Enhanced context reveals the scope of somatic missense mutations driving human cancers

Collin Tokheim[1,2] and Rachel Karchin[1,2,3,*]

1 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
2 Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA
3 Department of Oncology, Johns Hopkins University, Baltimore, MD 21204, USA

* corresponding author
Rachel Karchin, Ph.D.
217A Hackerman Hall
3400 N. Charles St.
Baltimore, MD USA 21204
ph: +1 410 516 5578
fax: +1 410 516 5294
karchin@jhu.edu

Keywords: cancer driver, missense mutation, rare drivers, TCGA

Abstract

Large-scale sequencing studies of patient cohorts enable identification of many cancer driver genes. However, not every mutation in a driver gene is necessarily a driver of cancer; thus requiring methods to discriminate whether an individual mutation is a driver or passenger. By completely re-working the CHASM algorithm, CHASMplus leverages multi-scale context to identify driver missense mutations, and consistently outperforms comparable methods across a wide variety of benchmarks – including *in vitro* experiments, *in vivo* experiments, and literature curation. Applied to 8,657 samples across 32 cancer types, CHASMplus identifies 3,527 unique driver mutations. Our results support a prominent emerging role for rare driver mutations. To our knowledge, this study is the first to systematically assess variability across cancer types with respect to the spectrum of common, intermediate, and rare frequency driver mutations. We show that the trajectory of driver discovery is systematically different across cancer types, depending on mutational prevalence and diversity.

Introduction

While somatic missense mutations are the most common protein-coding mutations in human cancers, their impact on cancer cell fitness is highly variable. Missense mutations may be critical to increasing the net growth of a cancer cell *in vivo* (drivers), while others are benign passengers[1]. Certain cancer types are known to be driven by well-established, highly prevalent missense mutations in oncogenes, such as KRAS G12 mutations in pancreatic ductal adenocarcinoma[2] or IDH1 R132 mutations in glioblastoma multiforme[3]. The OncoKB database[4], a literature curated knowledgebase, has compiled a list of approximately two thousand driver missense mutations. This database is likely incomplete because of the current limited throughput of functional validation assays for driver missense mutations. It has been hypothesized that cancer driver mutations exhibit a long tail phenomenon with few common drivers and many rare drivers[5,6], suggesting that numerous rare drivers remain to be discovered. Outstanding questions remain about the spectrum of driver missense mutations, their cancer type specificity, and their potential clinical relevance.

.

The task of identifying drivers from cancer sequencing studies has classically used statistical models that identify an excess number of mutations over expectation[7,8]. However, most genes are large and even within a driver gene numerous passenger mutations are expected to accumulate[9]. This leads to uncertainty on whether an individual mutation is a driver. Approaches to improve the specificity of driver discovery have focused on smaller intra-genic regions, such as protein domains[10], protein-protein interfaces[11,12], and individual codons[13].

An alternative approach is to apply machine learning to predict the cancer driver status of individual missense mutations by leveraging features characterizing the mutation, e.g., inter-species evolutionary conservation, features of the local protein environment, molecular function annotations, and biophysical characterizations of the amino acid substitution. Cancer-focused machine learning methods have previously tried to enhance performance by training cancer type specific models[14,15] or boosting data with synthetic passenger missense mutations[14]. While results have been promising, a recent systematic study comparing 15 such methods concluded that none of them were yet sufficiently reliable to guide high-cost experimental or clinical follow-through[16]. We and others have hypothesized that determining the impact of missense mutations requires proper context[17], which has not yet been systematically leveraged in the current generation of methods. Context includes both prior knowledge about the functional importance of genes or gene subregions in which a mutation occurs, and mutational patterns that are now evident from cancer sequencing studies of many thousands of patients.

In this work, we present a new driver missense mutation prediction method, CHASMplus, that uses machine learning to integrate missense mutation context at multiple scales. The new CHASMplus consistently outperforms comparable methods, including the original CHASM, on eight different benchmark sets -- including in vitro experiments, in vivo experiments and literature benchmarks. Encouraged by these results, we applied CHASMplus to 8,657 The Cancer Genome Atlas (TCGA) samples from 32 cancer types to systematically identify driver missense mutations. The resulting predictions support a high degree of variability in the contributions of rare, intermediate, and common driver missense mutations across cancer types. The prevalence and diversity of the driver mutation landscape differ by cancer type, suggesting which cancer types are saturated for driver discovery and those most likely to benefit by additional sequencing studies.

Result

Overview of CHASMplus

CHASMplus uses the Random Forest algorithm to discriminate somatic missense mutations (referred to hereafter as *missense mutations*) as either cancer drivers or passengers (Online Methods). A Random Forest is an ensemble of many randomized decision trees[18,19]. Each tree is trained on a random selection of training set examples and candidate features, via a recursive splitting process[20] (Fig 1a). CHASMplus is trained using somatic mutation calls from The Cancer Genome Atlas (TCGA) covering 8,657 samples in 32 cancer types (Table S1, Online Methods). Because there is no gold standard set of driver and passenger missense mutations, we developed a semi-supervised approach to assign class labels to missense mutations, taking advantage of Random Forest robustness to noisy class labels. Briefly, class labels are assigned so as to enrich the positive class for driver missense mutations (Online Methods, Fig S1). CHASMplus training is done with a rigorous gene holdout cross-validation protocol to avoid overfitting, by ensuring all mutations within a gene are within the same fold[21]. Therefore, missense mutations are never scored by a Random Forest trained on any missense mutation harbored by the same gene. Finally, predicted scores from CHASMplus are weighted by the 20/20+ driver gene score, producing gene-weighted (gwCHASMplus) scores (Fig 1b, Online Methods).

CHASMplus scores benefit from representation of missense mutation context at multiple scales. The Random Forest was trained on 95 features (Table S2), and the 34 most important features are shown in Fig 1c (Online Methods). Important features assess five broad categories: multi-resolution missense mutation hotspots (HotMAPS 1D algorithm[22]), evolutionary conservation/human germline variation, molecular function annotations (*e.g.*, protein-protein interface annotations from[23]), sequence biased regions, and gene-level covariates (*e.g.,* replication timing). Missense mutation context is further represented by the 20/20+ driver score of the gene harboring the missense mutation and the specific cancer type in which it was observed. While gene-level features have been previously applied to missense mutation driver prediction[24], to our knowledge, this is the first time that gene-level and missense mutation-level driver scores have been coupled in a cancer type-specific manner.

CHASMplus can also evaluate the statistical significance of cancer type-specific predictions for each of 32 cancer types from The Cancer Genome Atlas (TCGA), and *pan-cancer* predictions for all TCGA cancer types in aggregate (Fig 1d, Online Methods). Because Random Forests do not intrinsically include hypothesis testing techniques, we used simulated mutations to assess the statistical significance of scores. P-values were estimated from a simulated null distribution, controlling for sequence composition, and corrected for multiple testing with the Benjamini-Hochberg method (Online Methods). The resulting P-value distributions suggest our statistical model is well calibrated (Fig S1). Well-calibrated P-values enable quantitative estimates of false discovery rate and thus inform a user about how to select a suitable score threshold for predicted driver missense mutations.

