	307T
SMEK1	Q66H	370T
TBC1D17	P393L	370T
TENM2	R33C	337T
TGFBR1	E168K	350T
TRPC6	R464T	357T
TSC2	S1163G	302T
TYRO3	D706N	296T1

Legend

Truncations in known TSGs	Confirmed as drivers in original study
Other mutations in CGC members	Mutations in large muscle proteins

Figure 3.4. Comparing tool predictions in an independent dataset. ParsSNP, CanDrA and CHASM were applied to the Kakiuchi *et al* dataset, which consists of 2,988 protein-coding somatic mutations from 30 diffuse-type gastric carcinoma patients. For each tool, the top 30 predicted drivers (equivalent to 1% of the dataset) were extracted. The overlap between the candidate driver lists from each tool is diagramed (top left), and the candidate drivers themselves are listed according to the tools they were identified by.

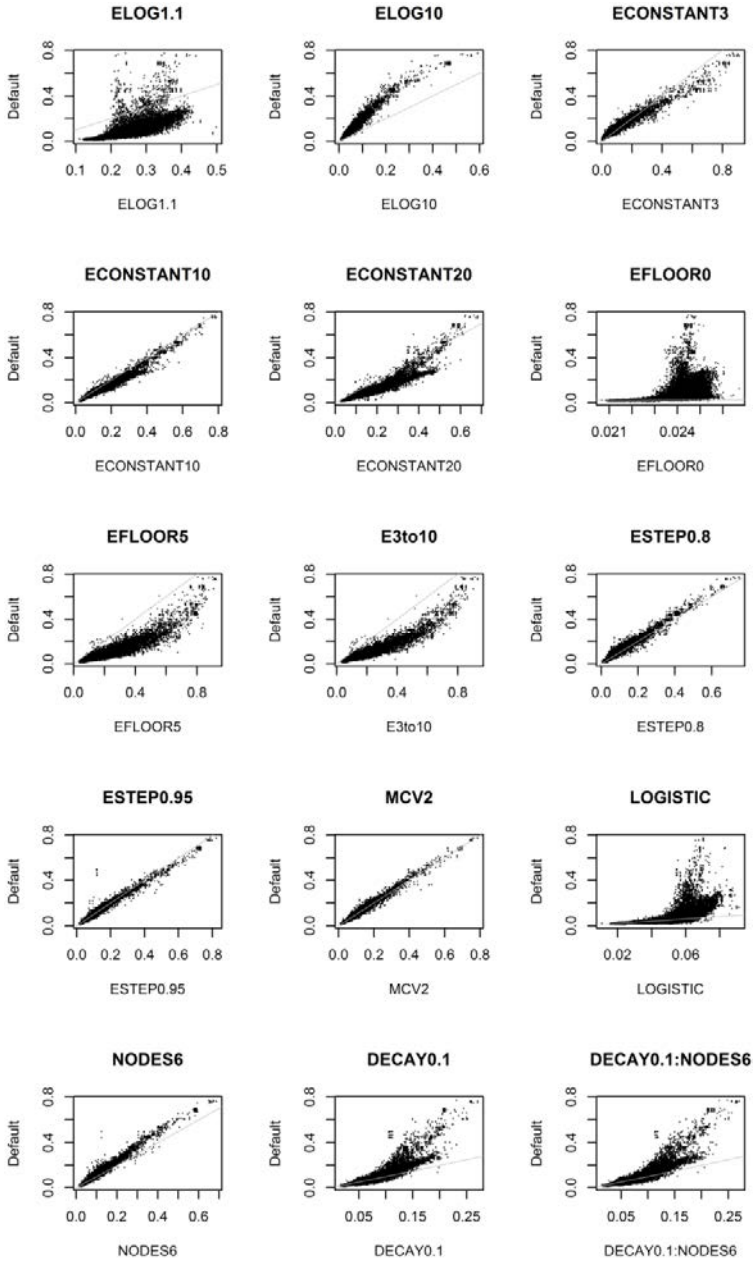


Figure S3.1. Comparison of reference and parameter variations during learning. The results of various alternative parameter settings are plotted against the reference labels in the training dataset (N=566,223). Most alternative settings produce predictions that are highly correlated with the default settings. Key: ELOG1.1, E-step uses a logarithmic upper-bound with base of 1.1 (default=2); ELOG10, logarithm base is 10; ECONSTANT3, E-step uses a constant upper-bound set to 3 (default upper-bound scales logarithmically in base 2); ECONSTANT10, constant upper bound of 10; ECONSTANT20, constant upper-bound of 20; EFLOOR0, E-step lower-bound set to 0 (default=1); EFLOOR5, lower-bound set to 5; E3to10, E-step uses lower and upper bounds of 3 and 10 for all samples; ESTEP0.8, E-step sliding bound calculated as 80% of current belief (default=90%); ESTEP0.95, sliding bound calculated as 95% of current belief; MCV2, M-step uses 2-fold cross validation (default=5); LOGISTIC, M-step uses logistic regression (default is a tuned neural network); NODES6, M-step uses neural network with only 6 hidden nodes (default is tuned, can use more than 6 nodes); DECAY0.1, M-step uses neural network with weight decay of 0.1 (default is tuned, can use less stringent decay); DECAY0.1; NODES6, M-step enforces use of a simpler neural network than default settings require.

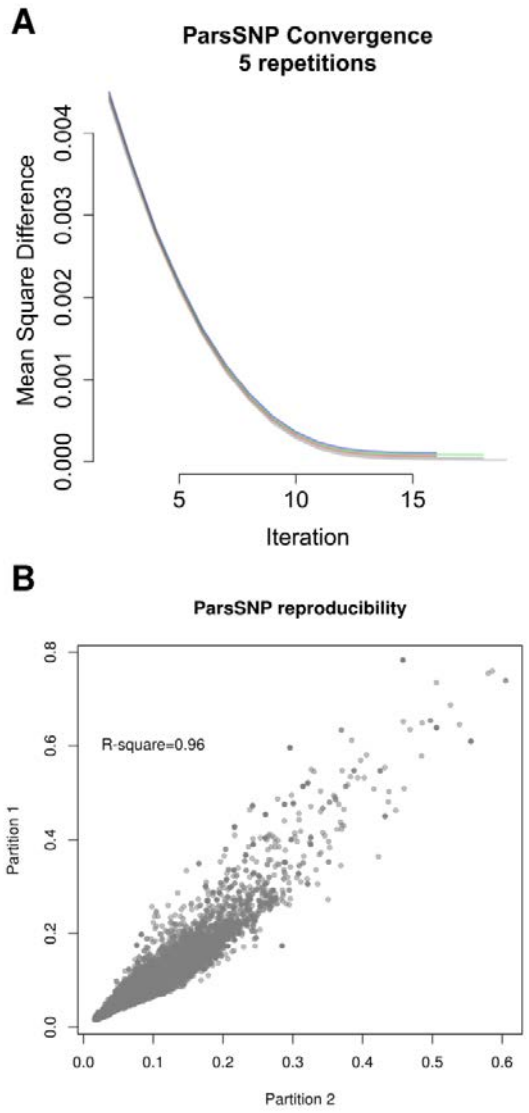


Figure S3.2. ParsSNP convergence and reproducibility. A) The EM portion of ParsSNP consistently converges in 15-20 iterations. Lines are offset slightly to aid visualization. B) The pan-cancer training set was partitioned randomly into two equally sized, independent halves. ParsSNP produces highly correlated scores when trained on independent but comparable datasets (N=566,223).

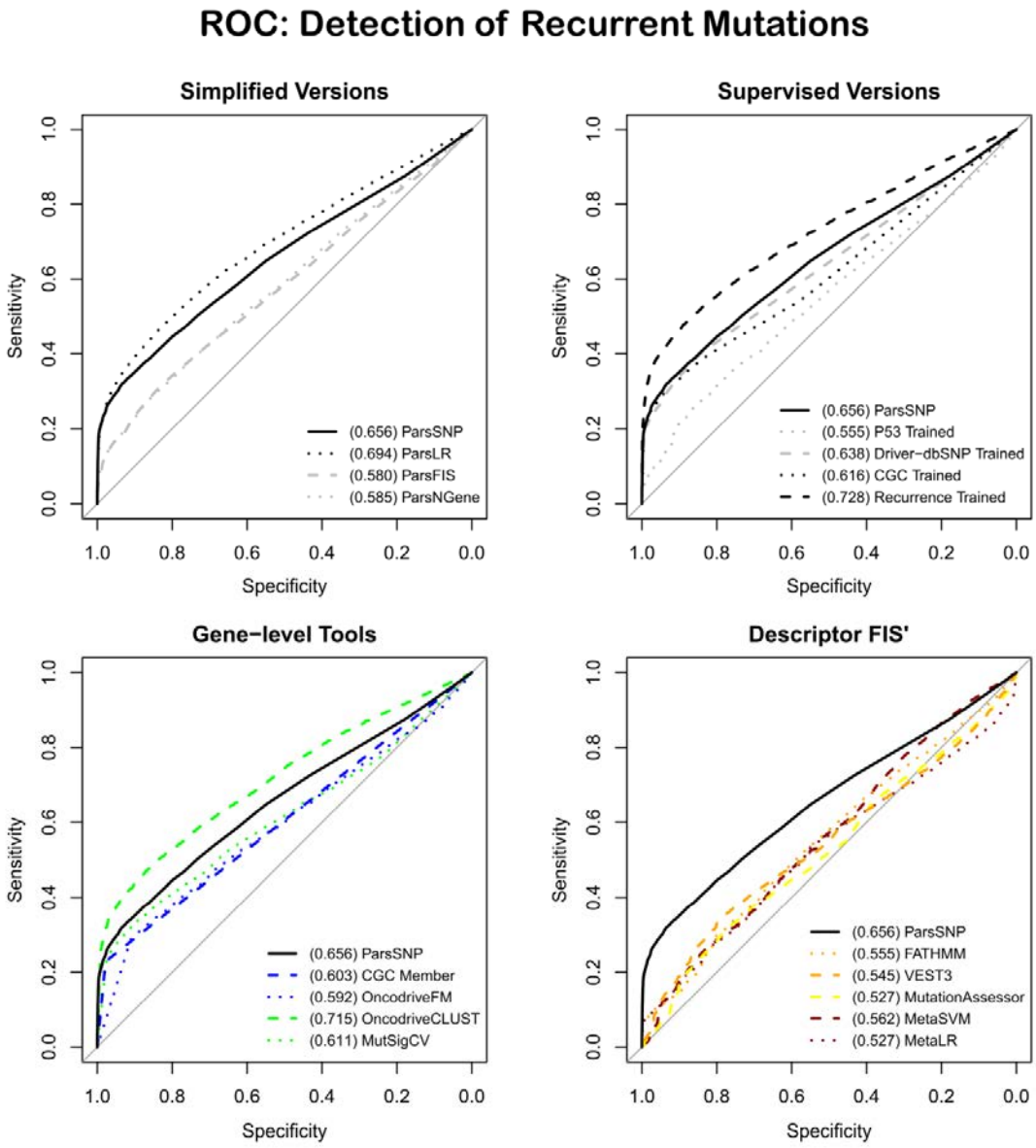


Figure S3.3. Methodological controls and recurrent missense mutations. ROC curves for methodologic controls. AUROCs are depicted.

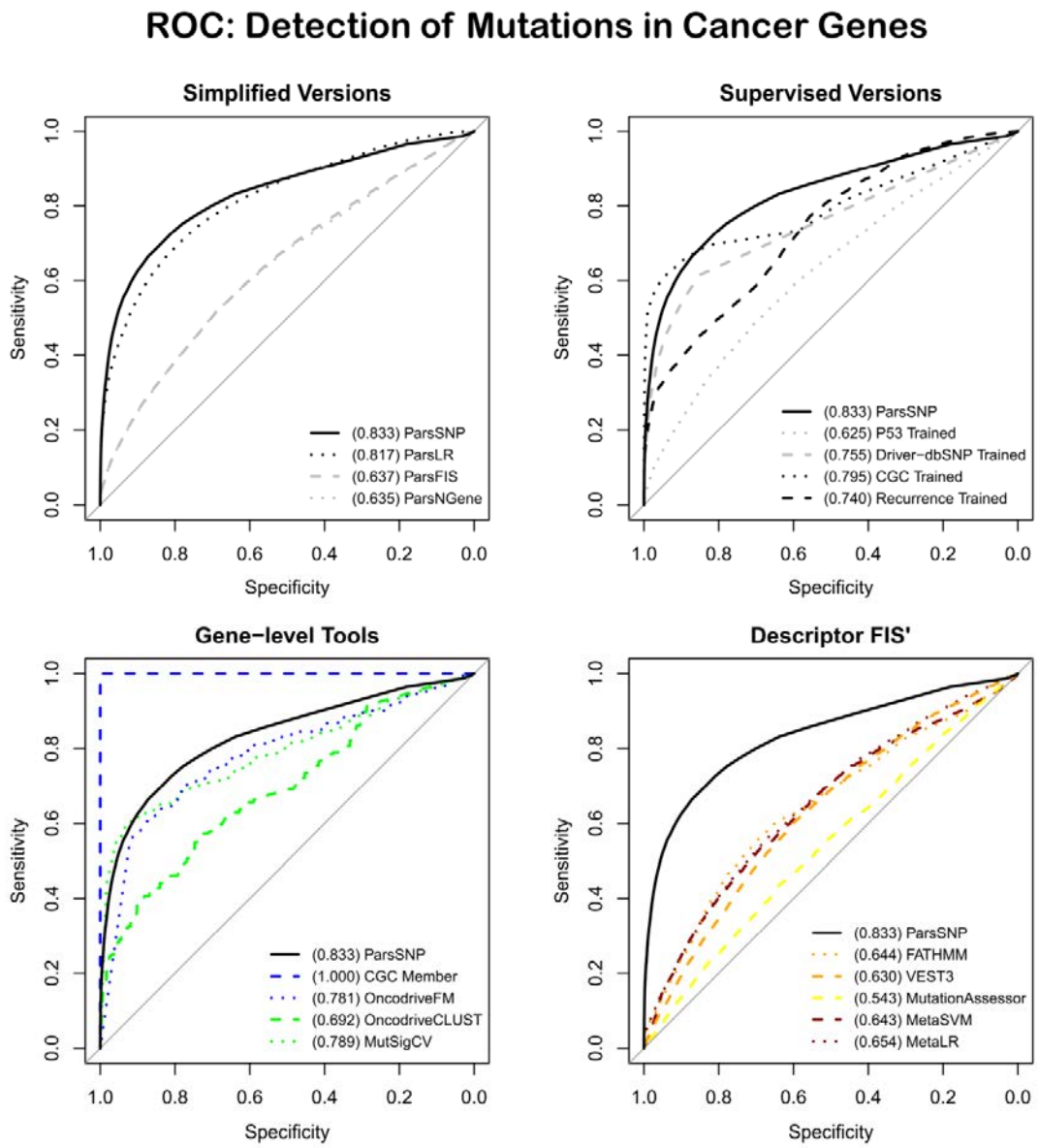


Figure S3.4. Methodological controls and non-recurrent CGC mutations. ROC curves for methodologic controls. AUROCs are depicted.

