CHASMplus 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

Summary

Large-scale cancer sequencing studies of patient cohorts have statistically implicated many cancer driver genes, with a long-tail of infrequently mutated genes. Here we present CHASMplus, a computational method to predict driver missense mutations, which is uniquely powered to identify rare driver mutations within the long-tail. We show that it substantially outperforms comparable methods across a wide variety of benchmark sets. Applied to 8,657 samples across 32 cancer types, CHASMplus identifies over 4,000 unique driver mutations in 240 genes, further distinguished by their specific cancer types. Our results support a prominent emerging role for rare driver mutations, with substantial variability in the frequency spectrum of drivers across cancer types. The trajectory of driver discovery may already be effectively saturated for certain cancer types, a finding with policy implications for future sequencing. As a resource to handle newly observed rare driver mutations, we systematically score every possible missense mutation across the genome.

Key words

cancer driver, missense mutation, rare drivers, TCGA

Significance

With the ever-growing pace of DNA sequencing of human tumors, the total number of detected mutations in cancer continues to accelerate. However, only a few mutations in each tumor may actually “drive” the growth of cancer, some of which can have value for diagnostic, prognostic, or therapeutic purposes. Based on a new rigorous statistical analysis of The Cancer Genome Atlas (TCGA), we find a prominent emerging role for rare missense mutations predicted to be “drivers” of cancer, which may have potential implications for genome-driven precision oncology, since rare driver mutations that are putatively actionable could be newly observed in a patient, thus, requiring personalized modeling and assessment. To extend beyond the TCGA, we provide a systematic resource to assess such newly observed missense mutations as cancer drivers. Lastly, we assess the driver landscape of human cancers and find that discovery for some cancer types are already approaching saturation.

Introduction

A growing set of mutations found in cancer genomes are recognized as clinically actionable (Hyman et al., 2017) with more putatively actionable mutations likely being verified in the future (Bailey et al., 2018). However, even interpreting the effect of somatic missense mutations, the most common type of protein-coding mutation found in human cancers(Vogelstein et al., 2013), is difficult because their impact on the fitness of cancer cells can be highly variable. Certain missense mutations are a critical step towards increasing the net growth of cells during the neoplastic process of cancer (drivers), while most others are benign passengers(Torkamani and Schork, 2008). For example, certain cancer types are known to be driven by well-established, highly prevalent missense mutations in oncogenes, such as KRAS G12D mutations in pancreatic ductal adenocarcinoma(Biankin et al., 2012) or IDH1 R132H mutations in gliomas (Parsons et al., 2008). Further, a literature curated database has compiled a list of approximately two thousand such driver missense mutations(Chakravarty et al., 2017), but due to the current limited throughput of functional validation assays it is likely incomplete. It has been hypothesized, though, that cancer driver mutations exhibit a long tail phenomenon with few common drivers and many rare drivers(Ding et al., 2010; Garraway and Lander, 2013), suggesting that numerous rare drivers remain to be discovered.

The task of identifying putative drivers from cancer sequencing studies has classically used statistical models that identify an excess number of mutations over expectation(Dees et al., 2012; Lawrence et al., 2013). However, most genes are large and, even within a driver gene, numerous passenger mutations are expected to accumulate by chance(Martincorena et al., 2017). 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(Yang et al., 2015), protein-protein interfaces(Engin et al., 2016; Porta-Pardo et al., 2015), and individual codons(Chang et al., 2016).

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 of the protein sequence, features of the local protein environment, annotations of molecular function, and biophysical characterizations of the amino acid substitution. Cancer-focused machine learning methods have previously tried to enhance performance by training cancer-specific background models (Carter et al., 2009; Mao et al., 2013) or boosting data with synthetic passenger missense mutations(Carter et al., 2009). 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(Martelotto et al., 2014). We and others have hypothesized that determining the impact of missense mutations requires proper context(Raphael et al., 2014), 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 statistically rigorous method, CHASMplus, for predicting the driver status of missense mutations. After careful benchmarking, we applied CHASMplus to 8,657 sequenced tumors from The Cancer Genome Atlas (TCGA) spanning 32 types of cancer. We explore the role for rare driver missense mutations in cancer and, when possible, relate predictions to supporting functional evidence. We provide an interactive resource for exploring driver missense mutations identified from the TCGA (http://karchinlab.github.io/CHASMplus) and a user-friendly tool (http://chasmplus.readthedocs.io/) to predict whether newly observed mutations from further sequencing are likely cancer drivers. Lastly, we examine the diversity of driver missense mutations across various types of cancer, which leads to a refined understanding of the likely trajectory of driver discovery with further sequencing.

Result

Overview of CHASMplus

We have developed a new method named CHASMplus that uses machine learning to discriminate somatic missense mutations (referred to hereafter as *missense mutations*) as either cancer drivers or passengers (Figure 1a, Methods). Unlike a recent analysis (Bailey et al., 2018), predictions can be done in a cancer type-specific manner or considered across multiple cancer types in aggregate ("pan-cancer"). Predictions utilize the Random Forest Algorithm, which consists of an ensemble of many randomized decision trees(Amit and Geman, 1997; Breiman, 2001), to score missense mutations. Each decision tree is trained on a random selection of training set examples and candidate features, via a recursive splitting process(Breiman, 1984) (Figure 1b). Missense mutations are only considered putative drivers if scores reach statistical significance, after controlling for sequence composition and multiple hypothesis testing with the Benjamini-Hochberg method (Figure S1a, methods). The resulting P-value distribution from CHASMplus suggest our statistical model is well calibrated (Figure S1).