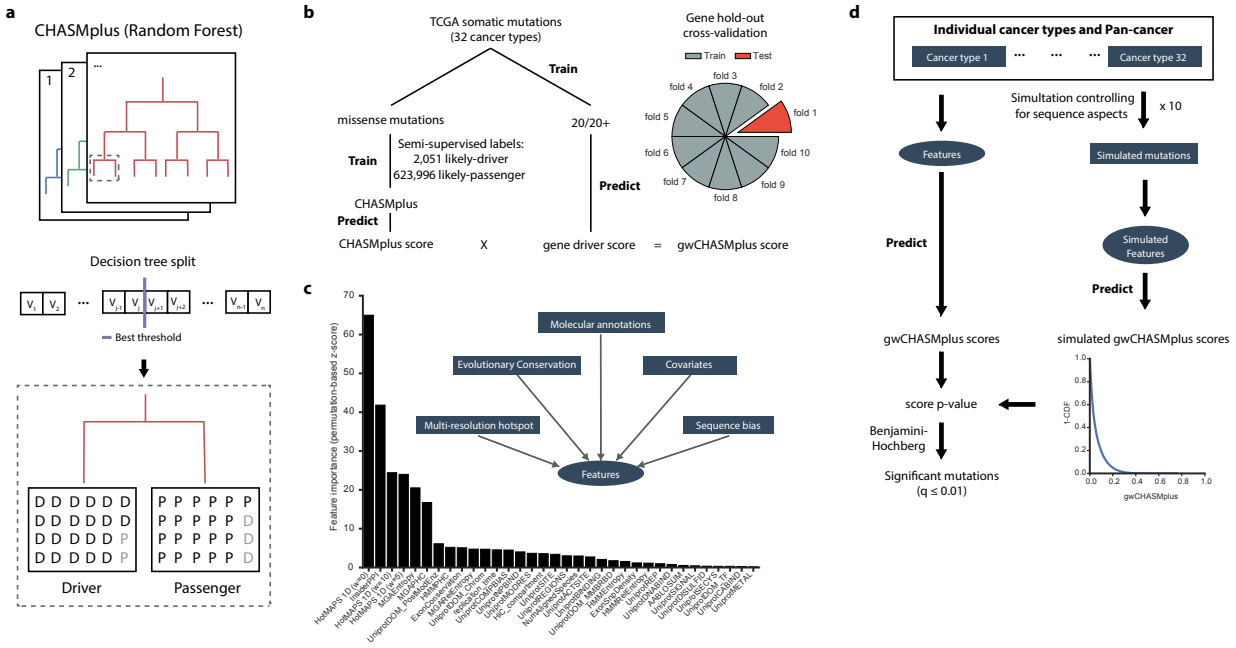

Figure 1. Overview of CHASMplus algorithm. a) CHASMplus predicts driver somatic missense mutations by using a random forest algorithm, consisting of an ensemble of decision trees. Each decision tree is constructed by selecting a random set of examples and features and recursively splitting examples by the best split criterion. **b)** Diagram of training and testing procedure by CHASMplus. **c)** Features with a net-positive feature importance by CHASMplus according to a permutation adjusted z-score. Boxed text indicates broad feature categories that were important. **d)** Diagram of how CHASMplus identifies statistically significant driver somatic missense mutations in each of the 32 cancer types individually and in aggregate (pan-cancer).

CHASMplus dramatically improves identification of somatic missense mutation drivers

We next sought to compare the performance of CHASMplus with respect to 12 comparable methods by using seven mutation-level benchmarks. Our benchmarks fall under three broad categories: *in vitro* experiments, high throughput *in vivo* screens, and curation from published literature. Each of these categories has weaknesses, but, in aggregate, they span multiple scales of evaluation and amount of supportive evidence (Fig 2a). For example, several benchmarks are limited to one or a few well-established driver genes, while others are exome-wide, but lack experimental support. A range of benchmarks is critical because missense mutations with the most established experimental support for a driver role tend to be in a few well-understood cancer driver genes. However, limiting benchmarking to these genes makes it difficult to assess the generalizability of a method's performance to missense mutations in other genes. All benchmark evaluations used the area under the Receiver Operating Characteristic Curve (auROC) as a metric (Fig 2b, Fig S2, Table S3). Overall, CHASMplus had a mean auROC of 0.09 higher than the next best method. This common metric is used in machine learning to describe how well predictions separate two classes without *a priori* selecting a score threshold, which for many methods is not well defined[25]. In our assessment, the two classes represent likely driver and passenger missense mutations. In general, auROC values range from 0.5 (random prediction performance) to 1.0 (perfect).

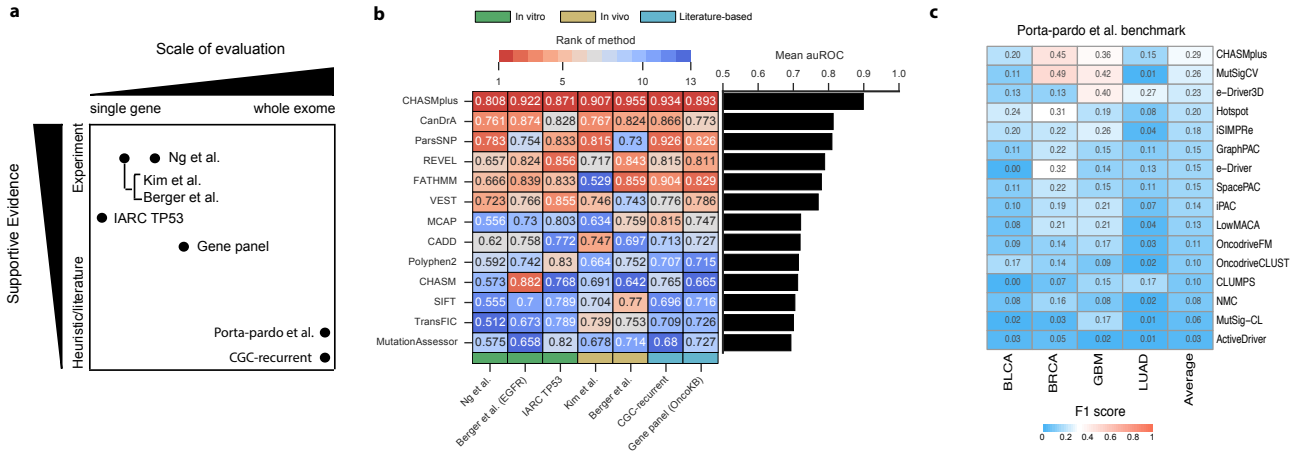

Figure 2. Cancer driver prediction benchmark. a) Conceptual diagram of how 8 benchmarks compare in terms of the scale of evaluation and amount of supportive evidence. **b)** A heatmap showing performance measured by the area under the Receiver Operating Characteristic Curve (auROC) on the 7 mutation-level benchmarks (shown in text). The color scale from red to blue indicates methods ranked from high to low performance. Benchmarks are categorized by in vitro (green), in vivo (yellow), and literature-based benchmarks (turquoise). The bar graph shows the mean auROC across the benchmarks. **c)** Heatmap showing performance (F1 score) on a cancer type specific benchmark. The overall performance on four cancer types (BLCA, BRCA, GBM, and LUAD) is measured by the average F1 score (right column) TCGA cancer type acronyms are in Table S1.