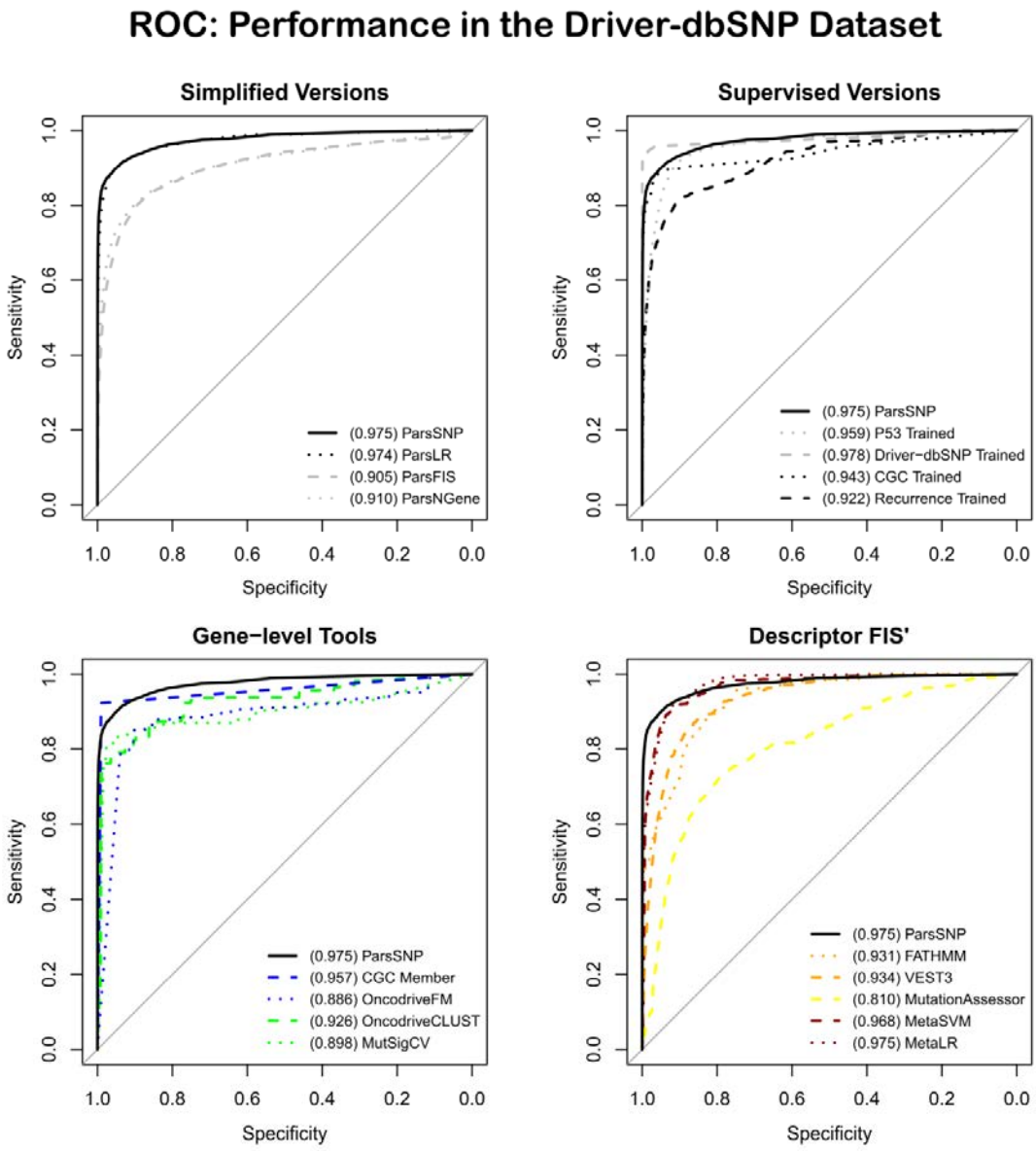


Figure S3.5. Methodological controls in the driver-dbSNP dataset. ROC curves for methodologic controls. AUROCs are depicted.

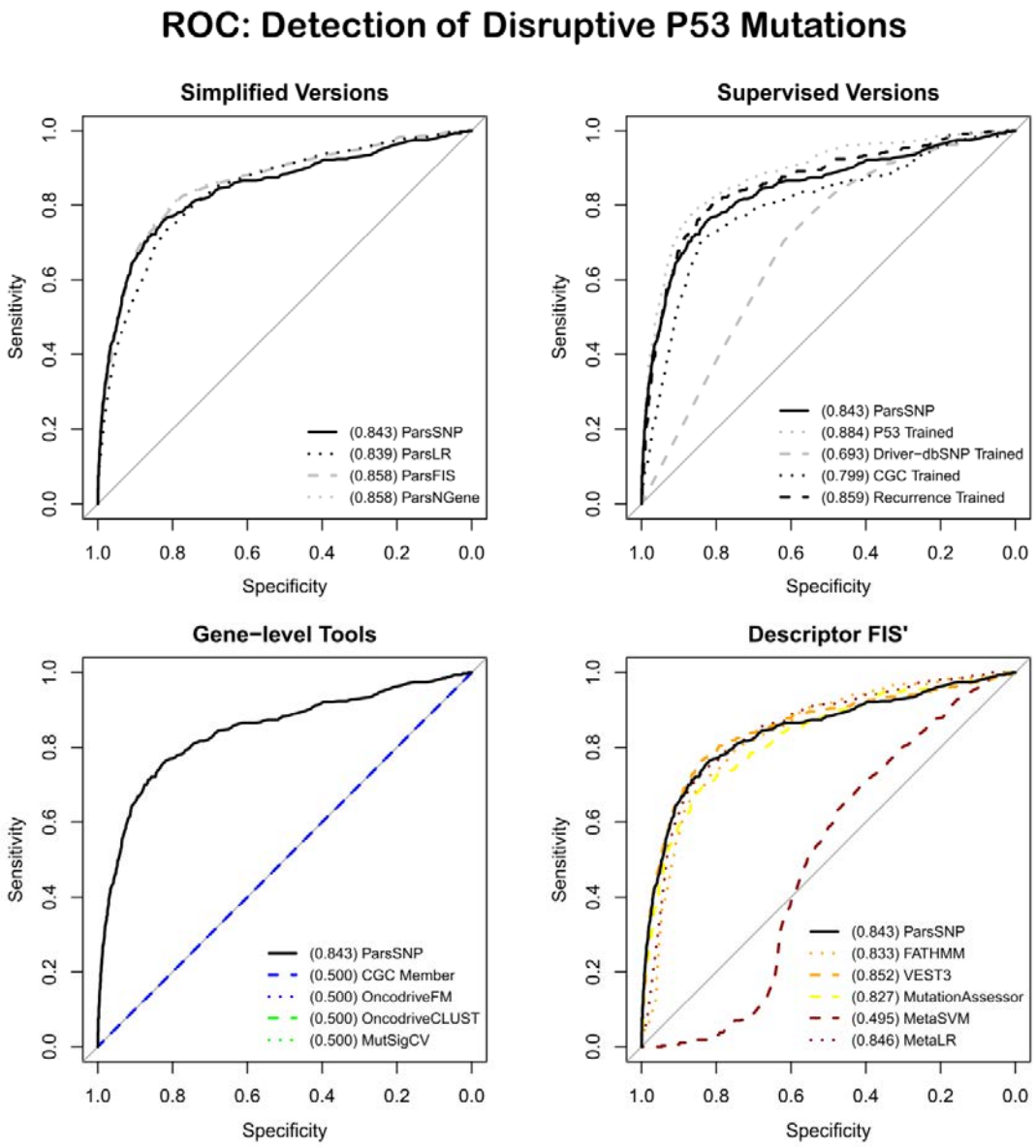


Figure S3.6. Methodological controls in IARC P53 dataset. ROC curves for methodologic controls. AUROCs are depicted.

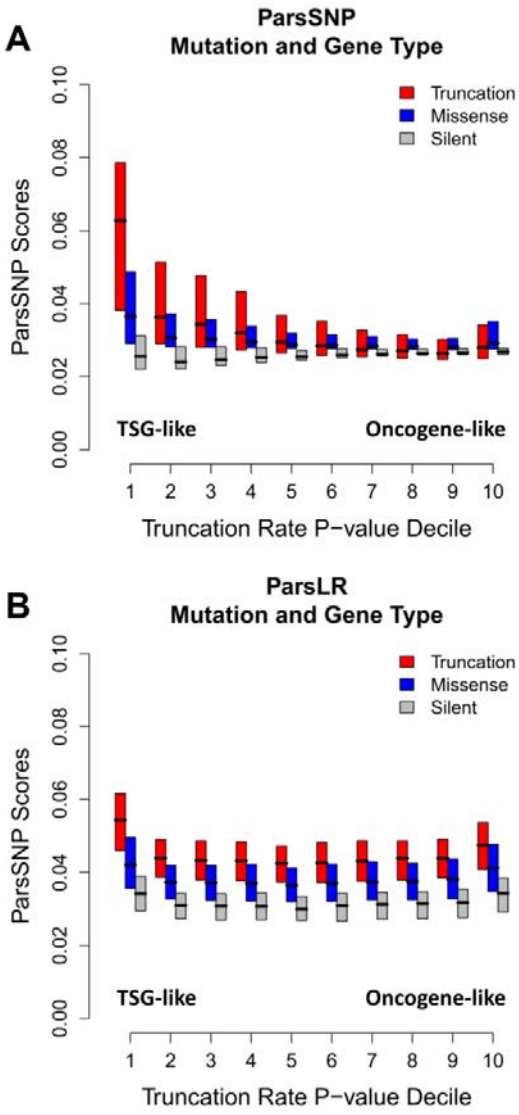


Figure S3.7. Distribution of ParsSNP scores by mutation and gene type. A) Truncation rate is a gene- level descriptor that assigns low p-values to genes enriched in truncations (TSG-like) and assigns high p-values to genes that are depleted in truncations (ONC-like). 'Truncation' events include frameshift, pre mature stop and nonstop changes. 'Missense' mutations include missense substitutions as well as inframe insertions/deletions. 'Silent' changes include synonymous nucleotide substitutions as well as non-coding variants. Truncations receive higher median scores in TSG-like genes, while missense mutations receive higher scores in both TSG-like and ONC-like genes. This represents a potential non-linear two-way interaction between ParsSNP descriptors (*Truncation Rate* and mutation type). Boxes enclose the inter-quartile range. B) ParsLR uses Logistic Regression rather than a neural network model, and does not exhibit the same properties as the full ParsSNP model.

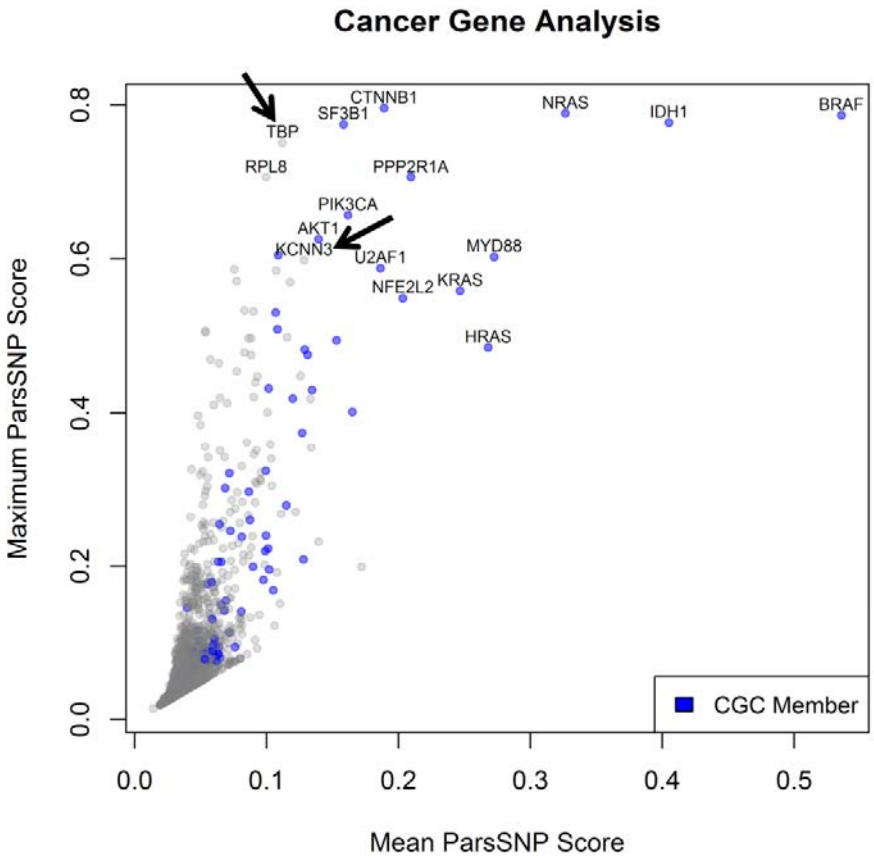


Figure S3.8. Identification of putative driver genes and mutations with ParsSNP. Genes are plotted by the average ParsSNP score of their mutations and their single highest score in the entire pan-cancer dataset (training+test+hypermutator). The top ParsSNP scoring mutations are generally found in members of the CGC. Two genes not belonging to the CGC have multiple exceptional mutations (arrows): TATA Box Binding Protein (TBP), and the calcium-activated potassium channel, KCNN3. Both have significantly higher median ParsSNP scores than expected by chance (Bonferroni corrected one-sample Wilcoxon $p<0.05$) and multiple mutations with exceptionally high ParsSNP scores, including: TBP A191T (ParsSNP=0.75) and R168Q (0.67), as well as KCNN3 R435C (0.60), L413Q (0.59), S517Y (0.53).