An adequate training and testing procedure is critical for any approach based on machine learning. CHASMplus is trained using somatic mutation calls from The Cancer Genome Atlas (TCGA) covering 8,657 samples in 32 cancer types (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 (Figure S1, Methods). CHASMplus predictions are done with a rigorous gene holdout cross-validation protocol to avoid overfitting, by ensuring all mutations within a gene are within the same fold(Capriotti and Altman, 2011). Therefore, missense mutations are never scored by a Random Forest trained on any missense mutation harbored by the same gene. Finally, predicted scores from CHASMplus are weighted by the 20/20+ driver gene score, producing gene-weighted (gwCHASMplus) scores (Figure 1c, Methods).

CHASMplus represents the context of missense mutations at multiple scales. The Random Forest was trained on 95 features (Table S1), and the 34 with net-positive feature importance are shown in Figure 1d (Methods). Important features assess five broad categories: multi-resolution missense mutation hotspots (HotMAPS 1D algorithm(Tokheim et al., 2016a)), evolutionary conservation/human germline variation, molecular function annotations (*e.g.*, protein-protein interface annotations from(Meyer et al., 2018)), 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(Kumar et al., 2016), 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 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 (Figure 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 performance metric (Figure 2b-c, Figure S2a-h), which has been used in many prior studies of variant effect prediction (Adzhubei et al., 2010; Ioannidis et al., 2016; Kircher et al., 2014; Kumar et al., 2016; Mao et al., 2013). Overall, CHASMplus had the highest auROC on each benchmark, with 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(Bradley, 1997). 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). An alternative metric called the area under the Precision-Recall curve yielded similar conclusions as auROC (Figure S2i-k, Methods).

We used three benchmarks based on *in vitro* experiments (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(Ng et al., 2018). 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(Berger et al., 2016), 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)(Petitjean et al., 2007), 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 tumor growth fitness in a competition assay. The first was performed from mutations observed in lung cancers (44 missense mutations)(Berger et al., 2016) and the second from mutations observed in 27 cancer types (71 missense mutations)(Kim et al., 2016). 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(Futreal et al., 2004) as drivers (Methods). We found that the gene weighted CHASMplus scores (auROC=0.908) were substantially better at this whole exome-wide prioritization task compared to the unweighted CHASMplus scores (auROC=0.856) (p<2.2e-16, delong test). CHASMplus scores were also significantly better than the next best method (ParsSNP) (p=0.008, 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 (Methods). CHASMplus significantly outperformed all other methods, the nearest being ParsSNP (p=7e-14, delong test).

CHASMplus identifies putative driver mutations in 32 cancer types

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(Cancer Genome Atlas Research et al., 2013).

Using CHASMplus, we identified 3,527 unique missense mutations as statistically significant putative drivers from our pan-cancer analysis at an estimated false discovery rate of 1% (Table S2). The pan-cancer results had a substantial overlap with a prior pan-cancer analysis done by the TCGA (Figure S3a-b) (Bailey et al., 2018). When applied to each cancer type individually, the number found significant varied substantially from 8 in thymoma to 572 in bladder urothelial carcinoma, with a median of 78 (Figure S3c-d, Table S3). In total, 479 unique driver missense mutations were only identified by the cancer type-specific analyses. 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 yet sufficiently characterized for inclusion in OncoKB. Moreover, CHASMplus had the best performance on a previously reported benchmark (Porta-Pardo et al., 2017) of cancer type-specific driver predictions (Figure S4, Methods). Altogether across the pan-cancer and cancer type-specific analyses, 4,006 unique driver missense mutations were identified by CHASMplus, of which 2,037 were not found by OncoKB or the official TCGA analysis, indicating a potentially expanded landscape of putative driver missense mutations of interest for further examination.

CHASMplus identifies both common and rare cancer drivers

The long tail hypothesis, initially proposed from examining the overall mutation frequency of driver genes(Ding et al., 2010; Garraway and Lander, 2013), 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%) frequency driver missense mutations varied substantially among cancer types (Figure 3a). 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 slightly greater than for common drivers (35.5% and 35.4%, respectively) and 4-fold greater than found by a previous method (8%, P<2.2e-16, Fisher's exact test)(Chang et al., 2016). After adjusting for sample size, we observed that the average tumor 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, Figure 3b). Given that driver mutations likely arise in tumors from a combination of clonal selection and the mutation rate to generate the mutation in the first place (Greaves and Maley, 2012), the latter may have a larger role for the origins of less frequent driver mutations compared to their common counterpart.

Conceivably, the different frequency spectra of driver missense mutations across cancer types could also arise from differential selection between subtypes within a given type of cancer. A common driver mutation in an uncommon subtype, could be perceived, overall, as rare. To test this, we analyzed whether driver missense mutations within a gene showed noticeable enrichment for cancer samples that are a particular subtype. For the 12 cancer types with available subtype information (Sanchez-Vega et al., 2018), 55 out of 223 genes (24.7%) found with significant missense mutations by CHASMplus were differentially enriched in cancer subtypes (q-value≤0.1, chi-squared test, Figure 3c, Table S4, Figure S5). The modest percentage of genes suggests that subtype-specificity only partly explains differences in the driver mutation frequency spectrum between cancer types. Several genes showed strong subtype specificity, consistent with prior literature, such as *NFE2L2* mutations in the squamous cell subtype in ESCA (Network, 2017), *TP53* mutations in Human Papillomavirus-negative tumors in HNSC (Network, 2015), and *KIT* mutations in testicular seminomas (Kemmer et al., 2004). The mutational subtype enrichment in breast invasive carcinoma (BRCA) is consistent with previously reported divergent gene expression patterns in tumors related to the P53 pathway and the PI3K/AKT/MTOR signaling across BRCA subtypes (Dinalankara et al., 2018). However, in some cancer types, specifically glioblastoma (GBM) and low grade glioma (LGG), the subtypes were defined based on driver mutation status of the genes IDH1/IDH2, so it was not surprising there was strong enrichment.