We used three benchmarks based on *in vitro* experiments (Online Methods). The first was a set of missense mutations assessed by an assay of cell viability in two growth-factor dependent cell lines, Ba/F3 and MCF10A (pro-B and breast epithelium cell lines), covering 747 mutations in 48 genes[26]. CHASMplus had significantly higher performance than the next best performing method (ParsSNP) ($p<0.05$, delong test). In the second benchmark, an *in vitro* assay of EGFR resistance to erlotinib from missense mutations observed in lung adenocarcinoma[27], CHASMplus (auROC=0.92) outperformed all other methods, with the next best method (CanDrA) having an auROC of 0.87. CHASMplus auROC was significantly better than that of 7 of the methods tested ($p<.05$, delong test). For the remaining 5 methods, the improvement was not significant, possibly due to lack of power given the small number of mutations (n=75) tested in the assay. In the third benchmark, an assay of reduced transactivation (<50% WT, median of 8 targets) in TP53 from the IARC database (n=2,314 mutations)[28], CHASMplus significantly outperformed the next best method (REVEL) ($p=0.02$, delong test).

To investigate whether CHASMplus would also perform well when compared to results of *in vivo* experiments, we considered two benchmarks based on pooled *in vivo* screens in mice that assessed mutation driver status by fitness in a competition assay. The first was performed from mutations observed in lung cancers (44 missense mutations)[27] and the second from mutations observed in 27 cancer types (71 missense mutations)[29]. CHASMplus had the highest auROC of the 13 tested methods on both benchmarks, with an increase in auROC by 0.09 and 0.1, respectively, compared to the next best methods (ParsSNP in the first benchmark and FATHMM in the second). The increase was significant in the second larger, benchmark ($p=0.03$, delong test, $n=72$), but not in the first, which may be the result of the smaller sample size. In the first benchmark, CHASMplus was significantly better than 9 out of 12 tested methods ($p<0.05$, delong test, $n=44$).

Experimental testing of mutations across large number of genes or the whole exome is currently not feasible. Therefore, evaluation of CHASMplus at larger scales relied on two benchmarks based on literature and database curation. The first benchmark in this category labeled recurrent missense mutations within genes in the Cancer Gene Census[30] as drivers (Online Methods). We found that the gene weighted CHASMplus scores (auROC=0.934) were substantially better at this whole exome-wide prioritization task compared to the unweighted CHASMplus scores (auROC=0.893) ($p<2.2e-16$, delong test). CHASMplus scores were also significantly better than the next best method (ParsSNP) ($p=0.001$, delong test). The second benchmark was derived from a large driver gene panel (MSK-IMPACT) (414 genes) and 10,130 sequenced cancer patients. Missense mutations were labeled as drivers if they were annotated as such in OncoKB, a knowledge-base that aggregates known literature (Online Methods). CHASMplus significantly outperformed all other methods, the nearest being ParsSNP ($p=7e-14$, delong test).

CHASMplus improves identification of cancer type specific driver genes

We evaluated the performance of CHASMplus on identifying cancer-type specific driver genes, using a previously published benchmark and assessment of 15 computational methods designed for this purpose[31]. Genes were labeled by their designations in the Cancer Gene Census as a cancer driver gene for a specific cancer type. Out of the 4 cancer type cohorts assessed (BLCA, BRCA, GBM, and LUAD), CHASMplus had the highest average F1 score, a balance between precision and recall that was used as a performance metric by[31] (Fig 2c). We

additionally note that of the methods tested, CHASMplus was the only one not primarily designed to predict driver genes that had high recall (average recall=.45) while maintaining precision (average precision=.23) (Fig S3).

CHASMplus identifies both common and rare cancer drivers

Certain cancer driver mutations primarily occur in a specific cancer type, while others appear in many cancer types. The power to detect driver mutations, which occur at low frequency in many cancer types, is increased when many cancer types are aggregated, known as a pan-cancer analysis. Conversely, driver mutations, which are specific to a particular cancer type, are best identified when cancer types are analyzed individually[32]. Using CHASMplus, we identified 3,527 unique missense mutations as statistically significant drivers by pan-cancer analysis at an estimated false discovery rate of 1% (Table S4). When applied to each cancer type individually, the number found significant varied substantially from 8 in thymoma to 572 in bladder urothelial carcinoma with a median of 78 (Fig 3a, Table S5). The median overlap with literature-based oncogenicity annotation from OncoKB was 53%, suggesting 47% of the driver missense mutations identified by CHASMplus either have not been previously characterized or not sufficiently characterized for inclusion in OncoKB. While OncoKB missense mutation annotations are not cancer-type specific, the genes with highest frequencies of cancer-type specific driver missense mutations identified by CHASMplus have well-known roles in cancer[33] (Fig 3b).

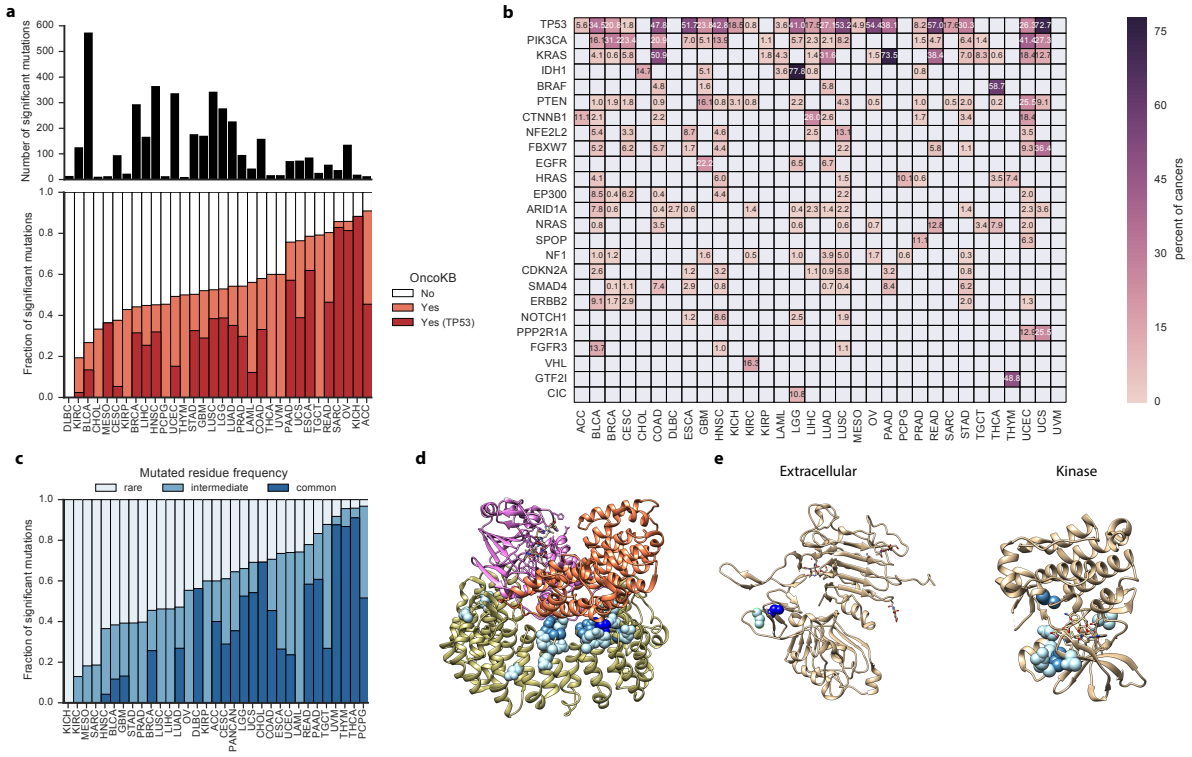

Figure 3. Discovery of driver somatic missense mutations. a) Bar graphs showing the number of unique driver somatic missense mutations (top) and the proportion previously known in OncoKB, a literature curated database (bottom). **b)** Heatmap of the top 25 genes containing the most frequent driver somatic missense mutations in TCGA across the cancer type specific analyses. Shown are the percentage of samples that are mutated. **c)** Proportion of overall frequency of driver somatic missense mutations found in rare (<1% of samples or singleton mutations), intermediate (1-5%), and common (>5%) driver somatic missense mutations. Correspondingly shown as light to dark blue. **d)** Structure of the Phosphatase 2A holoenzyme (PDB 2IAE). **e)** Structures of the ERBB2 extracellular domain (left, PDB 2A91) and kinase domain (right, PDB 3PP0). TCGA cancer type acronyms are in Table S1.