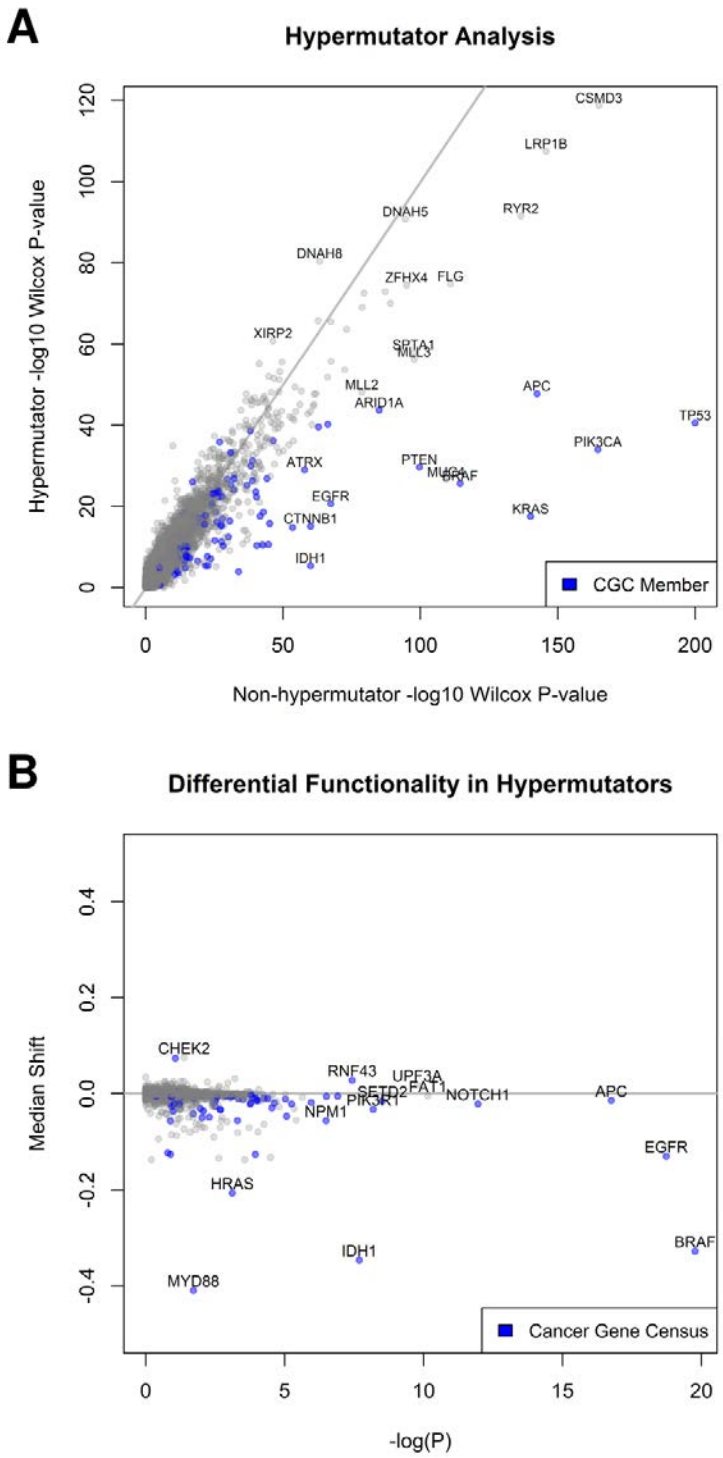


Figure S3.9. Differential functionality between hypermutators and non-hypermutators. A) A one-sample Wilcoxon test was performed on each gene in both the hypermutated and non-hypermutated (training + test) portions of the dataset using internal null distributions. The minus-log10 p-values of these tests are shown. As expected, many well-known cancer genes were more easily detected in the non-hypermutators. No genes were observed with elevated ParsSNP scores exclusively in the hypermutators. B) A two-sample Wilcoxon test was performed for each gene, comparing the ParsSNP scores assigned to it in the hypermutated and non-hypermutated segments. Genes are plotted by the magnitude of median shift (negative values indicate lower scores in the hypermutated samples) and the –log10 p-value. This analysis indicates that mutations in RNF43 and UPF3A have modestly but significantly elevated scores when observed in hypermutators. This suggests that these genes may be involved in the unique biology of these tumors.

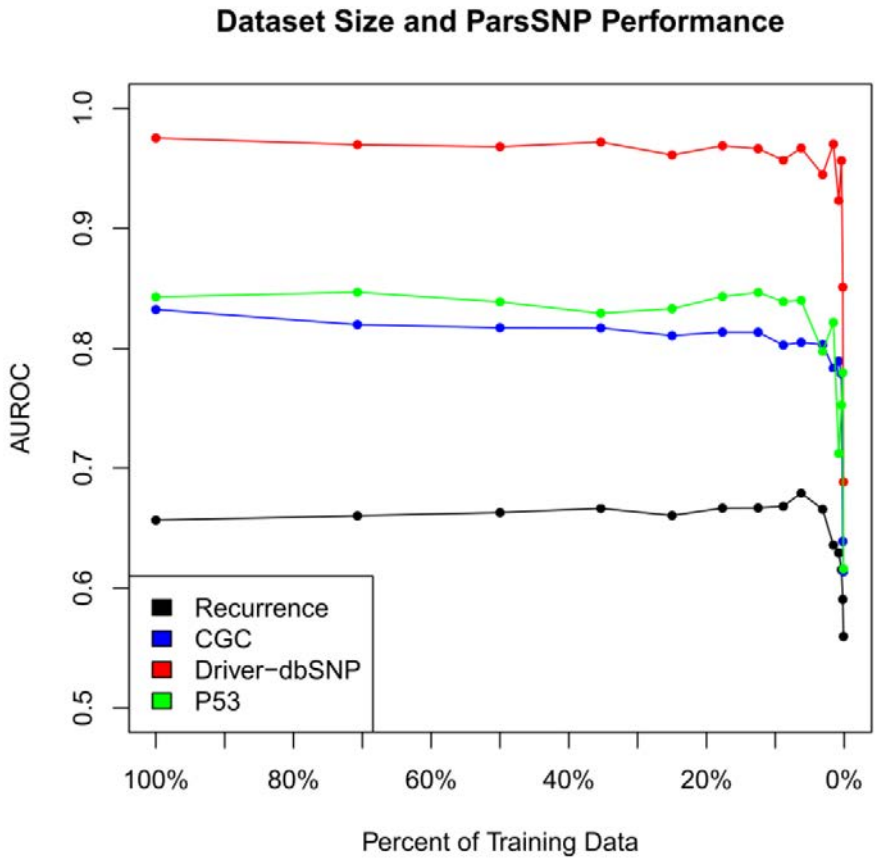


Figure S3.10. ParsSNP performance and dataset size. ParsSNP models were trained on progressively smaller subsets of the pan-cancer training data (N=566,223), and performance (AUROC) assessed for each classification task. Points represent average performance of 5 replicates.

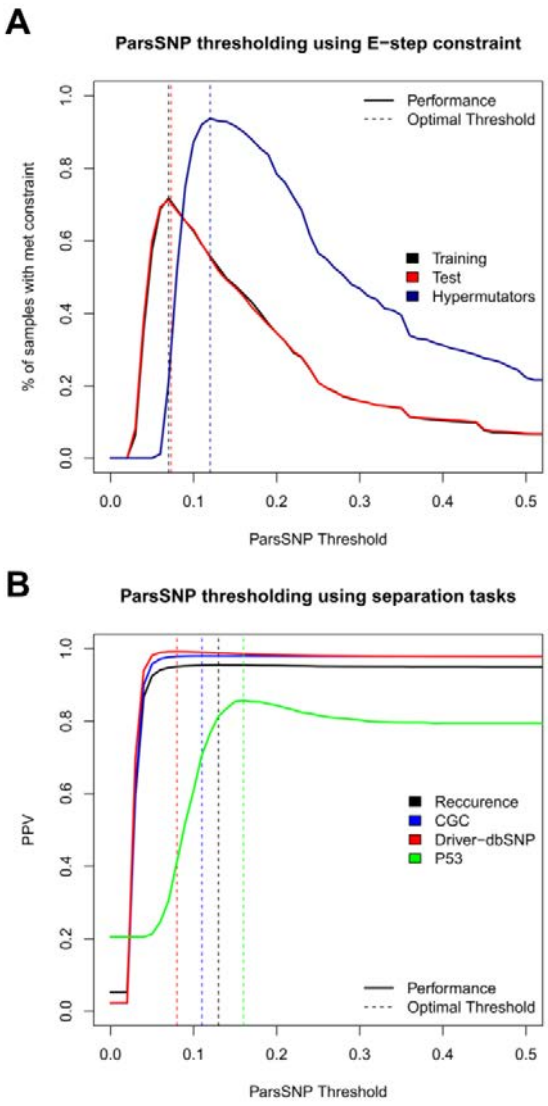


Figure S3.11. Criteria for thresholding ParsSNP scores. A) The E-step constraints are one possible objective criterion for thresholding ParsSNP scores. The value to be optimized is the percentage of samples receiving a number of driver mutations that is compatible with the E-step upper and lower bounds under the proposed threshold. B) Another approach is to select a threshold that optimizes accuracy in the classification tasks.

4. Identifying Drivers in Gene Families

This chapter is adapted from:

Kumar RD, Bose R. (2016). Analysis of somatic mutations across the kinome reveals loss-of-function mutations in multiple cancer types. In preparation.

4.1 Introduction

In this chapter, we develop new methods for analyzing somatic mutations by aggregating them across gene families. In previous chapters, we focused on methods that applied genome wide, either at the level of genes or individual mutations. However, these approaches do not account for knowledge of specific classes of genes and proteins. One gene family which is well studied and plays a large role in cancer development is the kinases. Previous studies make clear that driver events in these enzymes have unique qualities, and that kinases share some common mechanisms of activation and inactivation[72-80]. However, previous kinase-specific methods rely on annotations that are unique to kinases, and may have no analogues in other gene types . We identify and validate functional kinase mutations using a different approach. Rather than use kinase structural knowledge to find functional mutations, we first pursue the reverse task: using observed mutations and a kinase alignment to develop a functionality map of the human kinome. To develop the functionality map, we design a series of statistical tests to identify aligned positions experiencing non-random mutations. These tests are similar to those we developed in chapter 2. This functionality map of the human kinome points us towards relatively few homologous positions with non-random mutations, many of which likely have a biological effect. Using mamalian cell-culture based techniques, we test eleven mutations across four genes at these positions for functionality, and find that all cause some reduction-of-function (ROF). We conclude that ROF point mutations are relatively common in the kinome in several cancer types.

4.2 Materials and Methods

4.2.1 Development of Statistical Tests

We developed a panel of statistical tests which can be used to identify non-random sets of mutations that occur at homologous positions in human kinases. Several of these tests are adapted from the work presented in chapter 2 [96]. In many cases, null distributions are defined empirically (via permutation). Where needed, amino-acid substitution frequencies are defined by mutations that are outside kinase domains (but within genes bearing kinase domains). These mutations are generated by the same mutational processes that produced mutations within the kinase domains, but should not be systematically enriched for biologically non-neutral mutations, which we assume is the case for mutations at some of the aligned kinase-domain positions. In some cases, the null hypothesis is also conditioned on the alignment and aspects of the observed mutations (for instance, most tests assume a fixed number of mutations).

Careful consideration was given to recurrent mutations which occur in more than one patient. These mutations are often presumed to have a functional effect[76], but they may also be idiosyncratic to particular genes. Completely excluding recurrent mutations will likely remove many biologically important mutations from the dataset; but completely including them will likely make the analysis sensitive to positions with even a few recurrent mutations. Therefore, our panel includes tests that operate at three levels which reflect different ways of handling recurrent events. Mutation-level tests (*Mutation Number, Patients, Cancer Types*) include all mutations in the dataset, and consider recurrent events as non-redundant. Residue-level tests (*Reference Residues, Variant Residues*) treat identical amino-acid substitutions as redundant (e.g. CHEK2 K373E, which occurs 48 times in the dataset, is counted as a single event). Finally, gene-level tests (*Cancer Genes, Gene Relatedness*) treat mutations that occur at a single position

in a gene as redundant (e.g. CHEK2 S372F and CHEK2 S372Y are treated as a single event). This approach should balance the value of recurrent mutations in identifying important positions against the risk of finding positions that are not broadly important to kinase function.

Mutation Number: In this simple test, we identify aligned positions with a higher-than-expected number of total mutations. All mutations are used, and the null is set using only non-kinase-domain mutations. We begin by defining the expected number of mutations per residue type (*r*). For each, we calculated the expected number of mutations using the mutations and sequences that are outside of kinase domains:

$$E_r = \frac{O_r}{N_r}$$

where E_r is the expected number of mutations per residue of type *r*, O_r is the observed number of mutations affecting residues of type *r* outside of the kinase domains, and N_r is the total number of residues of type *r* present in gene sequences, but outside of their respective kinase domains. Once the expectations per residue type are set, we calculate the expected number of mutations at each aligned position (*a*):

$$E_a = \sum_r E_r R_{a,r}$$

Where E_a is the expected number of mutations at an aligned position *a*; E_r is the expected number of mutations per residue type *r*, and $R_{a,r}$ is the number of residues aligned at *a* of type *r*. We assume that the presence of mutations at each gene and aligned position can be modeled with a poisson distribution, parameterized by E_r for the appropriate residue type. It follows that the number of mutations for an entire aligned position is therefore also poisson distributed (since it is

a sum of poisson variables), and parameterized by E_a. By comparing the observed number of mutations at the position with the null distribution, we generate an upper tail p-value for the test.

Patients and Cancer Types: In these tests, we identify positions with mutations that are not randomly distributed among patients and cancer types, given the number of mutations observed at the position. They are calculated very similarly to one another, and are described in chapter 2[96]. Both are calculated as chi-square goodness-of-fit tests, although both use empirical rather than theoretical distributions. Both tests use all mutations at the aligned positions. Unlike the other tests, the null distribution *includes mutations in kinase domains, as well as mutations outside kinase domains.*

Each mutation can be assigned to a patient (and cancer type), each of which has a certain mutation count associated with it (c). The mutation count is simply the number of times the patient (or cancer type) occurs in the dataset. Once each mutation has been associated with a value of c, we calculate the test statistic for each aligned position (a):

$$X_a^2 = \sum_c \frac{(O_{a,c} - E_{a,c})^2}{E_{a,c}}$$

$$E_{a,c} = \frac{N_a N_c}{N}$$

Where $O_{a,c}$ is the observed number of mutations at the aligned position from patients (cancer types) with mutation count c, $E_{a,c}$ is the expected number of mutations at the aligned position from patients (cancer types) with mutation count c, N_a is the number of mutations at the position, N_c is the total number of mutations in the dataset from patients (cancer types) with mutation count c, and N is the total number of mutations in the dataset.

This statistic is compared to a null distribution, which is generated by calculating the statistic for random draws with replacement from the set of patient (cancer type) labels, holding the number of mutations fixed. The final output is an upper-tail p-value.

Reference Residues: This test identifies positions where mutated residues appear non-random. It is calculated as a chi-square goodness-of-fit test, but uses an empirical null distribution instead of a theoretical one. It is a residue-level test, and recurrent mutations with identical residue changes are removed. The null distribution is set with mutations from outside of kinase domains. We use the expected number of mutations per residue of each type (E_r) that was used in *Number of Mutations.* We then calculate the test statistic for each aligned position (a):

$$X_a^2 = \sum_r \frac{(O_{a,r} - E_{a,r})^2}{E_{a,r}}$$

$$E_{a,r} = R_{a,r} E_r$$

Where $O_{a,r}$ is the observed number of mutations at the aligned position from residues of type r, $E_{a,r}$ is the expected number of mutations at the aligned position at residues of type r, and $R_{a,r}$ is the number of residues at the aligned position a of type r.

This statistic is compared to a null distribution, which is generated by calculating the statistic for random draws with replacement from the set of amino acid types (weighted by $E_{a,r}$ for each residue type), holding the number of mutations fixed. The final output is an upper-tail p-value.