Rare driver missense mutations exist not only in rarely mutated 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(Jeong et al., 2016), contained common driver missense mutations in our pan-cancer analysis at residue positions 179 and 183, which is located at the protein interface composing the phosphatase 2A complex (Figure 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) (Figure 3e). This is supportive of previous experimental work implicating rare cancer driver mutations in commonly mutated cancer driver genes(Kim et al., 2016).

Truncating or likely loss-of-function mutations are typical hallmarks of tumor suppressor genes (Vogelstein et al., 2013). However, the role of driver missense mutations may be under characterized in tumor suppressor genes, since these mutations are more diverse and occur over a larger region than in oncogenes (Porta-Pardo et al., 2017; Tokheim et al., 2016a). As a case in point, the tumor suppressor gene *CASP8* contains many truncating variants, while all of the putative driver missense mutations identified by CHASMplus were considered rare (Figure 3f). *CASP8* is a member of the apoptosis pathway and recently has been associated with a potential role in immune evasion in cancer (Rooney et al., 2015; Thorsson et al., 2018).

We explored functional evidence to support whether the rare driver missense mutations in *CASP8*, predicted by CHASMplus, behaved similarly to truncating variants. Thorsson et al. previously characterized immune phenotypes in TCGA tumor samples, i.e., levels of immune cell infiltrates and expression of immune-related genes (Thorsson et al., 2018). For 12 immune-related phenotypes, we compared tumor samples with driver missense mutations or truncating mutations in *CASP8* to control samples with no mutations in *CASP8*. In Head & Neck Squamous Cell Carcinoma (HNSC), both types of mutated samples had higher estimated levels of leukocytes, CD8 T-cells and Th1 response than controls (p<0.001, Mann-Whitney U test, for all except truncating mutations in Th1 response p=5.06E-03), and significantly elevated expression of key genes involved in tumor immunity, i.e., PD-1 (PDCD1; missense p=6.59E-05, truncating p=1.70E-02), PD-L1 (CD274; missense p=2.39E-05, truncating p=2.19E-04), CD8A (missense p=2.65E-05, truncating p=3.33E-03), and CTLA4 (missense p=1.81E-03, truncating p=2.67E-02) (Figure 4). A similar trend was seen in other cancer types with significant mutations (Figure

S6). Conventional wisdom has suggested that because rare missense mutations in tumor suppressor genes do not tend to cluster in protein sequence, they are solely passenger events (Vogelstein et al., 2013). However, our work suggests that rare driver missense mutations in *CASP8* and perhaps in other tumor suppressor genes may be relevant to tumor immunophenotypes.

CHASMplus mimics saturation mutagenesis

To handle newly arising rare driver mutations not observed in large-scale sequencing studies, we have precomputed CHASMplus scores for all possible missense mutations across the whole genome. We compared these scores (Figure 5a) to two saturation mutagenesis experiments characterizing the known tumor suppressor gene *PTEN*, a lipid phosphatase of phosphatidylinositol (3,4,5)-trisphosphate, an important signaling molecule in the PI3K signaling pathway, which is often dysregulated in human cancers (Sanchez-Vega et al., 2018). The first study systematically measured lipid phosphatase activity (Mighell et al., 2018), while the second study measured intracellular PTEN protein abundance (Matreyek et al., 2018), potentially an indicator of thermodynamic stability. Concordant with its tumor suppressor role, driver scores from CHASMplus are negatively correlated with both lipid phosphatase activity and PTEN protein abundance (Figure 5b-c, Table S5), indicating CHASMplus identifies functionally damaging mutations. Interestingly, CHASMplus is more correlated with each mutagenesis study (lipid phosphatase activity spearman ρ=-0.52, protein abundance spearman ρ=-0.43) than they are to each other (spearman ρ=0.35), suggestive that CHASMplus is capturing multiple modes of damage in PTEN (Figure 5d).

We observed that driver missense mutations identified by CHASMplus in the TCGA, regardless of frequency, had significantly lower lipid phosphatase activity than other missense mutations in *PTEN* (common: p=0.008; intermediate: p=1.9e-9; rare: p=1.6e-18; Mann-Whitney U test, Figure 5e). Moreover, the median lipid phosphatase activity monotonically decreased as the observed frequency increased in the TCGA, ultimately, to comparable levels as loss-of-function mutations (truncating variants: p=1.6e-112, Mann-Whitney U test). A likely explanation is that the strength of the functional consequence on lipid phosphatase activity in PTEN impacts the degree of clonal selection in tumors, resulting in more damaging PTEN variants being more frequently observed. In contrast, common (significantly higher, p=3.6e-3, Mann-Whitney U test) or intermediate (p=0.17, Mann-Whitney U test) frequency driver missense mutations in PTEN did not have a lower protein abundance, indicating that lower protein abundance may only be a viable effect for rare driver missense mutations in PTEN (p=9.2e-5, Mann-Whitney U test, Figure 5f), while truncating variants had a pronounced lower protein abundance (p=5e-59, Mann-Whitney U test).