The long tail hypothesis, initially proposed from examining overall mutation frequency of driver genes[5,6], suggests there are few common drivers and many rare drivers. However, the overall mutation frequency of a gene does not account for the confounding presence of passenger mutations within a driver gene. From our mutation-level analysis, we observed that the spectrum of rare (<1% of cancer samples), intermediate (1-5%), and common (>5%) driver missense mutations varied substantially among cancer types (Fig 3c). For example, uveal melanoma was dominated by common driver missense mutations (88%), while head and neck squamous cell carcinoma (HNSC) was dominated by rare driver missense mutations (63%). Interestingly, from the pan-cancer analysis, the overall proportion of driver missense mutations considered rare was only slightly smaller than for common drivers (35.4% and 35.5%, respectively), but 4-fold greater than found by a previous method (8%, P<2.2e-16, Fisher's exact test)[13].

Rare driver missense mutations exist not only in rare driver genes, but also may be spatially proximal in protein structure to common driver missense mutations. For example, the protein phosphatase PPP2R1A, which has been implicated as a tumor suppressor gene in many tumor types[34], contained common driver missense mutations in our pan-cancer analysis at residue positions 179 and 183, which is located at the protein interface composing the phosphatase 2A complex (Fig 3d). It also had a much broader set of rare drivers throughout the protein interface, such as R105Q and R459C. Similarly, CHASMplus identified common driver missense mutations (S310A/F/Y) in the extracellular domain of the well-known oncogene ERBB2, but also finds rare driver missense mutations in both the extracellular and kinase domain (e.g., L313V and R678Q) (Fig 3e). This is supportive of previous experimental work implicating rare cancer driver mutations in common cancer driver genes[29].

Mutation hotspot detection approaches have limited power

A codon or small region of protein sequence or structure where recurrent mutations are observed is known as a hotspot. Similar to statistical methods for driver gene detection, hotspot detection identifies an excess number of mutations compared to expectation using a large number of cancer samples. We asked whether, given current cohort sizes, codon-based hotspot detection had sufficient statistical power to identify rare driver mutations. We assessed the number of samples required to detect driver mutations across a range of frequencies (proportion of samples in which a mutation occurs) and somatic background mutation rates. In Fig 4a, each of the 32 TCGA cancer types is placed according to its sample size and background mutation rate, relative to six curves which represent the required sample size to detect driver mutations of a certain frequency, with 90% power, using hotspot detection (Online Methods). For example, the TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) cohort has 274 samples and a background mutation rate of 3.5 mutations/Mb. This sample size is sufficient to detect driver mutations that occur in ~2% of the samples with 90% power.

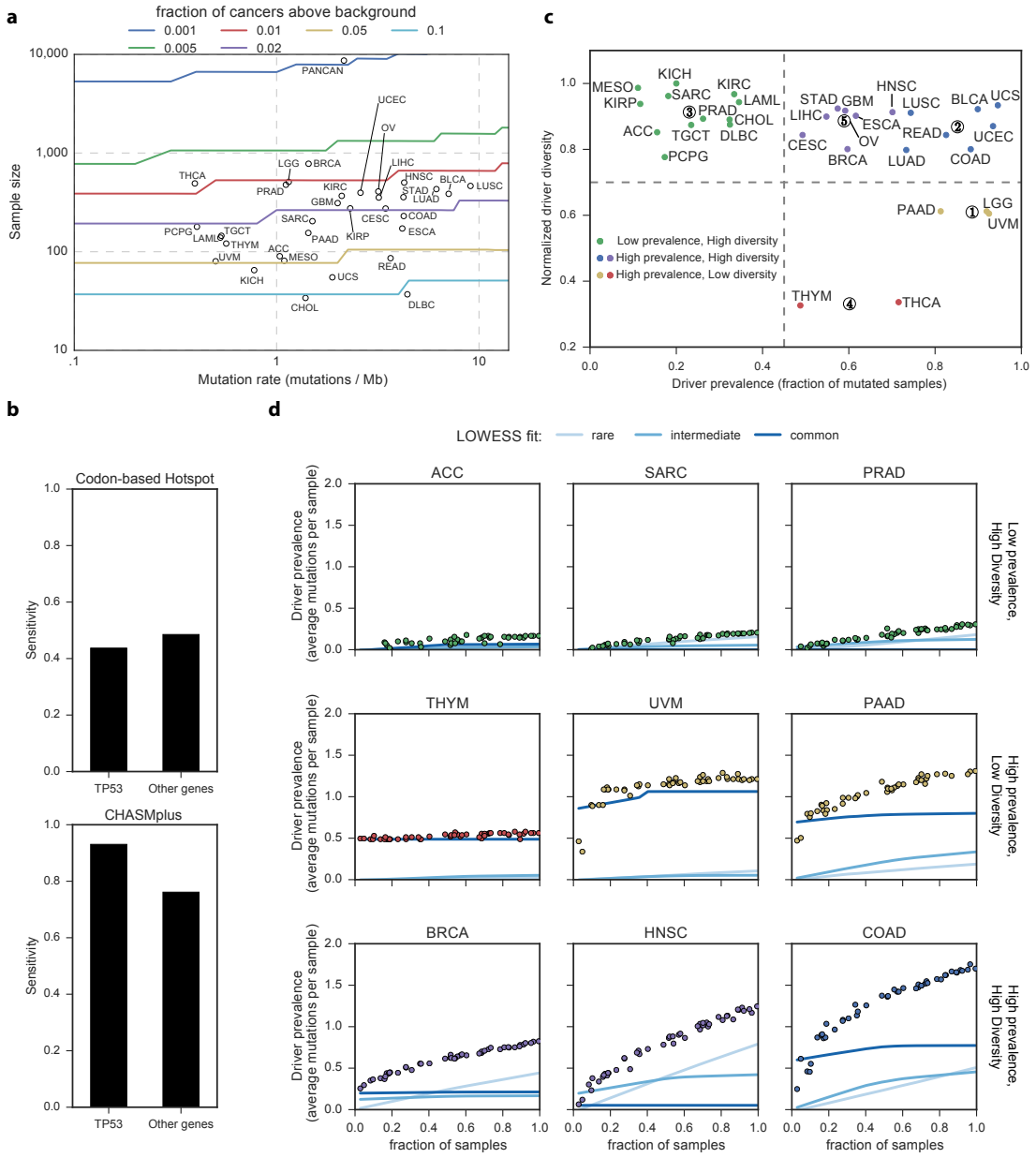

Figure 4. Saturation and characteristics of driver somatic missense mutations. a) Statistical power to detect significantly elevated non-silent mutations for individual codons as a function of sample size and mutation rate. Circles represent each cancer type from the TCGA and is placed according to sample size and median mutation rate. Curves are colored by the effect size of driver mutations (fraction of non-silent mutated cancer samples above the expected background mutation rate). **b)** Bar graph comparing sensitivity to detect labeled oncogenic driver missense mutations from OncoKB between CHASMplus and a hotspot detection approach. **c)** Plot displaying normalized driver diversity and driver prevalence (fraction of samples mutated) for driver somatic missense mutations in 32 cancer types. K-means clustering identified 5 clusters with centroids shown as numerically designated circles. **d)** Prevalence of driver somatic missense mutations as a function of sample size. Lines represent LOWESS fit to different rarities of driver somatic missense mutations. TCGA cancer type acronyms are in Table S1.