Variant Residues: This test is very similar to *Reference Residue Distribution*, but tests for positions where the newly produced amino acids appear non-random. It is calculated as a chi-square goodness-of-fit test, but uses an empirical null distribution instead of a theoretical one. It

is a residue-level test, and recurrent mutations with identical residue changes are removed. The null distribution is set with mutations from outside of kinase domains. We then calculate the test statistic for each aligned position (*a*):

$$X_a^2 = \sum_v \frac{(O_{a,v} - E_{a,v})^2}{E_{a,v}}$$

$$E_{a,v} = \sum_r P_{r,v}\, O_{a,r}$$

Where *v* is the type of variant residue and *r* is the type of reference residue. $P_{r,v}$ refers to the probability that a mutation occurring at a residue of type *r* will result in a residue of type *v* (calculated based on the amino acid substitution frequencies observed outside of kinase domains), and $O_{a,r}$ is the observed number of mutations at aligned position *a* with reference residues of type *r*.

This statistic is compared to a null distribution, which is generated by calculating the statistic for random draws with replacement of amino acid types (weighted by $E_{a,v}$), holding the number of mutations fixed. The final output is an upper-tail p-value.

Cancer Genes: This test identifies positions with mutations that tend to occur in predicted cancer genes. It is a gene-level test, and multiple mutations that affect a single gene at a single position are only counted once. We associate each gene with a score that represents how likely the gene is to be related to cancer. Cancer genes have smaller scores on average (this is the "Unknown Score" produced by the RF5 model in chapter 2[96]).

To perform the test, we calculate the average score for the genes that are mutated at a given aligned position. We generate a null distribution by calculating the average score for random

draws of genes (weighted by the E_r that corresponds to each gene's aligned residue at the given position). The result of the test is a lower-tail p-value.

Gene Relatedness: This test identifies positions where mutated genes have kinase domains that are more closely relate to one another on average than expected by chance, given the mutation patterns observed outside of kinase domains. It is a gene-level test, and mutations that affect a single gene at a given position are only counted once. The distance matrix of all kinase domains in the dataset was calculated from the phylogenetic tree produced by ClustalOmega when it produced the alignment.

To perform the test, we calculate the average pair-wise distance for all genes that are mutated at a given aligned position. We generate a null distribution by calculating the average pair-wise distance for random draws of genes (weighted by the E_r that corresponds to each gene's aligned residue at the given position). The result of the test is a lower-tail p-value.

4.2.2 Imputation of Missing Data

The only variable with notable missingness was Cancer Type, which ~20% of mutations lacked. We found that excluding these mutations from the *Caner Types* test or including them under a "missing/other" category produced virtually identical results. The final analysis includes them as a separate category.

4.2.3 Experimental Procedures and Reagents

Experiments were performed as previously described[40]. Briefly, cDNA for KDR, TGBFR1 and CHEK2 were purchased from Addgene. ERBB2 cDNA was a gift from Dr. Dan Leahy (Johns Hopkins University, Baltimore). Mutations were introduced using QuikChange II site-directed mutagenesis (Agilent). Constructs were then shuttled into the pCFG5 retroviral vector

(which includes a zeocin resistance marker and IRES-GFP sequence) using the In-Fusion HD cloning system kit (Clonetech), and verified by full-length Sanger sequencing. For KDR, TGFBR1 and CHEK2, a c-terminal FLAG tag was introduced. For ERBB2, TGFBR1 and KDR, retroviral particles were produced using ϕNX amphotrophic packaging cells. NIH 3T3 cells (and IMCE cells, in the case of ERBB2 mutations) were spin-infected with virus, and selected under 10μg/ml zeocin for 3 weeks. Fluorescence was confirmed at >95% by flow cytometry or >90% by microscopy. Cells were serum starved for 6hrs before lysate harvesting for each of these three genes. In the case of TGBFR1 and KDR, cells were treated or untreated with ligand prior to harvesting. In the case of CHEK2, transient transfections were performed using LTX and Plus reagent from Thermo Fisher, using the manufacturers standard protocol in HEK 293T cells. Cells were lysed 24hrs after transfection. Transfection efficiency was confirmed by microscopy as >50% in all cases.

ERBB2/HER2 signaling was assayed using pHER2 and pMAPK levels[40]. TGFBR1 activity was assayed using pSMAD2 levels[116, 117]. KDR activity was assayed using pKDR[118] and pMAPK levels. CHEK2 was assayed with pS516, which is both an autophosphorylation site and necessary for full activation of CHEK2, and has been used previously as a proxy of CHEK2 activity[119-121].

NIH 3T3 cells were acquired from the American Type Culture Collection (ATCC). IMCE cells were a gift from Dr. Robert Whitehead (Vanderbilt University, Nashville). HEK 293T cells were a gift from Dr. Akhilesh Pandey (Johns Hopkins University, Baltimore). Antibodies used include HER2 from Thermo-Fisher (Ab-17), phospho-HER2 (pY1248) from Millipore (06-229), p44/42 MAPK from Cell Signaling Technologies (CST, 137F5), phospho p44/42 MAPK from CST (20G11), FLAG from Sigma-Aldrich (F3165), phospho-KDR (pY1175) from CST (19A10),

phospho-SMAD2 (S465/467) from CST (138D4), SMAD2 from CST (D43B4), phospho-CHEK2 (pS516) from CST (#2669). Ligand included $VEGF_{165}$ (#8065, 10min induction, 10ng/ml) from CST and TGFβ (20min induction, 5ng/ml).

4.3 Results

4.3.1 Description of Data

We used dGene to identify genes that have kinase domains, ultimately drawing 486 kinase domain sequences from 471 unique genes from Uniprot[122, 123]. These kinase domains were aligned using ClustalOmega with default settings[124]. The default settings are quite permissive to gaps in the alignment; this is acceptable for our purposes, since the analysis assumes that aligned residues have homolgous functions, and a more stringent alignment may violate the assumption. The final alignment has 1808 positions.

We draw 64,554 point mutations in these genes from our previous study, updated with additional mutations from the cBio portal (Figure 4.1A) [96, 125]. 21,917 of the mutations map to the kinase domains, while the remainder are outside the kinase domain. Duplicate mutations from multiple sources were removed. We limit scope to just point mutations (missense and silent changes), because other types of mutations like insertions and deletions often cannot be mapped to a single position on the alignment. 14,665 silent mutations are included in all analyses. Positions that are systematically enriched or depleted for silent mutations may be under negative or positve selection, respectively, making these events a valuable source of information[126, 127]. Moreover, there is evidence that some silent mutations have important functional consequences at the protein level[128, 129]. The mutations of our dataset come from 8,674 distinct patients, although the number of patients sequenced to generate these mutations is likely 10-20% higher, since some patients will have no mutations in any protein kinase.

4.3.2 Testing Aligned Positions

Mutations were mapped onto the alignment of human kinase domains (Figure 4.1B). Mutations in these genes that are outside the kinase domain define the null hypotheses, since they are produced by the same mutational proceses as mutations within the kinase domains, but are unlikely to be systematically enriched for biologically active mutations as the aligned kinase domains are. We developed a series of seven statistical tests to identify homolgous positions with non-random mutation patterns. The tests are described in Table 4.1. These tests can be calculated using basic approaches outlined in section 4.2.1.

4.3.3 Making the Functionality Map

Since the tests require multiple mutations and genes to be calculated, they were applied to the 831 positions (of 1808 total) that had mutations in at least two genes. The p-values from the tests were then combined using the Fisher procedure to produce a single p-value for the position[48]. These Fisher p-values were then adjusted for multiple-testing to control the false discover rate (FDR)[130]. We found 23 significantly mutated positions (SMP) with FDRs less than 0.10 (Table 4.2).

When viewed against the known structure of kinase domains, these SMPs compose a map of regions that may be important to kinase function. In Figure 4.2, we map these positions onto EGFR kinase domain crystal structure. The largest contiguous section of SMPs correspond to postions 4 though 8 in Figure 4.2; these are all very well known activation loop (A-loop) residues, and many are known to host important functional mutations (Figure 4.2). Additional SMPs are distributed throughout the N and C-lobes of the kinase domain.

4.3.4 Selecting Mutations for Validation

We first narrowed focus to just 14,541 unique missense mutations in the kinase domains (Figure 4.1A). We further focus on the 42 protein kinases which we confirmed or predicted as cancer genes in chapter 2, reducing the candidates to 1894 mutations (genes had to have greater than even chance of being either an oncogene or tumor suppressor according to the cancer gene analysis) [96]. Finally, we limited to scope to the 23 SMPs, resulting in 218 candidate mutations.

We selected ten of these mutations for functional testing in cell culture (Table 4.3). We sought a mix of recurrent and non-recurrent events, as well as mutations from diverse areas of the kinase domain. In particular, we tried to test mutations at a variety of SMPs, and avoid mutations that were closely related to well studied functional mutations. The mutations we selected represent hypotheses suggested by the functionality map, but which would likely be de-prioritized under other criteria. The mutations we selected include mutations in TGFBR1, CHEK2 and KDR, as well as the ERBB2 R868W mutation (Table 4.3). Five are non-recurrent, and seven are not homologuous to known functional mutations, to our knowledge.

Our group specializes in ERBB2/HER2, and we have particular interest in mutations occuring in the terminal portion of the C-lobe. Since none of the mutations observed in this region occurred at an SMP, we sought out additional mutations that otherwise did not meet the selection criteria. We chose two additional candidates. Position 1430 of the alignment is one of the most downstream SMPs; although no mutation was observed in ERBB2 at this position, an R to C change occurred at this position 33 times in 23 different genes, including one observation of EGFR R958C. We therefore constructed ERBB2 R966C, which corresponds to this position. Of the mutations that *were* observed in this region, S974F occurs at the most highly ranked position,

with an FDR value of 0.20. Although it does not occur at an SMP as defined in our analyis, we also included this mutation for testing.

4.3.5 Experimental Results

Using a previously described retroviral transduction system[40], we produced NIH 3T3 cells stably overexpressing both mutant and wild-type proteins for each of TGFBR1, KDR and ERBB2 (see section 4.2.3 for details). We found that we could not stably overexpress wild type CHEK2 in this setting: cells retained the selection marker, but stopped expressing the construct. Instead, CHEK2 experiments were performed using transient transection in HEK293T cells. TGFBR1, CHEK2 and KDR constructs were tagged with FLAG. All experiments were performed in duplicate or triplicate.

*TGFBR1***.** TGFBR1 (Transforming Growth Factor Beta Receptor 1) is a receptor S/T kinase. It has well appreciated functions in immune regulation as well as tissue remodeling. It is generally thought of as a tumor suppressor and acts to arrest the cell cycle[131], although it can also act as a pro-tumor factor in later disease progression, particularly by causing increased cell invasiveness, proliferation and migration[116, 132]. We tested two mutations in this gene. We found that NIH 3T3 cells overexpressing TGFBR1 S241L and L354P had reduced signaling when exposed to the ligand TGFβ when compared with wild type (Figure 4.3A).

*CHEK2***:** Checkpoint 2 is a cytoplasmic S/T kinase that has important functions in cell cycle control, specifically in DNA damage and repair. Much like P53, it is a well appreciated tumor suppressor[133]. We transiently transfected HEK 293T cells with wild type CHEK2 and five variants. We confirmed previous observations that wild type CHEK2 is constitutively activated under these conditions, as judged by phosphorylation at the autophosphorylation site S516[121].

We found that CHEK2 S372F, S372Y, and A392V all had less than 15% of the wild type phosphorylation. The highly recurrent mutant K373E had 45% of wild type phosphorylation, while A392S had 70% (Figure 4.3B, Supplementary Figure S4.1).

KDR/VEGFR2. KDR/VEGFR2 (Vascular Endothelial Growth Factor Receptor-2) is a receptor tyrosine kinase (RTK). KDR is a well-established oncogene with crucial roles in angiogenesis, although there is evidence of an autocrine function as well[134]. We tested two mutations in this gene. We found that both the R1032Q and S1100F mutations markedly reduced function, as judged by levels of phospho-KDR and signaling through MAPK after exposure to the ligand VEGF (Figure 4.3C).

ERBB2/HER2. ERBB2/HER2 is a member of the EGFR family of RTKs and a well known oncogene. Our lab has shown that point mutations in the HER2 kinase domain can trigger increased signaling and cell transformation in both breast[40] and colorectal cell lines[41]. We found that HER2 R966C and R868W caused a reduction-of-function as judged by levels of phospho-HER2 and MAPK signaling (Figure 4.3D). HER2 S974F did not produce a notable change in HER2 function when compared to the wild-type construct. These results were also confirmed in IMCE cells (Supplementary Figure S4.2).

The mutations we tested encompass a total of 74 patients with more than 11 distinct cancers (Table 4.4). The CHEK2 K373E variant was split among many cancer types, but 17 patients with lung adenocarcinoma carried it. The KDR variants R1032Q and S1100F were predominantly observed in 11 melanoma patients. Finally, the TGFBR1 S241L and ERBB2 R868W mutations were found in colorectal patients.

4.4 Discussion

In this study, we hypothesized that somatic cancer mutations could be used to identify important functional regions within proteins. Specifically, we focused on the family of protein kinases, which are a conserved set of phosphotransferases that share homologous sequences and structural motifs. By mapping mutations onto the alignment of protein kinases and applying a panel of statistical tests, we were able to identify homologous positions that bear mutations which appear non-random. Since mutations are pooled across all family members, these positions should be broadly important to the function of many different protein kinases.

We found 23 significantly mutated positions (SMPs) within the kinase alignment. SMPs were found throughout the kinase domains, with a particular enrichment in and around the A-loop. Many of these SMPs contain well characterized activating mutations (Figure 4.2). We tested twelve distinct mutations found in several genes with diverse relationships with cancer development. We purposely focused on highly novel mutations, including many that are rare or non-recurrent, and avoiding mutations with that are closely related to well-studied functional mutations. Ten of these mutations were observed in the dataset at one of the SMPs, and an eleventh (ERBB2 R966C) was present at an SMP but not directly observed (we also tested ERBB2 S974F, which was not present at an SMP). All eleven mutations in SMPs noticeably reduced signaling through the corresponding kinase. These mutations were observed in 74 patients with eleven cancer types, with particularly large numbers of these mutations occurring in colorectal carcinomas, lung adenocarcinomas, and melanomas.

The fact that all eleven tested mutations found at SMPs reduced function is an important finding. It illustrates the importance of functional characterization of mutations, particularly given the diverse roles protein kinases play in cancer development. Our previous study showed that protein

kinases include many predicted tumor suppressors as well as oncogenes[96]. For instance, of the 42 predicted cancer genes we focused on when selecting mutations to test (Figure 4.1A), 17 are predicted to be tumor suppressors, and 25 are predicted oncogenes. In tumor suppressors, focus is often on truncating events like frameshift indels; in this study, we found that both highly recurrent mutations (like CHEK K373E) and rare mutations (like CHEK S372F/Y and A293V) in tumor suppressors can also cause loss- or reduction-of-function. Similarly, while it may be tempting to assume that missense mutations in oncogenes are either neutral or gain-of-function, this work shows that mutations in these genes can be loss-of-function (for instance, KDR R1032Q and S1100F). As it becomes more common for patients to have their tumors exome or genome sequenced, this knowledge will be crucial in identifying events that are most like to underpin their disease.