The trajectory of driver discovery across diverse cancer types

We next sought to understand whether cancer types fundamentally differed in their usage of driver missense mutations. We found the diversity and prevalence of driver missense mutations varied considerably across TCGA cancer types (Figure 6a, Methods). We defined diversity with respect to the distribution of driver missense mutations across codons and prevalence with respect to the overall frequency of 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, Methods). 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.

Are there substantially more cancer driver missense mutations yet to be discovered? If discovery was measured by the number of unique driver missense mutations identified, subsampling analysis showed all cancer types had a linear increase ($R^2 > 0.5$) with no evidence of saturation at current sample sizes (Figure S7a). However, we observed substantial variability in trajectories if discovery was measured by driver prevalence (average number of driver missense mutations per cancer sample) (Figure 6b), 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 (Armenia et al., 2018), 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 (Methods, Table

S6, Figure S7b-c). 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(Tomasetti et al., 2015). 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(Hyman et al., 2017). To address prior limitations of computational methods (Martelotto et al., 2014; Molina-Vila et al., 2014), 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. After careful evaluation, 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(Ding et al., 2010; Garraway and Lander, 2013) posits that there are many rare driver mutations in human cancers. However, a rigorous foundation for this hypothesis had been limited by the lack of statistical power to move beyond implicating genes towards understanding individual mutations (Methods, Figure S7). To overcome this limitation, we leveraged the improvements made in CHASMplus to systematically predict driver missense mutations in 8,657 cancer samples from the TCGA. Although individually rare, rare driver missense mutations, collectively, comprise a prominent emerging role in cancer, consistent with the long tail hypothesis. However, not every type of cancer is the same; our study is the first, to our knowledge, to systematically show that the prevalence and diversity of driver missense mutations is highly variable across cancer types. We find several factors likely influence the diversity of driver missense mutations in cancer, including tumor mutation burden, the type of gene (i.e., tumor suppressor genes), the functional strength of the mutation, and the mutation's subtype specificity. Other factors may also contribute, such as epistasis between mutations (Kent et al., 2015; Skoulidis et al., 2015), interactions between mutations and the (micro)environment (Marty et al., 2017; Rooney et al., 2015), selective pressures based on a broader cell-of-origin (Bailey et al., 2018), or competition from other types of somatic alterations. The diversity of driver missense mutations supports the critical role of understanding the overall contribution of rare 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, individually, they are so infrequent, rare driver mutations in newly sequenced tumors may not have been previously observed in the TCGA or other large-scale sequencing projects. A similar problem exists for interpreting rare germline variants impacting susceptibility to disease (Bomba et al., 2017; Consortium et al., 2015). One approach to this issue is saturation mutagenesis experiments, which functionally assess all mutations in genes of interest, regardless of whether they have been seen before (Fowler and Fields, 2014). These experiments have so far been limited to handful of well-characterized genes and are not yet available for most genes implicated in cancer (Bailey et al., 2018). Consequently, computational methods, like CHASMplus, are needed to prioritize mutations for low- and medium- throughput studies. We therefore have precomputed the score of every possible missense mutation across the genome, effectively an *in silico* saturation mutagenesis across all genes to score as of yet unobserved mutations that are potential cancer drivers. We provide mutation scores for each of the cancer types available in the TCGA, as well as pan-cancer scores (http://chasmplus.readthedocs.io/). We also have provided an interactive portal for exploring driver mutations (http://karchinlab.github.io/CHASMplus).

There are several limitations to our study. Although missense mutations are the most frequent somatic alteration in cancer (Vogelstein et al., 2013), CHASMplus only predicts missense mutations; however, in principle, our approach could be extended to other types of alterations. Further, CHASMplus is specifically optimized for somatic mutations in cancer, as such, it is not a general-purpose pathogenicity predictor of germline variants. Also, CHASMplus is trained using semi-supervised labels, emphasizing driver mutations in low mutation burden tumor samples, which may result in underperformance for driver mutations specific to high mutation burden cancer types or hypermutated tumors. We therefore performed predictions on melanoma separately (Table S7). Lastly, although CHASMplus can be applied to targeted gene panels, the estimation of statistical significance requires a correction for the specific genes that are targeted.

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(Armenia et al., 2018; Lawrence et al., 2014). 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 rarely mutated driver genes. Future work will further elucidate a broader range of driver mutations, including those within non-coding regions of the genome, at different stages of carcinogenesis, such as in pre-cancerous lesions, and in response to therapeutic treatment.

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.

Declaration of Interests

We have no competing financial interests to disclose.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Altmann, A., Tolosi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics *26*, 1340-1347.

Amit, Y., and Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. Neural Computation *9*, 1545-1588.

Armenia, J., Wankowicz, S.A., Liu, D., Gao, J., Kundra, R., Reznik, E., Chatila, W.K., Chakravarty, D., Han, G.C., and Coleman, I. (2018). The long tail of oncogenic drivers in prostate cancer. Nature genetics, 1.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B.*, et al.* (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371-385 e318.

Berger, A.H., Brooks, A.N., Wu, X., Shrestha, Y., Chouinard, C., Piccioni, F., Bagul, M., Kamburov, A., Imielinski, M., Hogstrom, L.*, et al.* (2016). High-throughput Phenotyping of Lung Cancer Somatic Mutations. Cancer Cell *30*, 214-228.