At current TCGA sample sizes, we found codon-based hotspot detection approaches were not well powered to identify driver mutations that occurred at less than 1% frequency in most cancer types. Exceptions were thyroid carcinoma (THCA), low grade glioma (LGG) and breast cancer (BRCA), which are seen to lie above (or close to) the curve representing 1% frequency (Fig 4a). Notably, these cohorts had large numbers of samples and low-to-medium background mutation rates. We also found that when cancer types were aggregated in pan-cancer analysis, power to detect codon-based hotspots improved substantially, but only when the recurrent mutations were shared in more than one cancer type. For these mutations, pan-cancer analysis using ~10,000 TCGA samples should enable detection of driver mutations at frequency as low as 0.1%.

In our pan-cancer analysis, CHASMplus had greater sensitivity to detect putatively oncogenic missense mutations than a recently published codon-based hotspot detection method. We compared the missense mutations in the TCGA pan-cancer cohort that were called statistically significant by CHASMplus and those called by a hotspot method described by[13] ($q<=0.01$). For both methods, we computed the overlap with well-curated oncogenic mutations in the OncoKB database. The sensitivity of CHASMplus to detect the OncoKB-labeled mutations was 0.83, which was significantly higher than the hotspot method (0.46, $p<2.2e-16$, McNemar's test, $n=896$). To minimize gene bias, we also repeated the analysis after excluding all 389 TP53 mutations, yielding sensitivity of 0.76 for CHASMplus and 0.49 for hotspot detection, a difference which is still statistically significant ($p<2.2e-16$, McNemar's test, $n=507$) (Fig 4b). Moreover, these results are also reflected in the number of significant predictions of the two methods. The codon-based hotspot method only identified 360 unique codons as significant in our TCGA data set, while CHASMplus found significant missense mutations in 2,588 codons. We believe that the increased sensitivity is the result of CHASMplus using a broad range of important features, including multi-resolution hotspot detection and weighting by driver gene scores (Fig 1c). Importantly, our increased sensitivity did not come at the cost of low specificity, as evidenced by our p-value calibration (Fig S1c) and extensive ROC analysis across seven benchmarked datasets (Fig 2b), which measures a balance of sensitivity and specificity.

Characterizing cancer types and the trajectory of discovery

The diversity and prevalence of driver missense mutations varied considerably across TCGA cancer types (Fig 4c, Online Methods). We defined diversity with respect to the distribution of

driver missense mutations across codons and prevalence with respect to the frequency of the mutations in tumor samples. High diversity indicated mutations were broadly distributed across codons, while high prevalence indicated driver missense mutations that occurred in a large number of tumor samples. Using K-means clustering, we found that cancer types grouped into high diversity and low prevalence (12 cancer types), high diversity and high prevalence (15 cancer types), and low diversity and high prevalence (5 cancer types). These differences were not associated with intra-tumor heterogeneity or normal contamination, as assessed by mean variant allele fraction (VAF) of a cancer type ($p>0.05$, correlation test). The differences also could not be associated only with TCGA sample size for a particular cancer type. For example, while both pancreatic ductal adenocarcinoma (PAAD) and sarcoma (SARC) had similar sample sizes (n=155, n=204 respectively), PAAD had high prevalence and low diversity, while SARC had low prevalence and high diversity. After adjusting for sample size, we observed that the average mutation burden for a cancer type positively correlated with the prevalence of rare (but not common) driver missense mutations (R=0.63, P=4.7e-5, likelihood ratio test).

Are there substantially more cancer driver missense mutations yet to be discovered? If discovery was measured by the number of unique driver missense mutations identified, subsampling analysis showed all cancer types had a linear increase ($R^2 > 0.5$) with no evidence of saturation at current sample sizes (Fig S5). However, we observed substantial variability in trajectories if discovery was measured by driver prevalence (average number of driver missense mutations per cancer sample) (Fig 4d), a metric which goes directly to utility of driver discovery in clinical practice (Discussion). For sarcoma (SARC), adrenocortical carcinoma (ACC), and prostate adenocarcinoma (PRAD), driver prevalence remained minimal as sample size increased. As a case in point, we repeated our analysis on data from a recently released PRAD study [35], which augmented the 477 TCGA PRAD samples with 536 additional samples. This resulted in only marginal increases in the overall prevalence of identified driver missense mutations, consistent with our predicted trajectory based only on TCGA samples (Supplementary Materials and Methods, Table S6, Fig S6). In contrast, thymoma (THYM), uveal melanoma (UVM), and pancreatic ductal adenocarcinoma (PAAD) contained common driver missense mutations that could be detected by using only a few samples from the cohort, *e.g.,* GTF2I L424H in THYM. Due to a lack of rare or intermediate driver missense mutations, we observed THYM and UVM saturated discovery as sample size increased. Although PAAD showed a growing set of intermediate/rare driver missense mutations, the overall driver prevalence exhibited a diminishing rate of discovery. In contrast, breast (BRCA), head and neck

squamous (HNSC), and colon cancers (COAD) harbored a full spectrum of driver missense mutations, with rare drivers increasing substantially as a function of sample size.

Discussion

Cancer genome interpretation is challenged by the reality that of all somatic mutations observed in cancer, only a small proportion are drivers[36]. Future insights into cancer evolution and its relevance for clinical care will increasingly rely on the precise interpretation of whether individual mutations are cancer drivers[37]. While many computational methods have been developed for this purpose, evaluations by external groups suggest that improvements are needed[16,38]. In this work, we used a large dataset (~9,000 cancer samples) from The Cancer Genome Atlas (TCGA) to train a new machine learning method called CHASMplus. CHASMplus was designed to better represent the context in which missense mutations occur by coupling prior information about a mutation's likely functional importance with mutational patterns evident from large cancer sequencing studies. We compared CHASMplus to 27 other computational methods, including the original CHASM, on eight benchmarks covering *in vivo* experiments, *in vitro* experiments, and literature curation. CHASMplus had the best performance at predicting drivers at each scale of evaluation – a whole exome, a targeted gene panel, and within a single gene. Although not perfect, we believe the application of multiple independent benchmarks spanning a wide array of genes is the current best practice for demonstrating effectiveness.