There are many potential extensions to this study, encompassing multiple fields. We have tested only a small fraction of the mutations at the SMPs we identified. Direct follow up studies, particularly on ROF mutations in the tumor suppressors TGFBR1 and CHEK2 will be necessary before these mutations can be confirmed as *bona fide* cancer drivers. Many other mutations are found at other SMPs, and our results suggest that testing these mutations could be fruitful, particularly if present in genes with therapeutic implications. Our results also have implications for the structural understanding of kinase signaling: for instance, the ERBB2 R966C mutation demonstrates the importance of the C-lobe to kinase function, but the exact role this region plays is not fully understood.

Our methods can also be applied in other settings. Although we have focused on the kinase family, none of our methods are kinase-specific. Our analysis is equally compatible with other conserved gene or domain families: for instance, other gene families of broad importance to

cancer development include nuclear hormone receptors[135] and G-protein coupled receptors[136]. Our methods will also become more precise as data volumes continue to increase. For instance, within the protein kinases, a more accurate alignment and better functional homology may be observed exclusively among tyrosine kinases, or serine/threonine kinases. With larger datasets, the number of genes necessary to complete this analysis will shrink, allowing increasing granularity. Our methods can even be adapted to single genes, provided a sufficient density of observed variants.

In conclusion, we have demonstrated the use of somatic mutations to identify functional positions and mutations within gene families. We developed several statistical approaches for identifying positions with non-random mutations, aggregating mutations across homologous positions in the human kinome to do so. We identified 23 significantly mutated positions, and tested eleven mutations found at these positions from several genes. We confirmed all eleven as causing reductions in kinase function. Mutations that reduce the function of tumor suppressors are particularly promising as candidate cancer drivers, though other mutations at these SMPs warrant study as well. Our methods are highly extensible, providing a framework for using somatic cancer data to identify functionally important regions in proteins, and eventually identifying mutations that are relevant to cancer development and growth.

Test	Description
Mutation Number	Detects elevated numbers of mutations using a poisson distribution.
Patients	Uses a chi-square statistic to detect deviations from expected patient distribution.
Cancer Types	Uses a chi-square statistic to detect deviations from expected cancer type distribution.
Reference Residues	Uses a chi-square statistic to detect deviations from expected distribution of mutated residues.
Variant Residues	Uses a chi-square statistic to detect deviations from expected distribution of variant residues.
Cancer Genes	Detects sets of mutated genes that are enriched in predicted cancer genes.
Gene Relatedness	Detects sets of mutated genes that are more related than expected.

Table 4.1. Summary of statistical tests for aligned gene families. See section 4.2.1 for details.

Aligned Position	Mutation Number	Patients	Cancer Types	Reference Residues	Variant Residues	Cancer Genes	Gene Relatedness	Combined Fisher P	Combined Fisher FDR
145	6.775E-02	8.350E-01	3.100E-02	1.860E-02	5.321E-01	7.120E-02	2.200E-03	3.111E-04	**2.468E-02**
200	3.498E-04	6.509E-01	2.402E-01	8.210E-02	5.279E-01	2.040E-02	5.726E-01	1.571E-03	**6.218E-02**
205	8.125E-04	3.960E-01	3.621E-01	5.356E-01	2.000E-03	1.600E-03	4.160E-02	4.103E-06	**4.871E-04**
246	1.078E-03	3.010E-02	1.843E-01	6.160E-02	6.437E-01	1.507E-01	6.587E-01	1.407E-03	**5.870E-02**
254	1.097E-03	7.889E-01	5.096E-01	1.493E-01	5.100E-02	5.920E-02	7.750E-02	1.051E-03	**5.125E-02**
258	1.948E-04	6.281E-01	2.952E-01	4.209E-01	6.859E-01	2.600E-03	2.032E-01	5.135E-04	**2.681E-02**
717	3.033E-02	1.100E-02	9.733E-01	4.447E-01	8.939E-01	1.600E-03	2.540E-02	4.961E-04	**2.681E-02**
731	3.217E-03	2.248E-01	4.645E-01	2.248E-01	1.192E-01	4.417E-01	8.600E-03	1.813E-03	**6.549E-02**
820	1.481E-04	4.870E-02	4.882E-01	1.126E-01	4.400E-03	3.271E-01	4.001E-01	5.100E-05	**5.297E-03**
828	5.983E-01	1.400E-02	8.800E-03	6.380E-02	8.210E-02	5.226E-01	1.174E-01	1.413E-03	**5.870E-02**
889	2.649E-03	3.000E-04	6.500E-02	7.361E-01	6.670E-02	2.000E-04	4.206E-01	2.278E-07	**3.786E-05**
891	6.903E-11	2.810E-02	5.600E-03	5.004E-01	2.874E-01	1.610E-02	2.105E-01	3.414E-11	**9.458E-09**
892	1.052E-07	5.348E-01	3.260E-02	5.362E-01	7.099E-01	5.655E-01	9.128E-01	7.155E-05	**6.606E-03**
893	5.663E-14	4.741E-01	7.290E-02	2.694E-01	3.064E-01	7.100E-03	7.039E-01	6.819E-12	**2.834E-09**
894	5.000E-05	5.000E-05	5.000E-05	4.430E-02	8.560E-02	8.300E-03	2.053E-01	6.821E-12	**2.834E-09**
895	2.471E-02	8.411E-01	1.624E-01	3.093E-01	5.500E-03	5.640E-02	9.540E-02	1.692E-03	**6.392E-02**
923	1.859E-05	9.208E-01	1.000E-03	2.688E-01	1.738E-01	1.546E-01	6.800E-03	6.802E-07	**9.420E-05**
941	1.744E-04	5.129E-01	5.700E-03	7.395E-01	6.822E-01	6.590E-01	1.940E-02	3.563E-04	**2.468E-02**
945	8.188E-12	1.306E-01	3.420E-02	7.542E-01	9.766E-01	3.250E-02	1.073E-01	3.912E-10	**8.128E-08**
1134	1.550E-04	4.703E-01	1.844E-01	9.618E-01	5.140E-01	6.230E-02	7.200E-03	3.322E-04	**2.468E-02**
1430	9.693E-02	2.227E-01	1.826E-01	5.084E-01	3.000E-04	4.830E-02	5.738E-01	1.110E-03	**5.125E-02**
1467	1.096E-02	6.619E-01	5.360E-02	2.370E-02	9.936E-01	2.780E-02	2.180E-02	5.163E-04	**2.681E-02**
1683	4.750E-04	5.380E-02	5.666E-01	3.820E-02	5.738E-01	2.745E-01	6.350E-02	5.150E-04	**2.681E-02**

Table 4.2. Test results for 23 significantly mutated positions. The p-values produced by the seven tests, as well as the Fisher p-value that combines the seven tests, are listed for each of the 23 significantly mutated positions (defined as FDR < 0.1).

Gene	Mutation	Occurrences	Region	Homologous mutations	Effect on activity	Aligned Position
TGFBR1	S241L	5	N-lobe		↓↓↓	246
TGFBR1	L354P	1	A-loop	EGFR L858R	↓↓↓	891
CHEK2	S372F	1	A-loop		↓↓↓	892
CHEK2	S372Y	1	A-loop		↓↓↓	892
CHEK2	K373E	48	A-loop	ERBB2 R868W, ALK R1275Q	↓↓	893
CHEK2	A392S	1	C-lobe		↓	945
CHEK2	A392V	2	C-lobe		↓↓↓	945
KDR	R1032Q	6	N-lobe		↓↓↓	820
KDR	S1100F	7	C-lobe		↓↓↓	1134
ERBB2	R868W	1	A-loop	CHEK2 K373E, ALK R1275Q	↓↓↓	893
ERBB2	R966C	0[†]	C-lobe	EGFR R958C	↓↓↓	1430
ERBB2	S974F	1	C-lobe		-	1462[‡]

Table 4.3. Summary of mutations to be functionally tested. Key: strongly inactivating (↓↓↓), moderately inactivating (↓↓), modestly inactivating (↓), neutral (-). [†]ERBB2 R966C was not directly observed in the dataset, but this amino acid substitution is common at this SMP in other genes. [‡]ERBB2 S974F is the most highly ranked of the observed ERBB2 C-lobe mutations as judged by position significance, but this position did not meet the FDR<0.1 criteria common to the other mutations listed. The alignment column corresponding to Table 4.2 is indicated.

Gene	Mutation	BLCA	CESC	COAD	HNSC	KIRC	LGG	LUAD	PRAD	SKCM	STAD	UCEC	Other	Recurrence
KDR	R1032Q			1						4			1	6
KDR	S1100F									7				7
TGFBR1	S241L			3									2	5
TGFBR1	L354P										1			1
CHEK2	S372F				1									1
CHEK2	S372Y												1	1
CHEK2	K373E	3	2	6	4	4	3	17	4	1	2	1	1	48
CHEK2	A392S												1	1
CHEK2	A392V							1		1				2
ERBB2	R868W			1										1
ERBB2	R966C													0
ERBB2	S974F										1			1
Patients Mutated		3	2	11	5	4	3	18	4	13	4	1	6	74
Total Patients*		254	39	435	341	369	197	809	245	511	264	231	2381	

Table 4.4. Tested mutations by cancer type. The distribution of each tested mutation among cancer types is listed. Abbreviations: BLCA=bladder carcinoma, CESC=cervical squamous cell carcinoma and endocervical adenocarincoma, COAD=colorectal adenocarcinoma, HNSC=head and neck squamous cell carcinoma, KIRC=kidney renal clear cell carcinoma, LGG=brain lower grade glioma, LUAD=lung adenocarcinoma, PRAD=prostate adenocarcinoma, SKCM=melanoma, STAD=stomach adenocarcinoma, UCEC=uterine corpus endometrial carcinoma. *The total patients reflect samples in the kinase dataset, which may be 10-20% fewer than were sequenced in the original studies, since some patients have no mutations in any kinase gene.

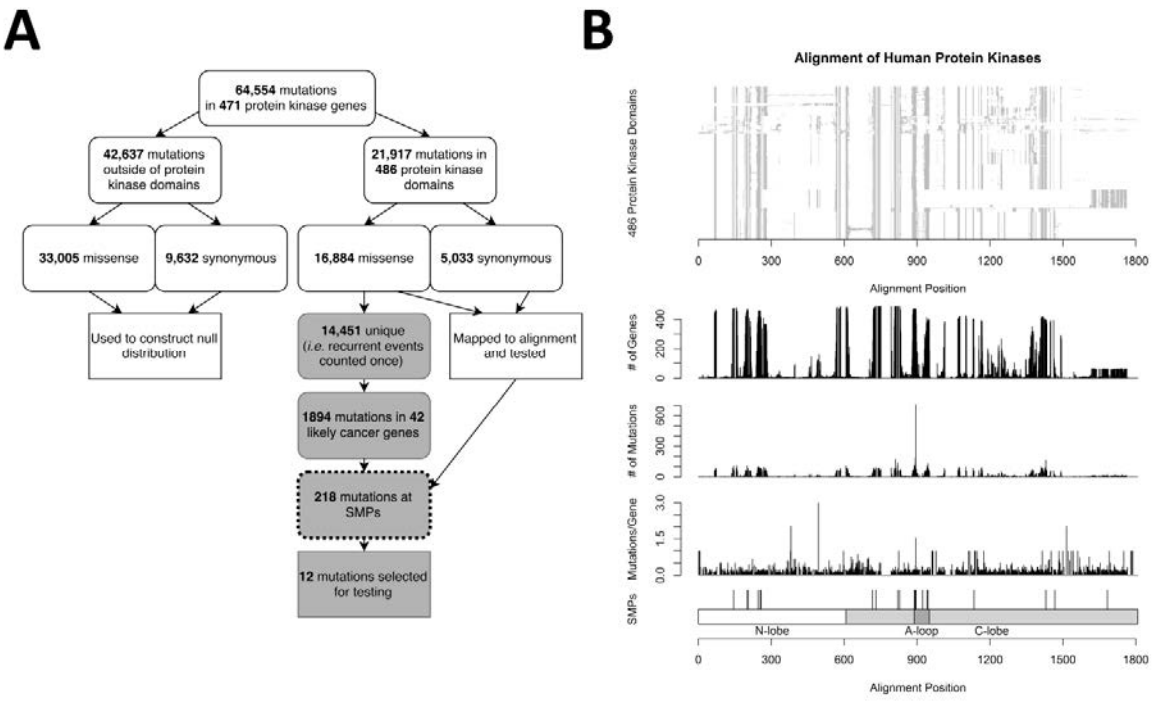


Figure 4.1. Summary of mutations in the kinome. A) The use of mutations in the study. The process of choosing mutations for experimentation is in grey; the use of Significantly Mutated Positions (SMPs) is outlined. B) Mapping of mutations to the protein kinase alignment. The location of 23 identified SMPs is indicated at the bottom, as well as the major regions of the aligned domains.

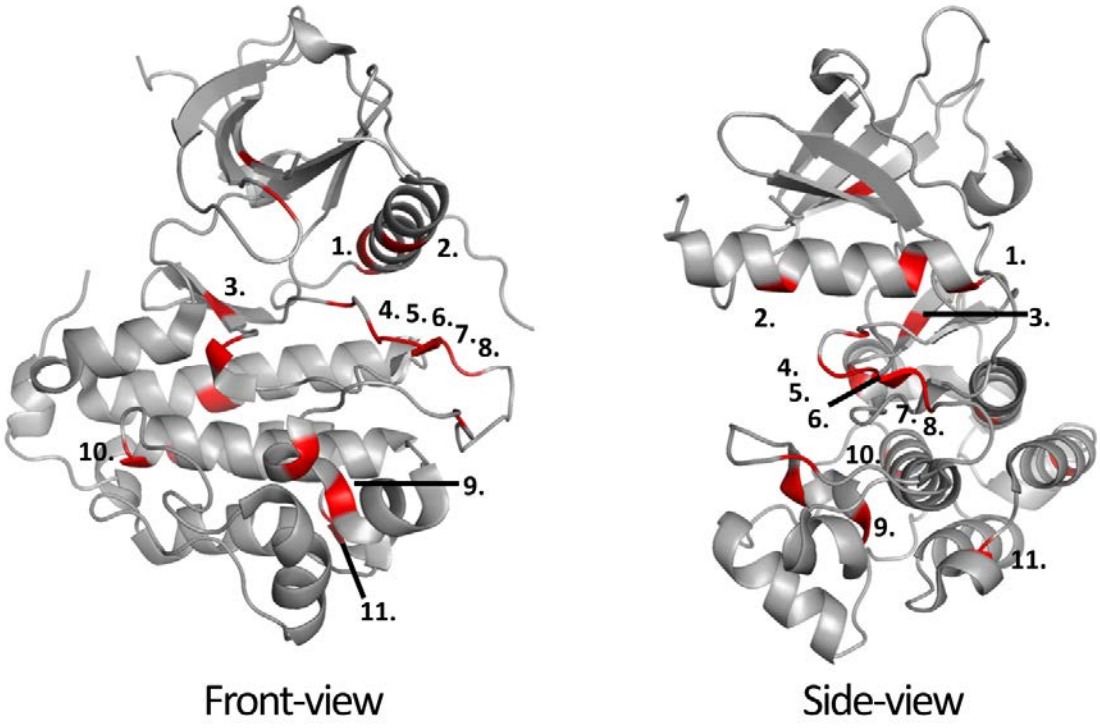


Figure 4.2. Significantly Mutated Positions as they appear on the EGFR kinase. The 20 of 23 SMPs to which EGFR aligns are portrayed. *Well-known functional mutations* and mutations to be validated *in vitro* are positioned as follows: 1) *ERBB2 V777L*. 2) TGFBR1 S241L. 3) KDR R1032Q. 4) *EGFR L858R*, TGFRB1 L354P. 5) CHEK2 S372F/Y. 6) *ALK R1275Q*, ERBB2 R868W, CHEK2 K373E. 7) *BRAF V600E*. 8) *BRAF K601E*. 9) CHEK2 A392S/V. 10) KDR S1100F. 11) ERBB2 R966C.