Biankin, A.V., Waddell, N., Kassahn, K.S., Gingras, M.C., Muthuswamy, L.B., Johns, A.L., Miller, D.K., Wilson, P.J., Patch, A.M., Wu, J.*, et al.* (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature *491*, 399-405.

Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. Genome Biol *18*, 77.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition *30*, 1145-1159.

Breiman, L. (1984). Classification and regression trees.

Breiman, L. (2001). Random Forests. Mach Learn *45*, 5-32.

Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113-1120.

Capriotti, E., and Altman, R.B. (2011). A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics *98*, 310-317.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res *69*, 6660-6667.

Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics *14 Suppl 3*, S3.

Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H.*, et al.* (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol *2017*.

Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B.*, et al.* (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol *34*, 155-163.

Consortium, U.K., Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M.*, et al.* (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82-90.

Davis, J., and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Paper presented at: Proceedings of the 23rd international conference on Machine learning (ACM).

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R.*, et al.* (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res *22*, 1589-1598.

Dinalankara, W., Ke, Q., Xu, Y., Ji, L., Pagane, N., Lien, A., Matam, T., Fertig, E.J., Price, N.D., Younes, L.*, et al.* (2018). Digitizing omics profiles by divergence from a baseline. Proc Natl Acad Sci U S A *115*, 4545-4552.

Ding, L., Wendl, M.C., Koboldt, D.C., and Mardis, E.R. (2010). Analysis of next-generation genomic data in cancer: accomplishments and challenges. Hum Mol Genet *19*, R188-196.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., and McLellan, M. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell systems *6*, 271-281. e277.

Engin, H.B., Kreisberg, J.F., and Carter, H. (2016). Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. PLoS One *11*, e0152929.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L.*, et al.* (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res *45*, D777-D783.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat Methods *11*, 801-807.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Garraway, L.A., and Lander, E.S. (2013). Lessons from the cancer genome. Cell *153*, 17-37.

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

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306.

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7.

Hulse, J.V., Khoshgoftaar, T.M., and Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th international conference on Machine learning (Corvalis, Oregon, USA: ACM), pp. 935-942.

Hyman, D.M., Taylor, B.S., and Baselga, J. (2017). Implementing Genome-Driven Oncology. Cell *168*, 584-599.

Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D.*, et al.* (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet *99*, 877-885.

Jagadeesh, K.A., Wenger, A.M., Berger, M.J., Guturu, H., Stenson, P.D., Cooper, D.N., Bernstein, J.A., and Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet *48*, 1581-1586.

Jeong, A.L., Han, S., Lee, S., Su Park, J., Lu, Y., Yu, S., Li, J., Chun, K.H., Mills, G.B., and Yang, Y. (2016). Patient derived mutation W257G of PPP2R1A enhances cancer cell migration through SRC-JNK-c-Jun pathway. Sci Rep *6*, 27391.

Kemmer, K., Corless, C.L., Fletcher, J.A., McGreevey, L., Haley, A., Griffith, D., Cummings, O.W., Wait, C., Town, A., and Heinrich, M.C. (2004). KIT mutations are common in testicular seminomas. The American journal of pathology *164*, 305-313.

Kent, D.G., Ortmann, C.A., and Green, A.R. (2015). Effect of mutation order on myeloproliferative neoplasms. N Engl J Med *372*, 1865-1866.

Kim, E., Ilic, N., Shrestha, Y., Zou, L., Kamburov, A., Zhu, C., Yang, X., Lubonja, R., Tran, N., Nguyen, C.*, et al.* (2016). Systematic Functional Interrogation of Rare Cancer Variants Identifies Oncogenic Alleles. Cancer Discov *6*, 714-726.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315.

Kumar, R.D., Swamidass, S.J., and Bose, R. (2016). Unsupervised detection of cancer driver mutations with parsimony-guided learning. Nat Genet *48*, 1288-1294.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A.*, et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Mao, Y., Chen, H., Liang, H., Meric-Bernstam, F., Mills, G.B., and Chen, K. (2013). CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLoS One *8*, e77945.

Martelotto, L.G., Ng, C.K., De Filippo, M.R., Zhang, Y., Piscuoglio, S., Lim, R.S., Shen, R., Norton, L., Reis-Filho, J.S., and Weigelt, B. (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol *15*, 484.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029-1041 e1021.

Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M.J., van de Haar, J., Engin, H.B., de Prisco, N., Ideker, T., Hildebrand, W.H., Font-Burgada, J.*, et al.* (2017). MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell *171*, 1272-1283 e1215.

Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J.*, et al.* (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet *50*, 874-882.

Meyer, M.J., Beltrán, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., and Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. Nature Methods.

Mighell, T.L., Evans-Dutson, S., and O'Roak, B.J. (2018). A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. Am J Hum Genet *102*, 943-955.

Molina-Vila, M.A., Nabau-Moretó, N., Tornador, C., Sabnis, A.J., Rosell, R., Estivill, X., Bivona, T.G., and Marino-Buslje, C. (2014). 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.

Network, C.G.A. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature *517*, 576.

Network, C.G.A.R. (2017). Integrated genomic characterization of oesophageal carcinoma. Nature *541*, 169.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat Methods *12*, 453-457.

Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res *11*, 863-874.

Ng, P.K., Li, J., Jeong, K.J., Shao, S., Chen, H., Tsang, Y.H., Sengupta, S., Wang, Z., Bhavana, V.H., Tran, R.*, et al.* (2018). Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell *33*, 450-462.e410.

Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L.*, et al.* (2008). An integrated genomic analysis of human glioblastoma multiforme. Science *321*, 1807-1812.

Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S.V., Hainaut, P., and Olivier, M. (2007). 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*, 622-629.

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. PLoS Comput Biol *11*, e1004518.

Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., Lopez-Bigas, N., Getz, G., and Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nat Methods *14*, 782-788.

Raphael, B.J., Dobson, J.R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med *6*, 5.

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

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E.*, et al.* (2015). 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*, 405-424.

Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell *160*, 48-61.

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one *10*, e0118432.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., and Saghafinia, S. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell *173*, 321-337. e310.

Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics *29*, 1504-1510.

Skoulidis, F., Byers, L.A., Diao, L., Papadimitrakopoulou, V.A., Tong, P., Izzo, J., Behrens, C., Kadara, H., Parra, E.R., Canales, J.R.*, et al.* (2015). Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. Cancer Discov *5*, 860-877.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A.*, et al.* (2018). The Immune Landscape of Cancer. Immunity *48*, 812-830 e814.

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D.M., Kim, R., Ryan, M., Masica, D.L., and Karchin, R. (2016a). Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res *76*, 3719-3731.

Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016b). Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A *113*, 14330-14335.

Tomasetti, C., Marchionni, L., Nowak, M.A., Parmigiani, G., and Vogelstein, B. (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. Proc Natl Acad Sci U S A *112*, 118-123.

Torkamani, A., and Schork, N.J. (2008). Prediction of cancer driver mutations in protein kinases. Cancer Res *68*, 1675-1682.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Wong, W.C., Kim, D., Carter, H., Diekhans, M., Ryan, M.C., and Karchin, R. (2011). CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics *27*, 2147-2148.

Yang, F., Petsalaki, E., Rolland, T., Hill, D.E., Vidal, M., and Roth, F.P. (2015). Protein domain-level landscape of cancer-type-specific somatic mutations. PLoS Comput Biol *11*, e1004147.

Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M.*, et al.* (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med *23*, 703-713.

Figure Legends

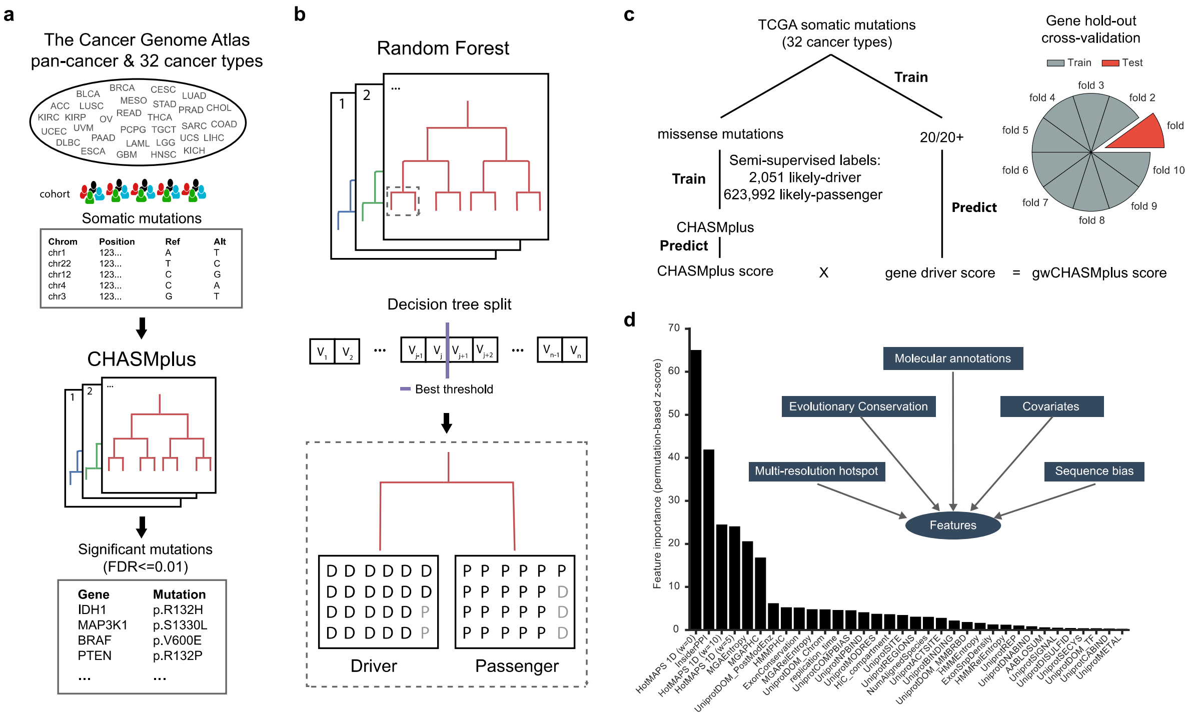
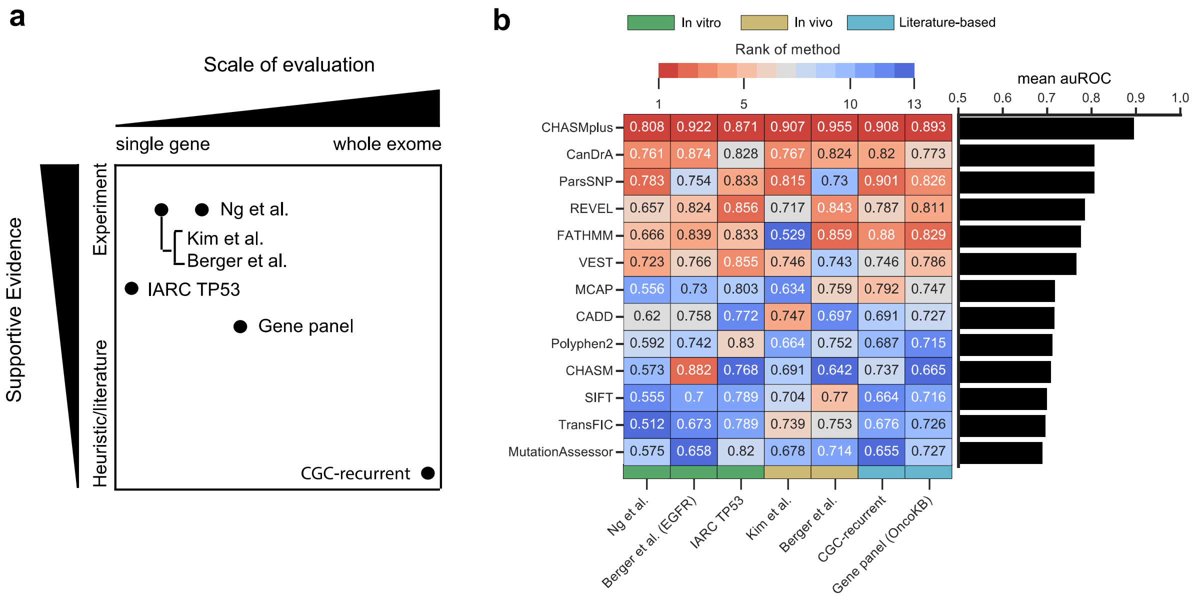