The long tail hypothesis[5,6] posits that there are many rare driver mutations in human cancers. Insight into rare driver mutations has been limited by the lack of statistical power from traditional hotspot detection approaches. To overcome this limitation, we leveraged the improvements made in CHASMplus to systematically predict driver missense mutations in 8,657 samples from the TCGA. Although individually rare, we found a prominent emerging role for rare driver missense mutations in aggregate, consistent with the long tail hypothesis. This result supports the critical role of understanding the prevalence of driver mutations -- failure to capture and identify rare driver mutations, which occur in aggregate at reasonable prevalence, may result in crucial missed opportunities for interpreting a patient's cancer. Because high-throughput functional validation studies of missense mutations are not yet widespread, computational methods, like CHASMplus, are needed to prioritize mutations for low- and medium-throughput studies. A benefit of CHASMplus is that we have precomputed a score for every possible missense mutation, forming an *in silico* saturation mutagenesis across all genes to score as of

yet unobserved mutations that are potential cancer drivers. We provide scores for each of the 32 TCGA cancer types, as well as pan-cancer scores (available on request from the authors).

To our knowledge, ours is the first study to systematically show that the prevalence and diversity of driver missense mutations is highly variable across the cancer types represented in the TCGA. We observed that mutation burden for a cancer type positively correlated with prevalence of rare (but not common) driver missense mutations, even after correcting for sample size, suggesting that accumulating a greater number of mutations in a cancer may increase the competitiveness of rare drivers. More research into the origins of rare driver mutations is warranted, because differences in the rarity of driver missense mutations could arise from a variety of factors, including the driver's strength, dependence on genetic or environmental factors, competition from other types of tumor-derived alterations, or role in cancer subtypes.

We expect that an increasingly complete catalog of driver missense mutations will be generated by a combination improved computational methods and cumulative growth of available samples from cancer sequencing studies. The multi-faceted features used in CHASMplus yield an improvement in statistical power to effectively identify these mutations. However, for some cancer types, discovery may already be effectively saturated. We observed that the rate of new driver discoveries with greater sample size may decay because of the rarity of newly identified driver missense mutations; indicating the trajectory of driver discovery is more complicated than previously envisioned by an analysis of driver genes[35,39]. The distinction of predicting drivers at the mutation-level is important, otherwise estimates will increasingly be inflated by the relatively greater proportion of passenger mutations within rare driver genes. Future work will further elucidate a broader range of driver mutations, including those within non-coding regions of the genome, and at different stages of carcinogenesis, such as in pre-cancerous lesions.

Online Methods

TCGA Mutation dataset

We collected a set of 1,225,917 somatic mutations in 8,657 samples from The Cancer Genome Atlas (TCGA) somatic mutation calls (v0.2.8, https://synapse.org/MC3)[40]. We analyzed 32 cancer types with abbreviations for the cancer types are listed in Table S1. We further filtered mutations by restricting to only mutations with an annotated 'PASS' filter, except for OV and LAML where mutations with only whole genome amplified ('wga') status was allowed because otherwise the majority of samples were filtered. We additionally removed hypermutated samples, as they tend to have an adverse effect on statistical power. We identified hypermutated samples as having more mutations than 1.5 times the interquartile range above the third quartile (Tukey's condition) of samples within the same cancer type. Because some relatively low mutation rate cancer types contained outliers, we additionally required the sample to have at least 1,000 mutations to be considered hypermutated.

CHASMplus

The code for CHASMplus is available on github (https://github.com/KarchinLab/CHASMplus).

Features

CHASMplus uses features spanning somatic mutation hotspot detection, evolutionary conservation, genetic variation, molecular features, sequence-based features, amino acid substitution scores, and other covariates (Table S2). Additional new features include features 1-10. Original features used in CHASM were obtained from an updated SNVBox MySQL database (features 11-95)[41].

Training Set

Using the TCGA mutation dataset, we established training labels with a semi-supervised approach, designed to minimize bias. The positive class (likely-driver missense mutations) was selected by the following criteria: 1) missense mutations had to occur in a curated set of 125 pan-cancer driver genes[42]; 2) for each of the 32 TCGA cancer types, missense mutations found

in that cancer type had to occur in a significantly mutated gene for that cancer type according to MutSigCV v1.4[39]. We ran MutSigCV using recommended settings and a full sequencing coverage file (http://archive.broadinstitute.org/cancer/cga/mutsig). Importantly, MutSigCV v1.4 only assess the total number of mutations in a gene, and not any characteristics of those mutations; thus, we avoid making strong assumptions about the properties of a particular driver mutation; 3) missense mutations had to occur in samples with relatively low mutation rate (less than 500 mutations, half the minimum hypermutator threshold as defined above). This filter was intended to limit the number of passenger mutations mislabeled as drivers. The negative class (likely-passenger missense mutations) consisted of the remaining missense mutations in the TCGA mutation set. For training purposes, we only used unique mutations to avoid double counting a mutation seen more than once. If, however, the same mutation consequence observed in different cancer types had contradictory labels, we regarded the mutation as a driver because mutation recurrence is often cited as supportive evidence for a cancer driver role. This established a set of 2,051 likely-driver missense mutations and 623,996 likely-passenger missense mutations, for which we found sufficient annotation to compute our selected features.

20/20+ driver gene prediction

Briefly, the driver gene predictions by 20/20+ (v1.2.0, https://github.com/KarchinLab/2020plus) were carried out as previously described[43]. We used all somatic mutations from the TCGA data to train a pan-cancer model. Driver gene scores for each cancer type or pan-cancer were then computed based on predictions from the trained model. The driver gene score represents the fraction of decision trees predicting driver for a particular gene in the random forest.

Random forest algorithm

We used random forests[18,19], a machine learning technique, to predict whether a missense mutation is a cancer driver. We trained a random forest using the *randomForest* R package. To ameliorate the problematic imbalance in the training set, we used a stratified down sampling approach within the bagging procedure of the random forest. Random undersampling has been previously recommended for random forests based on empirical performance[44]. The imbalance occurred on two levels, there were substantially more labeled passenger missense mutations than drivers, and among drivers it was concentrated in a few genes. We first calculated the median number of labeled driver missense mutations within genes containing at least one driver

missense mutation label. If a gene contained more labeled driver missense mutations than the median, we set the number of driver missense mutations sampled from that gene to the median. Passenger missense mutations were sampled at an equal frequency as driver missense mutations after the gene-based median correction.

Since missense mutations in the same gene may have overlapping feature representations which result in classifier overfitting[21], we performed prediction using a 10-fold gene hold-out cross-validation procedure for both CHASMplus and 20/20+. This involved creating 10 random folds for cross-validation but ensuring all mutations within a gene are within a single fold. The CHASMplus score represents the fraction of decision trees which vote for the mutation being a driver. We calculate the gene-weighted CHASMplus score (gwCHASMplus) by multiplying the random forest score of CHASMplus by the driver gene score from 20/20+.

Estimation of statistical significance

For each gene, the somatic mutation simulation procedure as previously reported[43] was repeated 10 times, and for each simulation all features were computed (probabilistic2020 python package, v1.2.0). Because the MSK-IMPACT gene panel did not contain silent mutations, we likewise dropped all simulated mutations resulting in a silent mutation for the MSK-IMPACT simulations. Next, each simulated missense mutation and gene was scored with the CHASMplus and 20/20+ models that were previously trained on the real data. The resulting CHASMplus and gwCHASMplus scores for all simulations were used as an empirical null distribution. To compute a P value for a score, we used the fraction of simulated mutations with a score equal to or greater than the actual score. P values were adjusted by the Benjamini–Hochberg method for multiple hypotheses. We considered a missense mutation to be significant at a q-value threshold of 0.01.