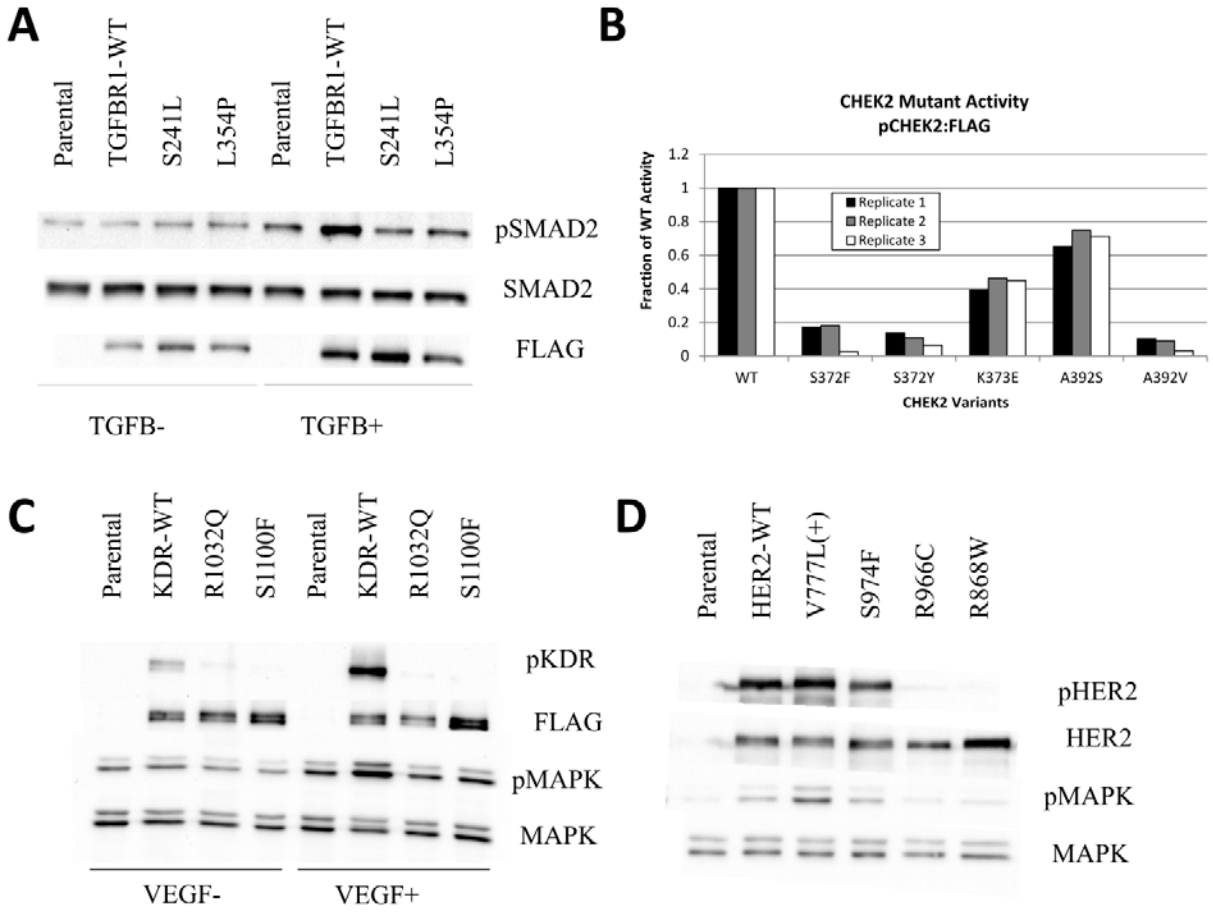


Figure 4.3. Functional validation of TGFBR1, CHEK2, KDR and ERB2 mutations. A) The mutations TGFBR1 S241L and L354P were tested in NIH 3T3 cells in the absence and presence of ligand. B) The mutations CHEK S372F/Y, K373E, and A392S/V were tested by transient transfection of HEK 293T cells. C) The mutations KDR R1032Q and S1100F were tested in NIH 3T3 cells in the absence and presence of ligand. D) The mutations ERBB2 R868W, R966C and S974 were tested in NIH 3T3 cells. See section 4.2.3 for details.

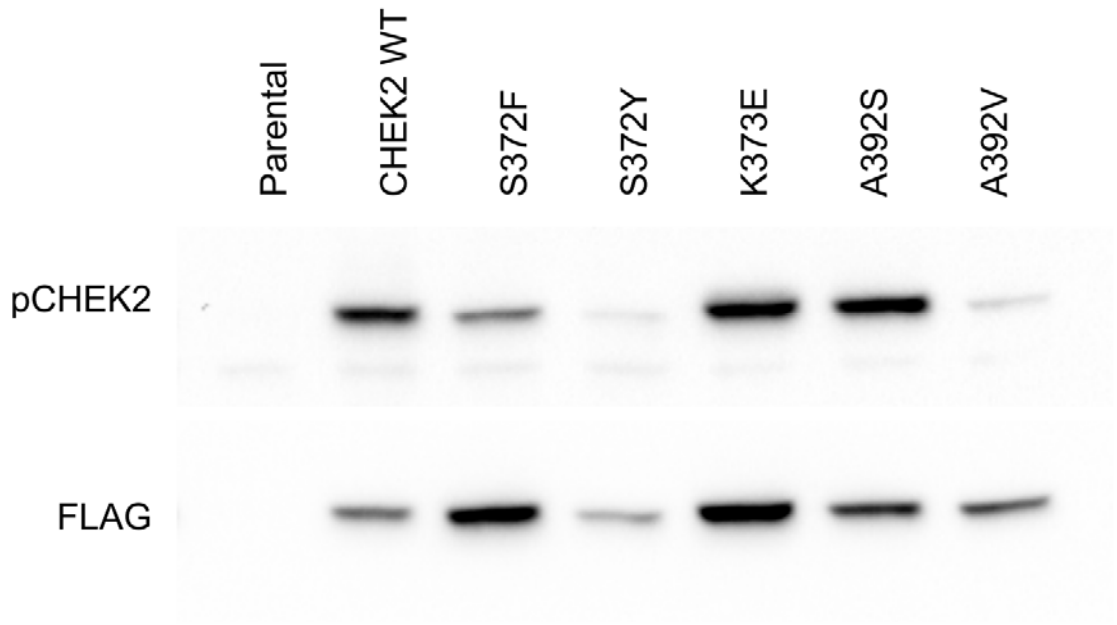


Figure S4.1. CHEK2 activity in HEK 293T cells. CHEK-FLAG constructs were transiently transfected into cells and basal levels of phosphorylation assayed without induction. Image is representative of three replicates.

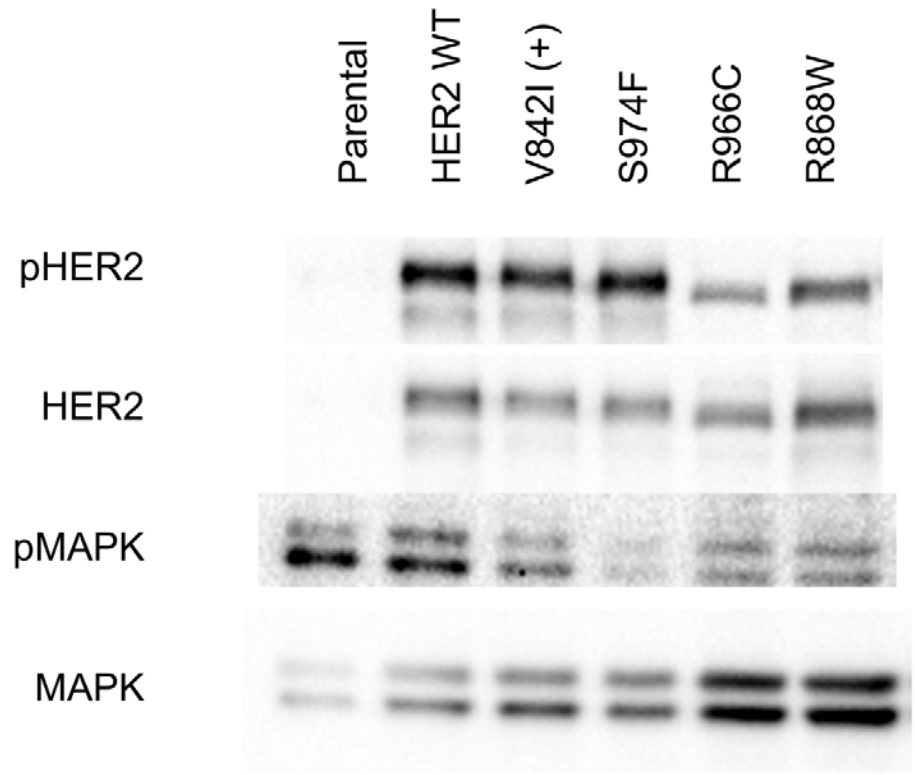


Figure S4.2. Confirmation of ERBB2/HER2 results in IMCE cells. We found that IMCE cells that stably overexpressed R966C and R868W HER2 had lower phospho-HER2 levels than cells expressing the wild type protein, confirming results from NIH 3T3 cells. The portrayed results are representative of two independent experiments.

5. Conclusions

5.1 Summary of Results

There is great promise in cancer genome sequencing. By scrutinizing the genetic alterations that occur within tumors, the mechanisms of cancer development can be catalogued, new drug targets identified, and better treatments can be developed. As whole-genome and exome sequencing grows more economical, this data will become increasingly common, and researchers will have the ability to study cancers with both greater granularity and a population-wide perspective. RNA sequencing and proteomic datasets will likely be added to these data, producing an unprecedented portrait of intracellular processes during disease development and progression.

However, to convert this information into improved treatments and outcomes for patients will likely require targeted, experimental follow-up. Whereas producing sequencing data is increasingly high-throughput and inexpensive, validating and elucidating biological mechanisms remains low-throughput and labor-intensive. Connecting the one to the other is akin to attaching a firehose to a drinking straw.

In the context of cancer genome sequencing, part of the solution to this dilemma is recognizing that most observed somatic mutations are incidental to tumor development. While “drivers” of tumor development often need to be explored through experimental approaches, the far more numerous “passengers” can be de-prioritized, as they provide little direct information regarding mechanisms of tumor development. In this dissertation, we developed three interrelated approaches to the problem of filtering out passenger events and prioritizing drivers for experimental follow-up.

In chapter 2, we developed new methods for using somatic mutations to identify putative cancer genes that possess non-random sets of mutations. We designed a series of five simple statistical tests which identify cancer genes based on several signals of positive selection and compared them with existing methods. We found that known tumor suppressors were easier to detect than oncogenes, but that our new methods outperformed existing approaches in detecting both gene classes. In particular, *Patient Distribution*, which identifies genes with mutations occurring in non-random sets of patients, outperformed existing methods in identifying the gene panel as a whole, with particularly stark improvements in detecting oncogenes. However, other tests had complementary strengths. For instance, *Truncation Rate*, which detects genes with either higher or lower than usual rates of truncating events, could separate tumor suppressors and oncogenes from each other very effectively, whereas *Patient Distribution* could not. A single model which incorporated all five tests was able to identify known cancer genes better than existing methods, and could also separate putative cancer genes into likely oncogenes or tumor suppressors. The ability to segregate putative cancer genes into likely oncogenes and likely tumor suppressors is critical, since the anticipated role of a gene will often inform how it is handled in the lab (for instance, putative oncogenes may be studied as potential therapeutic targets, whereas putative tumor suppressors could be useful for developing model systems).

In chapter 3, we developed ParsSNP, which uses a new paradigm for identifying likely driver mutations based on various mutation descriptors without the need for pre-defined training labels. ParsSNP uses an expectation-maximization framework: the E-step operates on the assumption that driver mutations are more equitably distributed among samples than mutations in general, while the M-step ensures that drivers are definable in terms of the descriptors. It ultimately identifies a parsimonious set of putative driver mutations that explain cancer incidence broadly.

We tested ParsSNP in four benchmarks, each representing qualities expected of driver mutations. We found that ParsSNP consistently outperformed existing methods. In particular, ParsSNP performed very well at identifying somatic mutations that are recurrent in multiple samples, and mutations that occur in known or likely cancer genes as defined by the Cancer Gene Census. It also performed as well or better than existing methods in detecting rare, experimentally validated driver events among numerous common polymorphisms, and in predicting which mutations disrupt P53 function. We conclude that ParsSNP is superior to existing methods for identifying drivers in cancer. ParsSNP could be used as a screening method to rule out large numbers of likely passengers using a permissive threshold, or it could be used as a final test to choose between mutations that have already been deemed as likely drivers. It can provide guidance regarding a single mutation, or a set of mutations observed in a single patient. Since it can be applied on a mutation-by-mutation basis, ParsSNP can be used in a wide variety of contexts.

In chapter 4, we developed methods for analyzing mutations across gene families, focusing in particular on the human protein kinases. We adapted many of the ideas from chapter 2 to the problem of identifying homologous positions in kinase domains that possess non-random sets of mutations. Specifically, we developed a panel of seven statistical tests, each sensitive to a particular signal of selection. Using these tests, we were able to identify 23 significantly mutated positions, which together create a functionality map of a stereotypical kinase. This map highlighted important residues throughout the N and C lobes of the kinase, as well as the regions abutting the A-loop. We tested eleven mutations from these positions, distributed between several known and predicted cancer genes as judged by the results of chapter 2. We found that all eleven reduced signaling through the affected kinase. We concluded that mutations that reduce kinase signaling are common in cancer and that functional characterization of missense

mutations is crucial. The effect of these mutations should not be assumed, even if the affected gene is well-characterized as a tumor suppressor or oncogene. The methods we developed can also be applied to other conserved domains and structural motifs, allowing functional insights to be gleaned from somatic mutation patterns. For instance, this type of homology-based approach could be appropriate for other well-studied gene families such as nuclear hormone receptors or G-protein coupled receptors. For groups that specialize in particular gene families, the methods developed in this chapter are a powerful means of prioritizing mutations for investigation, particularly when combined with the cancer gene analysis of chapter 2.