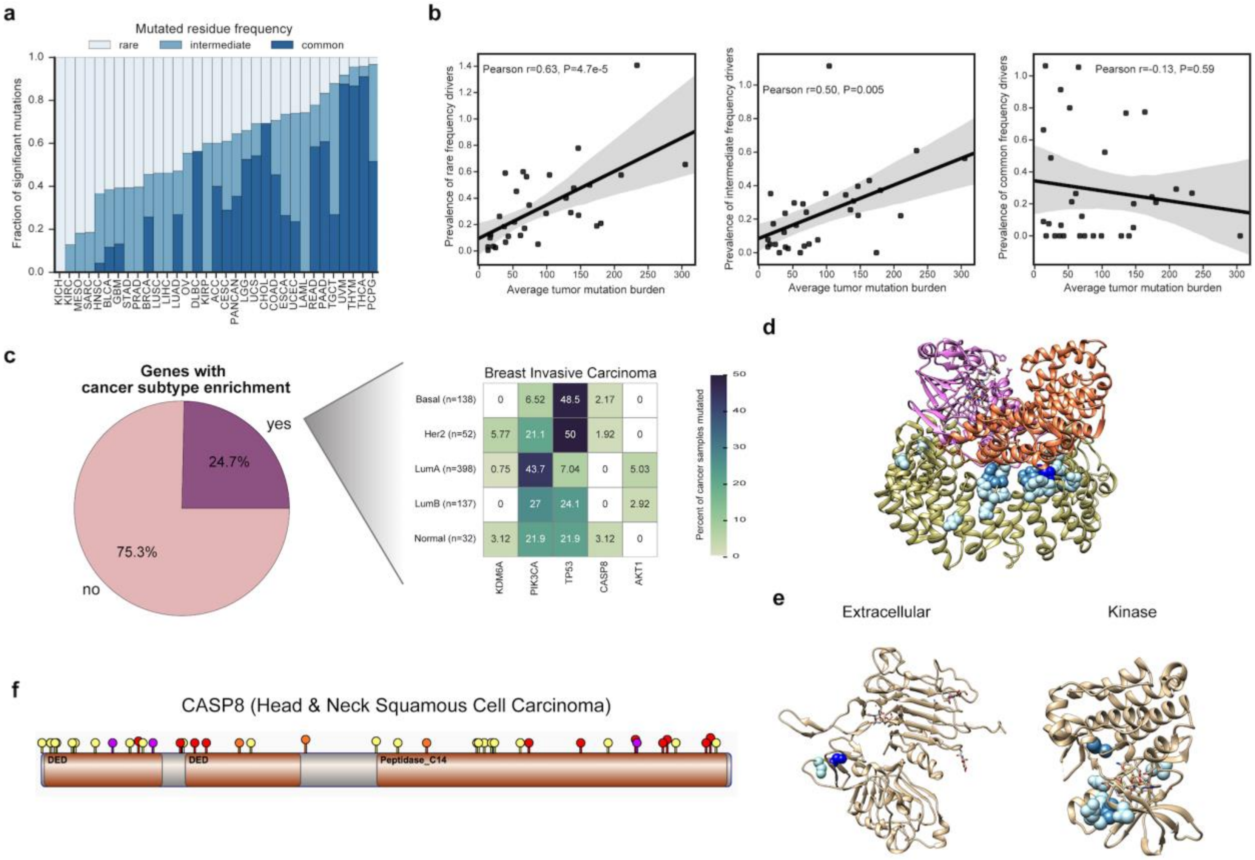
Figure 1. Overview of CHASMplus. a) CHASMplus was applied to somatic missense mutations in tumors from 32 different cancer types found in The Cancer Genome Atlas (TCGA). Significant putative driver missense mutations were identified at a False Discovery Rate (FDR) threshold of 1%. **b)** CHASMplus predictions utilize the 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. **c)** Diagram of training and testing procedure by CHASMplus. **d)** 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.