Feature importance

We used the Mean Decrease in Gini Index (MDGI) as a measure of feature importance in the random forest. This measurement, however, has been previously noted to favor continuous features over discrete features with a small number of possible values[45]. We compensated for this phenomenon by calculating an adjusted z-score using a permutation-based approach. This involved calculating MDGI for each feature for 1,000 permutations and calculating the z-score of the observed data by using the mean and standard deviation of the permutations. The

permutations were carried out as follows: First, we randomly permuted the list of unique genes containing any missense mutation. Second, we assigned the first gene in the permuted list with the same fraction of driver/passenger mutations as our first gene containing labeled driver mutations in our training set. We then proceeded to the next gene in the list, and repeated the procedure, until the same number of labeled driver mutations as in our actual training set was reached. All other genes had their mutations labeled as passengers. Finally, we computed the MDGI for each feature based on the CHASMplus model on the permuted training data. Grouping by gene in the permutation was done to mimic the heuristic on how training data was labeled, and to avoid gene-level features from having artificially high feature importance.

Compared methods

We compared CHASMplus to 12 other methods at prioritizing likely cancer driver missense mutations (VEST[46], CADD[47], FATHMM cancer[48], SIFT[49], MutationAssessor[50], REVEL[51], MCAP[52], ParsSNP[24], CHASM[14], Polyphen2[53], transFIC[54] and CanDrA[15]). Scores were obtained by means made available by each of the methods. We used ANNOVAR to obtain scores for 7 of the methods (VEST, CADD, SIFT, MutationAssessor, REVEL, MCAP, and Polyphen2) from dbNSFP using the ljbb26_all annotation, except for REVEL and MCAP, which we used the revel and mcap annotations, respectively. TransFIC was obtained (http://bbglab.irbbarcelona.org/transfic/home) and ran locally using the scores from SIFT as input. Two versions of CanDrA were tested as the preferred version was not clear (http://bioinformatics.mdanderson.org/main/CanDrA), version 1.0 and version plus. With version plus, we used the “cancer-in-general” scores, but this was not available for version 1.0, so instead we used the ovarian scores. CHASM was executed using a 10-fold gene-holdout cross-validation procedure also using an ovarian passenger distribution. We executed ParsSNP using the provided pre-computed model where the input score annotations were obtained from ANNOVAR. FATHMM cancer scores were obtained directly from the available website (http://fathmm.biocompute.org.uk/). Inputs to each of the methods were prepared using custom python scripts.

CHASMplus was also compared to a codon-based hotspot method (v0.6)[13], with respect to its ability to identify cancer type-specific driver genes, sensitivity at discovering oncogenic codons, and its calibration of p-values. The hotspot method was run using default parameters on the TCGA mutation dataset. For each gene, we used the biomart R package to measure its protein

length. To assess p-value calibration, we collected all p-values for all codons, except for mutated codons in genes found in the Cancer Gene Census[30]. The code was obtained from github (https://github.com/taylor-lab/hotspots).

Driver mutation benchmarks

In each benchmark, we define a 'positive' (more driver-like) and 'negative' (more passenger-like) class for mutations to evaluate discriminating performance. We define the annotation of class and mutation data used for each benchmark below. Only missense mutations were used for each of the benchmarks. For reproducibility, all data, results, and analysis code are available on github (https://github.com/KarchinLab/Tokheim_biorxiv_2018).

CGC-recurrent

We examined driver prioritization on an exome-scale for our TCGA mutation dataset (see above) through a combined literature/heuristic evaluation. We first obtained a set of curated likely driver genes from the Cancer Gene Census (CGC, COSMIC v79) [33]. We restricted to only CGC genes that were labeled as somatic and marked as relevant for missense mutations. We labeled all recurrent missense mutations (n>1) in the CGC genes as the positive class, and remaining mutations as the negative class[33].

MSK impact gene panel and OncoKB

We obtained all missense mutations from the MSK-impact gene panel of 414 cancer-related genes[55]. Mutations were annotated against OncoKB (downloaded 4/3/2017), if the oncogenicity annotation was available for an individual mutation[4]. We regarded any mutation labeled as 'Oncogenic' or 'Likely Oncogenic' as the positive class for evaluation with remaining mutations considered as negative.

Pooled *in vivo* screen in mice

A previous study by Kim et al [29] used a competitive screen of mutations in mice to assess the oncogenicity of mutations. The study selected based on their presence in sequenced human tumors. Mutations were then transduced into HA1E-M cells, and pools of cells with different mutations were then injected into mice and then later assessed for representation of the mutation. 71 promising alleles were then subsequently validated, individually, from the screen in

NCR-Nu mice. We directly used the annotation of 'functional' (positive class) and 'neutral' (negative class) from the authors in terms of the individually validated alleles.

Multiplexed xenograft tumorigenesis assay and In vitro EGFR resistance

Berger *et al* tested a subset of lung adenocarcinoma somatic mutations suspected as likely cancer drivers. We regarded a missense mutation as 'negative' for benchmarking if they were labeled 'neutral' by the expression-based method (eVIP) and did not appear as a hit in functional assays. We benchmarked these neutral mutations against two functional assays, an *in vitro* EGFR resistance and a xenograft tumorigenesis assay. The former is an erlotinib-rescue assay using PC9 cells treated at two erlotinib concentrations (300 nM and 3 μM), which we required resistance at both concentrations to be labeled a positive. The multiplexed xenograft tumorigenesis assay (TumorPlex) used pooled barcoded alleles to assess allele tumor formation capability by comparing barcode representation to pre-injection levels. We used the author defined threshold of a TumorPlex hit to label a mutation as a positive for benchmarking purposes.

TP53 transactivation from the IARC TP53 database

We assessed each methods ability to distinguish TP53 mutations with low transactivation (positive class) versus all other TP53 mutations (negative class). We evaluated all missense mutations (n=2,314) for TP53 from the IARC TP53 database[56]. Low transactivation was considered as less than 50% wildtype, as indicated by the median of 8 different targets (WAF1, MDM2, BAX, h1433s, AIP1, GADD45, NOXA, and P53R2).

Cell viability *in vitro* assay

We evaluated missense mutations (n=747) from a prior medium-throughput *in vitro* experiment on two growth-factor dependent cell lines, Ba/F3 and MCF10A[26]. We assessed each method's ability to distinguish mutations resulting in increased cell viability (labeled 'activating'; positive class) versus those that did not (labeled 'neutral'; negative class). The experiment assumes that mutations that provide a growth advantage to cells with growth factors withdrawn reflect cancer drivers. The study considered mutations as validated if the cell viability with the mutation was higher than wild type in either cell line (2 negative controls, 3 positive controls, and wild type).

Cancer type-specific benchmark

We evaluated how well methods could distinguish cancer type-specific driver genes based on a previously established benchmark[31]. Since the cancer type-specific benchmark measures performance by gene, we indicated a gene as a cancer driver if any missense mutation was found significant by CHASMplus. CHASMplus was executed on the same mutation data used for the other methods in the benchmark[13]. Class definitions were used as is from the authors. Evaluation of performance was carried out by modifying the benchmarking scripts and data provided by the author to include CHASMplus (https://github.com/eduardporta/sub-gene_resolution).