Although we have generally focused on each method in isolation, there are many scenarios in which a combination of these methods could be used to direct experimental projects. For instance, our group has particular competence in the study of kinases, especially receptor tyrosine kinases. Using the results from chapter 2, a large set of candidate mutations can be quickly limited to just those occurring in known and predicted cancer genes. These mutations can then be analyzed using the methods outlined in chapter 4, yielding a subset of mutations that occur at positions which are likely to have functional importance to the enzyme. If a large number of mutations remain, a functional impact score like ParsSNP, discussed in chapter 3, can be used to select final candidates for experimental characterization. The methods outlined in this dissertation are a powerful and flexible toolset, which should be broadly useful to experimental biologists studying cancer genome sequencing results in a variety of genes, cancer types and model systems.

5.2 Future Directions

Iterative improvements to the methods described in this dissertation are warranted, as is biological characterization of many of the genes and mutations we have identified as putative cancer drivers. These ideas are discussed in the relevant chapter sections. Here we will focus on larger trends and how this dissertation can remain relevant in a quickly changing field.

The most important trend is that greater volumes and diversity of data are becoming available. Since the work of this dissertation began, major advances in sequencing and annotating non-protein-coding variants have been made[61, 115, 137]. Moreover, in addition to being DNA sequenced, tumor samples are increasingly annotated with transcriptomic, proteomic, or epigenomic data from a variety of platforms[14, 15, 21]. The methods discussed in this dissertation were developed for somatic protein-coding mutations, but are designed to be modular and flexible. Adapting and elaborating them to work with broader data types will likely be a fruitful endeavor.

The methods outlined in chapter 2 can be adapted to new data in several ways. This was a gene-based analysis, but our methods are largely applicable to other functional units of DNA. For instance, as non-protein coding variants become increasingly available, this type of analysis could help to identify regulatory regions (*e.g.* enhancers) or non-coding transcripts (*e.g.* lncRNAs) that are experiencing selection during tumor progression. Some of the tests we developed, such as *Cancer Type Distribution* and *Patient Distribution,* could be applied directly to any arbitrary set of mutations, while others could be adapted or replaced with relatively modest effort. Another area of advancement would be to leave the focus on genes but adapt the methods to other data types. For instance, tests that detect genes with systematic copy-number alterations, expression level changes, epigenomic alterations or even post-translational

modifications could be incorporated into the analysis, allowing cancer genes to be identified through the integration of multiple platforms.

The algorithm described in chapter 3 can also be adapted along similar lines. ParsSNP relies on a descriptor table to make predictions. As long as appropriate descriptors can be assembled, any set of DNA alterations can be analyzed in our expectation-maximization framework. For instance, as the ability to informatively annotate non-protein-coding variants grows, ParsSNP could be adapted to whole-genome sequencing datasets. There is also no limit to the types of data that can be incorporated into the descriptor table. For instance, an expanded descriptor table could include functional data from sources like ENCODE, or annotations from transcriptomic or proteomic datasets.

Much like the cancer gene analysis of chapter 2, the gene family analysis of chapter 4 can be adapted to larger and more diverse datasets. In principle, the analysis applies to any conserved DNA elements where homologous sequences imply similar functions. This could include regulatory regions of DNA or non-protein-coding RNA transcripts, for instance. The tests underpinning the analysis are completely modular; while many are designed to function in the context of gene families, they can be adapted or replaced with moderate effort, allowing the analysis to be used on other units of DNA.

The field of cancer genomics is quickly growing to encompass larger volumes of diverse data types. The methods we have described in this dissertation are substantially modular, built with simple descriptors and statistical tests, and meant to be used alone or in combination with one another. With creativity and diligence, they can be adapted to a wide variety of settings.

5.3 Final Thoughts

It has been 45 years since President Nixon declared the beginning to the "War on Cancer," and at times it can seem as though there is no end in sight. But the scientists of that era could not have imagined the tools available to us today. The first cancer genome was only sequenced in 2008, and new insights will undoubtedly lead to many new and exciting opportunities.

Making sense of the data produced in sequencing efforts is a major challenge for traditional biologists. Millions of somatic mutations have been observed in thousands of patients, only a tiny fraction of which will ever be followed-up experimentally. The fact that so many mutations occur incidentally to tumor development is both a challenge and an opportunity. These passenger events are challenging because they can discourage risk-taking when selecting mutations for further study. But they provide an opportunity, because ruling them out up-front with *in silico* approaches would allow the few remaining drivers to be studied in-depth.

We have described several interrelated approaches to this problem. In chapter 2, we developed methods for identifying cancer genes, whose somatic mutations display evidence of positive selection. In chapter 3, we explored unsupervised methods for identifying driver mutations, allowing us to produce more generalizable results and better predictions than existing methods are capable of. Finally, in chapter 4 we mapped mutations to the aligned human kinome, allowing us to use mutations to identify functional regions in these enzymes and prioritize mutations at those positions for experimental validation. Together, these approaches are a framework for identifying the highest-priority mutations for experimental study, ultimately leading to new treatments and better outcomes for cancer patients.

References

1. Balmain, A., *Cancer genetics: from Boveri and Mendel to microarrays.* Nat Rev Cancer, 2001. **1**(1): p. 77-82.

2. Stehelin, D., et al., *DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA.* Nature, 1976. **260**(5547): p. 170-173.

3. Friend, S.H., et al., *A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma.* Nature, 1986. **323**(6089): p. 643-6.

4. Ley, T.J., et al., *DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.* Nature, 2008. **456**(7218): p. 66-72.

5. The Cancer Genome Atlas Research Network, *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-1068.

6. The Cancer Genome Atlas Research Network, *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-615.

7. The Cancer Genome Atlas Research Network, *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**: p. 61-70.

8. The Cancer Genome Atlas Research Network, *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-337.

9. The Cancer Genome Atlas Research Network, *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-525.

10. The Cancer Genome Atlas Research Network, *Integrated genomic characterization of endometrial carcinoma.* Nature, 2013. **497**(7447): p. 67-73.

11. The Cancer Genome Atlas Research Network, *Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia*. New England Journal of Medicine, 2013. **368**(22): p. 2059-2074.

12. Brennan, Cameron W., et al., *The Somatic Genomic Landscape of Glioblastoma.* Cell, 2013. **155**(2): p. 462-477.

13. Davis, Caleb F., et al., *The Somatic Genomic Landscape of Chromophobe Renal Cell Carcinoma.* Cancer Cell, 2014. **26**(3): p. 319-330.

14. Ceccarelli, M., et al., *Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.* Cell, 2016. **164**(3): p. 550-563.

15. Abeshouse, A., et al., *The Molecular Taxonomy of Primary Prostate Cancer.* Cell, 2015. **163**(4): p. 1011-1025.

16. Agrawal, N., et al., *Integrated Genomic Characterization of Papillary Thyroid Carcinoma.* Cell, 2014. **159**(3): p. 676-690.

17. Akbani, R., et al., *Genomic Classification of Cutaneous Melanoma.* Cell, 2015. **161**(7): p. 1681-1696.

18. The Cancer Genome Atlas Research Network, *Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas.* New England Journal of Medicine, 2015. **372**(26): p. 2481-2498.

19. Ciriello, G., et al., *Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer.* Cell, 2015. **163**(2): p. 506-519.

20. The Cancer Genome Atlas Research Network, *Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma.* New England Journal of Medicine, 2015. **374**(2): p. 135-145.

21. Zheng, S., et al., *Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma.* Cancer Cell, 2016. **29**(5): p. 723-736.

22. The Cancer Genome Atlas Research Network, *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-582.

23. The Cancer Genome Atlas Research Network, *Comprehensive molecular characterization of clear cell renal cell carcinoma.* Nature, 2013. **499**(7456): p. 43-49.

24. The Cancer Genome Atlas Research Network, *Comprehensive molecular characterization of gastric adenocarcinoma.* Nature, 2014. **513**(7517): p. 202-209.

25. The Cancer Genome Atlas Research Network, *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-550.

26. The Cancer Genome Atlas Research Network, *Comprehensive molecular characterization of urothelial bladder carcinoma.* Nature, 2014. **507**(7492): p. 315-322.

27. Pleasance, E.D., et al., *A small-cell lung cancer genome with complex signatures of tobacco exposure.* Nature, 2010. **463**(7278): p. 184-190.

28. Ellis, M.J., et al., *Whole-genome analysis informs breast cancer response to aromatase inhibition.* Nature, 2012. **486**(7403): p. 353-60.

29. Shah, S.P., et al., *The clonal and mutational evolution spectrum of primary triple-negative breast cancers.* Nature, 2012. **486**(7403): p. 395-399.

30. Stephens, P.J., et al., *The landscape of cancer genes and mutational processes in breast cancer.* Nature, 2012. **486**(7403): p. 400-4.

31. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.

32. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333-339.

33. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic Acids Research, 2011. **39**: p. D945-D950.

34. Hoadley, K.A., et al., *Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin.* Cell, 2014. **158**(4): p. 929-944.

35. Pabinger, S., et al., *A survey of tools for variant analysis of next-generation genome sequencing data.* Brief Bioinform, 2013. **15**(2): p. 256-78.

36. Zhang, J., et al., *The impact of next-generation sequencing on genomics.* Journal of Genetics and Genomics, 2011. **38**(3): p. 95-109.

37. Mardis, E.R., *A decade/'s perspective on DNA sequencing technology.* Nature, 2011. **470**(7333): p. 198-203.

38. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.

39. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing*. Genome Res, 2012. **22**(3): p. 568-76.

40. Bose, R., et al., *Activating HER2 Mutations in HER2 Gene Amplification Negative Breast Cancer.* Cancer Discovery, 2013. **3**(2): p. 224-237.

41. Kavuri, S.M., et al., *HER2 activating mutations are targets for colorectal cancer treatment.* Cancer Discov, 2015. **5**(8): p. 832-41.

42. Ben-Baruch, N.E., et al., *HER2-Mutated Breast Cancer Responds to Treatment With Single-Agent Neratinib, a Second-Generation HER2/EGFR Tyrosine Kinase Inhibitor.* J Natl Compr Canc Netw, 2015. **13**(9): p. 1061-4.

43. Arteaga, C.L., et al., *Treatment of HER2-positive breast cancer: current status and future perspectives.* Nat Rev Clin Oncol, 2012. **9**(1): p. 16-32.

44. Vogelstein, B., et al., *Cancer Genome Landscapes.* Science, 2013. **339**(6127): p. 1546-1558.

45. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.

46. Russler-Germain, D.A., et al., *The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers.* Cancer Cell, 2014. **25**(4): p. 442-54.

47. Dees, N.D., et al., *MuSiC: identifying mutational significance in cancer genomes.* Genome Res, 2012. **22**(8): p. 1589-98.

48. Gonzalez-Perez, A. and N. Lopez-Bigas, *Functional impact bias reveals cancer drivers.* Nucleic Acids Research, 2012. **40**(21): p. e169.

49. Tamborero, D., A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.* Bioinformatics, 2013. **29**(18): p. 2238-2244.

50. Reimand, J. and G.D. Bader, *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.* Molecular systems biology, 2013. **9**: p. 637.

51. Hua, X., et al., *DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies.* Am J Hum Genet, 2013. **93**(3): p. 439-51.

52. Gu, Y., et al., *Systematic interpretation of comutated genes in large-scale cancer mutation profiles.* Mol Cancer Ther, 2010. **9**(8): p. 2186-95.

53. Youn, A. and R. Simon, *Identifying cancer driver genes in tumor genome sequencing studies.* Bioinformatics, 2011. **27**(2): p. 175-181.

54. Davoli, T., et al., *Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome.* Cell, 2013. **155**(4): p. 948-62.

55. Tamborero, D., et al., *Comprehensive identification of mutational cancer driver genes across 12 tumor types.* Sci Rep, 2013. **3**: p. 2650.

56. Schroeder, M.P., et al., *OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action.* Bioinformatics, 2014. **30**(17): p. i549-55.

57. Tarca, A.L., et al., *Machine learning and its applications to biology.* PLoS Comput Biol, 2007. **3**(6): p. e116.

58. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat. Protocols, 2009. **4**(8): p. 1073-1081.

59. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **Chapter 7**: p. Unit7.20.

60. Carter, H., et al., *Identifying Mendelian disease genes with the variant effect scoring tool.* BMC Genomics, 2013. **14 Suppl 3**: p. S3.

61. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nature genetics, 2014. **46**(3): p. 310-315.

62. Mao, Y., et al., *CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features.* PLoS ONE, 2013. **8**(10): p. e77945.

63. Carter, H., et al., *Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.* Cancer Res, 2009. **69**(16): p. 6660-7.

64. Ionita-Laza, I., et al., *A spectral approach integrating functional genomic annotations for coding and noncoding variants.* Nat Genet, 2016. **48**(2): p. 214-220.

65. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), 1977: p. 1-38.

66. Lawrence, C.E. and A.A. Reilly, *An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.* Proteins, 1990. **7**(1): p. 41-51.

67. Tomasetti, C., et al., *Only three driver gene mutations are required for the development of lung and colorectal cancers.* Proceedings of the National Academy of Sciences, 2015. **112**(1): p. 118-123.

68. Ciriello, G., et al., *Mutual exclusivity analysis identifies oncogenic network modules.* Genome Res, 2012. **22**(2): p. 398-406.

69. Porta-Pardo, E. and A. Godzik, *e-Driver: a novel method to identify protein regions driving cancer.* Bioinformatics, 2014. **30**(21): p. 3109-3114.

70. Torkamani, A., G. Verkhivker, and N.J. Schork, *Cancer driver mutations in protein kinase genes.* Cancer Letters, 2009. **281**(2): p. 117-127.

71. Lahiry, P., et al., *Kinase mutations in human disease: interpreting genotype-phenotype relationships.* Nat Rev Genet, 2010. **11**(1): p. 60-74.

72. Torkamani, A., et al., *Congenital disease SNPs target lineage specific structural elements in protein kinases.* Proc Natl Acad Sci U S A, 2008. **105**(26): p. 9011-6.

73. Torkamani, A. and N.J. Schork, *Accurate prediction of deleterious protein kinase polymorphisms.* Bioinformatics, 2007. **23**(21): p. 2918-2925.

74. Torkamani, A. and N.J. Schork, *Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family.* Genomics, 2007. **90**(1): p. 49-58.

75. Torkamani, A. and N.J. Schork, *Prediction of Cancer Driver Mutations in Protein Kinases.* Cancer Research, 2008. **68**(6): p. 1675-1682.

76. ManChon, U., et al., *Prediction and prioritization of rare oncogenic mutations in the cancer Kinome using novel features and multiple classifiers.* PLoS Comput Biol, 2014. **10**(4): p. e1003545.

77. Dixit, A. and G.M. Verkhivker, *Hierarchical modeling of activation mechanisms in the ABL and EGFR kinase domains: thermodynamic and mechanistic catalysts of kinase activation by cancer mutations.* PLoS Comput Biol, 2009. **5**(8): p. e1000487.