Calculation of driver mutation frequency

Mutation frequency was calculated based on the fraction of cancer samples that contain a driver mutation in a particular cancer type. Estimates for pan-cancer analysis, which analyzes 32 cancer types, are based on the maximum frequency observed over each of the cancer types individually. The mutation frequency calculation uses the sum of driver mutations observed within the same codon, because the American College of Medical Genetics and Genomics (ACMG) guideline indicates other pathogenic mutations in the same codon provides moderate support for the pathogenicity of a mutation[57]. All mutations within a codon are then classified as rare (<1% of cancer samples), intermediate (1-5%), or common (>5%). As a result of certain cancer types having a low total number of cancer samples, we also regarded singleton mutations as rare.

Statistical power analysis

We estimated the statistical power to find frequently mutated codons within driver genes by using a binomial model, as done previously[43]. In contrast with gene-level estimates, the length in the model represents a codon (L_c=3 bases) and not a typical coding DNA sequence length (L_g=1,500 bases). Consistent with previous gene-level power analysis[39], we use a Bonferroni family wise error rate threshold of 0.1 for statistical significance of testing all codons within putative driver genes. This threshold depends on the number of designated driver genes. We assume 206 genes (n=206), for this purpose, as this is the number of significant genes found by 20/20+ in the pan-cancer analysis (q-value <= 0.1). The alpha-level (α) to establish statistical significance is than as follows,

$$\alpha = \frac{0.1}{n * L_g / L_c}$$

Other parameters used to estimate statistical power were the same as done previously [43].

Clustering of cancer types

We clustered TCGA cancer types according to two features, prevalence (fraction of mutated samples) and normalized diversity (normalized entropy) among predicted missense mutation drivers (q <= 0.01). The normalized entropy score was calculated based on the codon-level, as follows,

$$E = \frac{-\sum_{i=1}^{k} p(i) \log_2 p(i)}{\log_2 k}$$

where there are k codons containing significant mutations, and the fraction of significant mutations in the i'th codon is p(i). We performed clustering using the k-means algorithm (scikit learn v0.18.0) where k, the number of clusters, was selected by the maximum silhouette score (k=5; varied between 2 and 10). Each parameterization was run ten times with different initial conditions to avoid local optimums by choosing the best run, defined as the lowest sum of distances to the closest centroid.

Subsampling procedure

We performed driver missense mutation predictions on random subsamples of each of 9 representative cancer types (ACC, SARC, PRAD, THYM, UVM, PAAD, BRCA, HNSC, and COAD), using CHASMplus. Subsampling was performed by randomly selecting a certain fraction of cancer samples without replacement. The designated fraction of samples for each iteration was randomly selected from a uniform distribution bounded between 0 and 1. We then ran CHASMplus using a 10-fold gene-holdout cross-validation model previously trained on the TCGA pan-cancer data. The number of unique driver missense mutations and overall driver prevalence (average number of driver missense mutations per cancer sample) were then calculated based on significant CHASMplus predictions (q<=0.01). The prevalence of a mutated residue within a particular sub-sampled result was measured against the full cohort.

Acknowledgements

Funding: National Cancer Institute (NCI) Grant F31CA200266 (to C.T.); NCI Grant U24CA204817 (to R.K.)

Author Contributions

CT and RK conceived of the study. CT and RK drafted and edited the manuscript. C.T. developed the CHASMplus algorithm and analyzed results.

Competing Financial Interests

We have no competing financial interests to disclose.

References

1. Torkamani, A. & Schork, N.J. Prediction of cancer driver mutations in protein kinases. *Cancer Res* **68**, 1675-82 (2008).
2. Biankin, A.V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399-405 (2012).
3. Parsons, D.W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-12 (2008).
4. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**(2017).
5. Ding, L., Wendl, M.C., Koboldt, D.C. & Mardis, E.R. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum Mol Genet* **19**, R188-96 (2010).
6. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17-37 (2013).
7. Dees, N.D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-98 (2012).
8. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
9. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*.
10. Yang, F. *et al.* Protein domain-level landscape of cancer-type-specific somatic mutations. *PLoS Comput Biol* **11**, e1004147 (2015).
11. Engin, H.B., Kreisberg, J.F. & Carter, H. Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PLoS One* **11**, e0152929 (2016).
12. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. & Godzik, A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput Biol* **11**, e1004518 (2015).
13. Chang, M.T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* **34**, 155-63 (2016).
14. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660-7 (2009).
15. Mao, Y. *et al.* CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS One* **8**, e77945 (2013).
16. Martelotto, L.G. *et al.* Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol* **15**, 484 (2014).
17. Raphael, B.J., Dobson, J.R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* **6**, 5 (2014).
18. Amit, Y. & Geman, D. Shape Quantization and Recognition with Randomized Trees. *Neural Computation* **9**, 1545-1588 (1997).
19. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5-32 (2001).
20. Breiman, L. Classification and regression trees. (1984).
21. Capriotti, E. & Altman, R.B. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics* **98**, 310-7 (2011).

22. Tokheim, C. *et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res* **76**, 3719-31 (2016).
23. Meyer, M.J. *et al.* Interactome INSIDER: a structural interactome browser for genomic studies. *Nature Methods* (2018).
24. Kumar, R.D., Swamidass, S.J. & Bose, R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. *Nat Genet* **48**, 1288-94 (2016).
25. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30**, 1145-1159 (1997).
26. Ng, P.K. *et al.* Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* **33**, 450-462.e10 (2018).
27. Berger, A.H. *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* **30**, 214-228 (2016).
28. 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* **28**, 622-9 (2007).
29. Kim, E. *et al.* Systematic Functional Interrogation of Rare Cancer Variants Identifies Oncogenic Alleles. *Cancer Discov* **6**, 714-26 (2016).
30. Futreal, P.A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).
31. Porta-Pardo, E. *et al.* Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat Methods* **14**, 782-788 (2017).
32. Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-20 (2013).
33. Forbes, S.A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**, D777-D783 (2017).
34. Jeong, A.L. *et al.* Patient derived mutation W257G of PPP2R1A enhances cancer cell migration through SRC-JNK-c-Jun pathway. *Sci Rep* **6**, 27391 (2016).
35. Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nature genetics*, 1 (2018).
36. Tomasetti, C., Marchionni, L., Nowak, M.A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A* **112**, 118-23 (2015).
37. Hyman, D.M., Taylor, B.S. & Baselga, J. Implementing Genome-Driven Oncology. *Cell* **168**, 584-599 (2017).
38. Molina-Vila, M.A. *et al.* Activating mutations cluster in the “molecular brake” regions of protein kinases and do not associate with conserved or catalytic residues. *Human mutation* **35**, 318-328 (2014).
39. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
40. Ellrott, K. *et al.* Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell systems* **6**, 271-281. e7 (2018).
41. Wong, W.C. *et al.* CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27**, 2147-8 (2011).
42. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).

43. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**, 14330-14335 (2016).
44. Hulse, J.V., Khoshgoftaar, T.M. & Napolitano, A. Experimental perspectives on learning from imbalanced data. in *Proceedings of the 24th international conference on Machine learning* 935-942 (ACM, Corvalis, Oregon, USA, 2007).
45. Altmann, A., Tolosi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340-7 (2010).
46. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
47. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
48. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N. & Gaunt, T.R. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504-10 (2013).
49. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res* **11**, 863-74 (2001).
50. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118 (2011).
51. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016).
52. Jagadeesh, K.A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581-1586 (2016).
53. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
54. Gonzalez-Perez, A., Deu-Pons, J. & Lopez-Bigas, N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med* **4**, 89 (2012).
55. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* **23**, 703-713 (2017).
56. Olivier, M. *et al.* The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat* **19**, 607-14 (2002).
57. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-24 (2015).