78. Dixit, A., et al., *Sequence and structure signatures of cancer mutation hotspots in protein kinases.* PLoS One, 2009. **4**(10): p. e7485.

79. Dixit, A. and G.M. Verkhivker, *The energy landscape analysis of cancer mutations in protein kinases.* PLoS One, 2011. **6**(10): p. e26071.

80. Dixit, A. and G.M. Verkhivker, *Structure-Functional Prediction and Analysis of Cancer Mutation Effects in Protein Kinases.* Computational and Mathematical Methods in Medicine, 2014. **2014**: p. 24.

81. Fujita, P.A., et al., *The UCSC Genome Browser database: update 2011.* Nucleic Acids Research, 2011. **39**(suppl 1): p. D876-D882.

82. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic acids research, 2010. **38**(16): p. e164-e164.

83. Futreal, P.A., et al., *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-183.

84. Hanahan, D. and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.

85. Zhao, M., J. Sun, and Z. Zhao, *TSGene: a web resource for tumor suppressor genes.* Nucleic Acids Res, 2013. **41**(Database issue): p. D970-6.

86. Simonetti, F.L., et al., *Kin-Driver: a database of driver mutations in protein kinases.* Database (Oxford), 2014. **2014**: p. bau104.

87. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics, 1988: p. 837-845.

88. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

89. Zhang, J., et al., *The N-CoR-HDAC3 nuclear receptor corepressor complex inhibits the JNK pathway through the integral subunit GPS2.* Mol Cell, 2002. **9**(3): p. 611-23.

90. Towhid, S.T., et al., *Inhibition of colonic tumor growth by the selective SGK inhibitor EMD638683.* Cell Physiol Biochem, 2013. **32**(4): p. 838-48.

91. Kato, U., et al., *Role for phospholipid flippase complex of ATP8A1 and CDC50A proteins in cell migration.* J Biol Chem, 2013. **288**(7): p. 4922-34.

92. Sutherland, L.C., K. Wang, and A.G. Robinson, *RBM5 as a putative tumor suppressor gene for lung cancer.* J Thorac Oncol, 2010. **5**(3): p. 294-8.

93. Cai, Y., et al., *The NuRD complex cooperates with DNMTs to maintain silencing of key colorectal tumor suppressor genes.* Oncogene, 2014. **33**(17): p. 2157-2168.

94. Nagarajan, P., et al., *Role of chromodomain helicase DNA-binding protein 2 in DNA damage response signaling and tumorigenesis.* Oncogene, 2009. **28**(8): p. 1053-62.

95. Griffith, M., et al., *DGIdb: mining the druggable genome.* Nat Meth, 2013. **10**(12): p. 1209-1210.

96. Kumar, R.D., et al., *Statistically Identifying Tumor Suppressors and Oncogenes from Pan-Cancer Genome Sequencing Data.* Bioinformatics, 2015. **31**(22): p. 3561-3568.

97. Petitjean, A., et al., *Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database.* Hum Mutat, 2007. **28**(6): p. 622-9.

98. Reva, B., Y. Antipin, and C. Sander, *Predicting the functional impact of protein mutations: application to cancer genomics.* Nucleic Acids Research, 2011. **39**(17): p. e118.

99. Martelotto, L.G., et al., *Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations.* Genome Biol, 2014. **15**(10): p. 484.

100. Kakiuchi, M., et al., *Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma.* Nat Genet, 2014. **46**(6): p. 583-7.

101. Basheer, I.A. and M. Hajmeer, *Artificial neural networks: fundamentals, computing, design, and application.* Journal of Microbiological Methods, 2000. **43**(1): p. 3-31.

102. Zaretzki, J.M., et al., *Extending P450 site-of-metabolism models with region-resolution data.* Bioinformatics, 2015. **31**(12): p. 1966-1973

103. Hong, Y., *On computing the distribution function for the sum of independent and nonidentical random indicators.* Dep. Statit., Virginia Tech, Blacksburg, VA, USA, Tech. Rep. 11_2, 2011.

104. Venables, W.N. and B.D. Ripley, *Modern applied statistics with S.* 2002: Springer Science & Business Media.

105. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**: p. 77.

106. Shihab, H.A., et al., *Predicting the functional consequences of cancer-associated amino acid substitutions.* Bioinformatics, 2013. **29**(12): p. 1504-10.

107. Gonzalez-Perez, A., J. Deu-Pons, and N. Lopez-Bigas, *Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation.* Genome Med, 2012. **4**(11): p. 89.

108. Gonzalez-Perez, A. and N. Lopez-Bigas, *Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.* Am J Hum Genet, 2011. **88**(4): p. 440-9.

109. Olden, J.D. and D.A. Jackson, *Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks.* Ecological modelling, 2002. **154**(1): p. 135-150.

110. Guan, B., T.-L. Wang, and I.-M. Shih, *ARID1A, a Factor That Promotes Formation of SWI/SNF-Mediated Chromatin Remodeling, Is a Tumor Suppressor in Gynecologic Cancers.* Cancer Research, 2011. **71**(21): p. 6718-6727.

111. Kang, S., A.G. Bader, and P.K. Vogt, *Phosphatidylinositol 3-kinase mutations identified in human cancer are oncogenic.* Proc Natl Acad Sci U S A, 2005. **102**(3): p. 802-7.

112. Koo, B.-K., et al., *Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors.* Nature, 2012. **488**(7413): p. 665-669.

113. Kim, V.N., N. Kataoka, and G. Dreyfuss, *Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex.* Science, 2001. **293**(5536): p. 1832-6.

114. Huang, F.W., et al., *Highly recurrent TERT promoter mutations in human melanoma.* Science, 2013. **339**(6122): p. 957-9.

115. Lee, D., et al., *A method to predict the impact of regulatory variants from DNA sequence.* Nat Genet, 2015. **47**(8): p. 955-961.

116. Kojima, Y., et al., *Autocrine TGF-beta and stromal cell-derived factor-1 (SDF-1) signaling drives the evolution of tumor-promoting mammary stromal myofibroblasts.* Proc Natl Acad Sci U S A, 2010. **107**(46): p. 20009-14.

117. Kong, B., et al., *AZGP1 is a tumor suppressor in pancreatic cancer inducing mesenchymal-to-epithelial transdifferentiation by inhibiting TGF-beta-mediated ERK signaling.* Oncogene, 2010. **29**(37): p. 5146-58.

118. Antonescu, C.R., et al., *KDR Activating Mutations in Human Angiosarcomas are Sensitive to Specific Kinase Inhibitors.* Cancer research, 2009. **69**(18): p. 7175-7179.

119. Anderson, V.E., et al., *CCT241533 is a potent and selective inhibitor of CHK2 that potentiates the cytotoxicity of PARP inhibitors.* Cancer Res, 2011. **71**(2): p. 463-72.

120. Gire, V., et al., *DNA damage checkpoint kinase Chk2 triggers replicative senescence.* Embo j, 2004. **23**(13): p. 2554-63.

121. Schwarz, J.K., C.M. Lovly, and H. Piwnica-Worms, *Regulation of the Chk2 protein kinase by oligomerization-mediated cis- and trans-phosphorylation.* Mol Cancer Res, 2003. **1**(8): p. 598-609.

122. Kumar, R.D., et al., *Prioritizing Potentially Druggable Mutations with dGene: An Annotation Tool for Cancer Genome Sequencing Data.* PLoS One, 2013. **8**(6): p. e67980.

123. UniProt Consortium., *Activities at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2014. **42**(Database issue): p. D191-8.

124. Sievers, F. and D.G. Higgins, *Clustal Omega, accurate alignment of very large numbers of sequences.* Methods Mol Biol, 2014. **1079**: p. 105-16.

125. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.

126. Yang, Z., *Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.* Molecular biology and evolution, 1998. **15**(5): p. 568-573.

127. Ostrow, S.L., et al., *Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes.* PLoS Genetics, 2014. **10**(3): p. e1004239.

128. Supek, F., et al., *Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers.* Cell, 2014. **156**(6): p. 1324-1335.

129. Kimchi-Sarfaty, C., et al., *A" silent" polymorphism in the MDR1 gene changes substrate specificity.* Science, 2007. **315**(5811): p. 525-528.

130. Benjamini, Y. 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), 1995: p. 289-300.

131. Moore-Smith, L. and B. Pasche, *TGFBR1 signaling and breast cancer.* J Mammary Gland Biol Neoplasia, 2011. **16**(2): p. 89-95.

132. Ikushima, H., et al., *Autocrine TGF-β Signaling Maintains Tumorigenicity of Glioma-Initiating Cells through Sry-Related HMG-Box Factors.* Cell Stem Cell, 2009. **5**(5): p. 504-514.

133. Craig, A.L. and T.R. Hupp, *The regulation of CHK2 in human cancer.* Oncogene, 2004. **23**(52): p. 8411-8418.

134. Guo, S., et al., *Vascular endothelial growth factor receptor-2 in breast cancer.* Biochim Biophys Acta, 2010. **1806**(1): p. 108-21.

135. Baek, S.H. and K.I. Kim, *Emerging Roles of Orphan Nuclear Receptors in Cancer.* Annual Review of Physiology, 2014. **76**(1): p. 177-195.

136. Dorsam, R.T. and J.S. Gutkind, *G-protein-coupled receptors and cancer.* Nat Rev Cancer, 2007. **7**(2): p. 79-94.

137. ENCODE Project Consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

CURRICULUM VITAE

RUNJUN D. KUMAR, HBSc
June 2016

Email Address: runjun.d.kumar@gmail.com **Citizenship:** Canada

Present Position:

2010-2018 Washington University School of Medicine
Medical Scientist Training Program MD/PhD candidate
Graduate Year 4, Computational and Systems Biology
Division of Biology and Biomedical Science

Education:

2010 University of Toronto, Toronto, Ontario, Canada
Honours Bachelor of Science, Pathobiology & Statistics
with High Distinction

Research Experience:

2012-2016 Washington University in St. Louis
Doctoral Research, Computational and Systems Biology
Thesis Advisor: Ron Bose, MD, PhD

May-Aug. 2010 European Molecular Biology Laboratories, Heidelberg, Germany
Student Research Fellow
Advisor: Peer Bork, PhD

May-Aug. 2009 Harvard Medical School
Student Research Fellow
Advisor: Paul de Bakker, PhD

2008-2009 University of Toronto
Undergraduate Research Fellow
Advisor: William Navarre, PhD

Research Support:

2014-Present Canadian Institutes of Health Research
Doctoral Foreign Study Award (#DFS-134967)
(National Institutes of Health F31 equivalent)

2010-2014 Washington University in St. Louis
Medical Scientist Training Program
(Funded by National Institutes of Health)

May-Aug. 2010	European Molecular Biology Laboratories Visiting Scholar Grant
May-Aug. 2009	Heart and Stroke Foundation of Canada John. D. Schultz Scholarship
2008-2009	University of Toronto Undergraduate Research Fellowship

Teaching & Curriculum Development:

2013-2014	Washington University in St. Louis Medical Histology Teaching Assistant Responsibilities: leading laboratories, designing and delivering review sessions, grading exams. Class size: 120
2013-2015	Washington University in St. Louis Medical Scientific Methods I Course Assistant Responsibilities: redeveloping, delivering, assessing and revising new small group sessions and final exam format. Class Size: 120

Leadership

2013 – Present	Washington University in St. Louis School of Medicine Musical Producer Responsibilities: organizing auditions, rehearsals, venue and budget. Liaising with administrators and other student groups.
2010-2013	Washington University in St. Louis School of Medicine CPR Instructor Responsibilities: Organizing and leading required CPR certification for incoming pre-clinical and outgoing clinical students. Class Size: 15-50
2007-2010	University of Toronto Pathobiology Student Union President Responsibilities: Acted as student representative to the departmental and faculty curriculum renewal committees. Class Size: 150

Honors and Awards

2015	Molecular Genetics & Genomics Retreat Best Talk Award
2010	Alan Gornall Award (Pathobiology Gold Medal)
2006-2010	Canadian Millennium Foundation Merit Award
2006	Canadian National Biology Competition Award
2006	Governor General's Award

Refereed Articles

Chudnovski Y*, **Kumar RD***, Schrock AB, Gowen K, Frampton GM, Connelly C, Stephens PJ, Miller VA, Ross JS, Ali SM, Bose R. Response of a Metastatic Breast Carcinoma with a Previously Uncharacterized ERBB2 G776V Mutation to Human Epidermal Growth Factor Receptor 2–Targeted Therapy. 2017. *JCO Precision Oncology*. 1,1-9. *equal contributors

Kumar RD, Bose R. 2017. Analysis of somatic mutations across kinome reveals loss-of-function mutations in multiple cancer types. *Scientific Reports*. 7.

Kumar RD, Swamidass SJ, Bose R. 2016. Unsupervised detection of cancer driver mutations with parsimony-guided learning. *Nature Genetics*. 48(10).

Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, Bose R. 2015. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome sequencing data. Bioinformatics. 31(22).

Arking DE, Pulit SL, Crotti L, van der Harst P, Munroe PB, …**Kumar RD**, … de Bakker PIW, Newton-Cheh C. 2014. Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. Nature Genetics. 46(8).

Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, Koval J, Das I, Callaway MB, Eldred JM, Miller CA, Subramanian J, Govindan R, **Kumar RD**, Bose R, Ding L, Walker JR, Larson DE, Dooling DJ, Smith SM, Ley TJ, Mardis ER, Wilson RK. 2013. DGIdb – Mining the druggable genome. Nature Methods. 10(12).

Kumar RD, Chang L-W, Ellis MJ, and Bose R. 2013. Prioritizing potentially druggable mutations with dGene: an Annotation Tool for Cancer Genome Sequencing Data. PlosONE. 8 (6).

Minguez P, Parca L, Diella F, Mende DR, **Kumar R**, Helmer-Citterich M, Gavin AC, Van Noort V, Bork P. 2012. Deciphering a Global Network of Functionally Associated Post-Translational Modifications. 2012. Molecular Systems Biology. 8(599).

Van Noort V, Seebacher J, Bader S, Mohammed S, Vonkova I, Betts MJ, Kuhner S, **Kumar R**, Maier T, O'Flaherty M, Rybin V, Schmeisky A, Yus E, Stulke J, Serrano L, Russell RB, Heck AJR, Bork P, Gavin AC. 2012. Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. Molecular Systems Biology. 8(571).

Navarre WW, Zou SB, Roy H, Xie JL, Savchenko A, Singer A, Edvokimova E, Prost LR, **Kumar R**, Ibba M, Fang FC. 2010. PoxA, YjeK, and elongation factor P coordinately modulate virulence and drug resistance in Salmonella enterica. Molecular Cell. 39 (2).

Conference Proceedings

Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, Bose R. A panel of novel statistical tests identifies tumor suppressors and oncogenes from pan-cancer genome sequencing data. In the *28th meeting on the Biology of Genomes*. May 5-9th, 2015, Cold Spring Harbor, NY.

Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, Bose R. Improved Detection of Cancer Genes from Pan-Cancer Genome Sequencing Data. In the *11th meeting of the American Physician Scientists Association.* April 24-26th, 2015, Chicago, Il