c

Method	Ng et al.	Berger et al. (EGFR)	IARC TP53	Kim et al.	Berger et al.	CGC-recurrent	Gene panel (OncoKB)
CHASMplus	1	1	1	1	1	1	1
CanDrA	0.007	0.2	3.00E-08	0.02	0.05	2.00E-107	8.00E-47
ParsSNP	0.05	0.001	2.00E-07	0.03	0.002	0.008	7.00E-14
REVEL	4.00E-11	0.06	0.02	0.002	0.1	3.00E-200	3.00E-27
FATHMM	2.00E-10	0.2	1.00E-06	6.00E-10	0.1	0.0002	4.00E-16
VEST	3.00E-05	0.004	0.03	0.007	0.008	0	3.00E-39
MCAP	1.00E-21	0.01	2.00E-17	0.0002	0.01	5.00E-172	2.00E-67
CADD	3.00E-15	0.03	3.00E-25	0.007	0.004	0	2.00E-75
Polyphen2	3.00E-19	0.01	1.00E-15	8.00E-07	0.006	0	6.00E-100
CHASM	3.00E-18	0.4	6.00E-30	0.0007	0.0004	1.00E-261	5.00E-86
SIFT	1.00E-19	0.003	1.00E-23	0.0008	0.01	0	9.00E-78
TransFIC	8.00E-24	0.004	1.00E-23	0.005	0.01	0	3.00E-57
MutationAssessor	3.00E-19	0.0002	1.00E-09	0.0002	0.02	0	1.00E-61

Figure 2. Benchmarking cancer driver prediction. 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)** Table of reported P values (delong test) from comparing the auROC of each method against that of CHASMplus.



	KDM6A	PIK3CA	TP53	CASP8	AKT1
Basal (n=138)	0	6.52	48.5	2.17	0
Her2 (n=52)	5.77	21.1	50	1.92	0
LumA (n=398)	0.75	43.7	7.04	0	5.03
LumB (n=137)	0	27	24.1	0	2.92
Normal (n=32)	3.12	21.9	21.9	3.12	0

Figure 3. Frequency landscape of driver missense mutations. a) Proportion of overall frequency of driver somatic missense mutations found to be 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). **f)** Lollipop plot of driver missense mutations identified by CHASMplus (yellow), and likely truncating variants (frameshift insertion or deletion: purple, nonsense mutation: red, and splice site mutation: orange) in *CASP8* for Head & Neck squamous cell carcinoma (HNSC). TCGA cancer type acronyms are listed in methods.

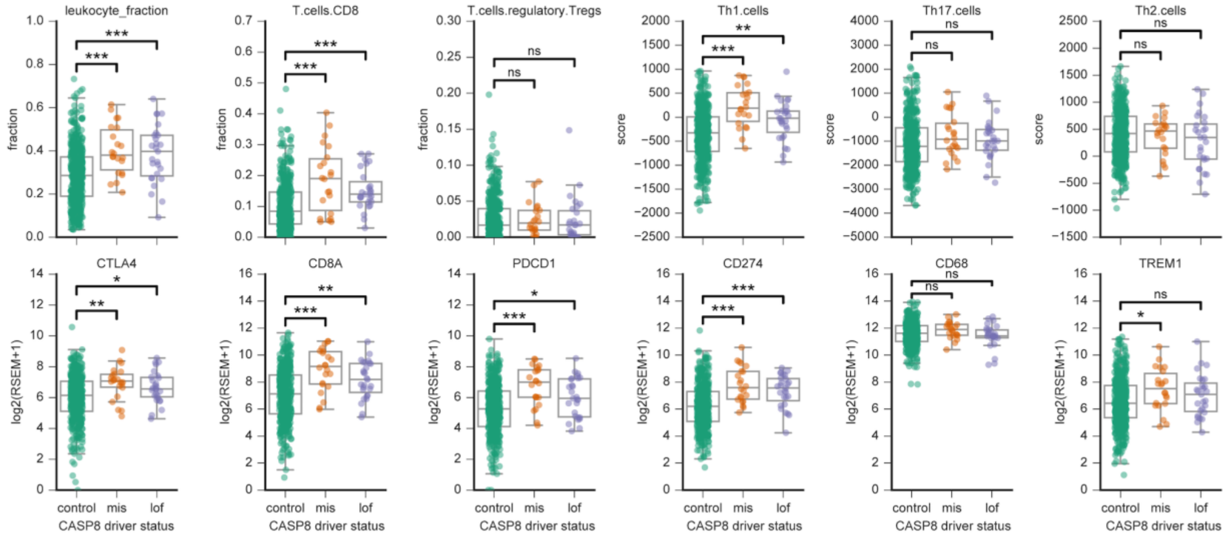


Figure 4. Effect of rare driver missense mutations on immune phenotypes. Correlation of putative driver mutations in *CASP8* with immune phenotypes in HNSC tumors, where control samples have no *CASP8* mutations, "mis" indicates samples with driver missense mutations identified by CHASMplus, and "lof" is likely loss-of-function variants (nonsense, frameshift insertion/deletions, splice site, translation start site, and nonstop mutations). Top row, immune cell/phenotype response inferred from DNA methylation or gene expression from (Thorsson et al., 2018). Bottom row, gene expression values from RNASeq for several important immune-related genes reported in (Thorsson et al., 2018). Mann Whitney U test: *=p<0.05, **=p<0.01, and ***=p<0.001.

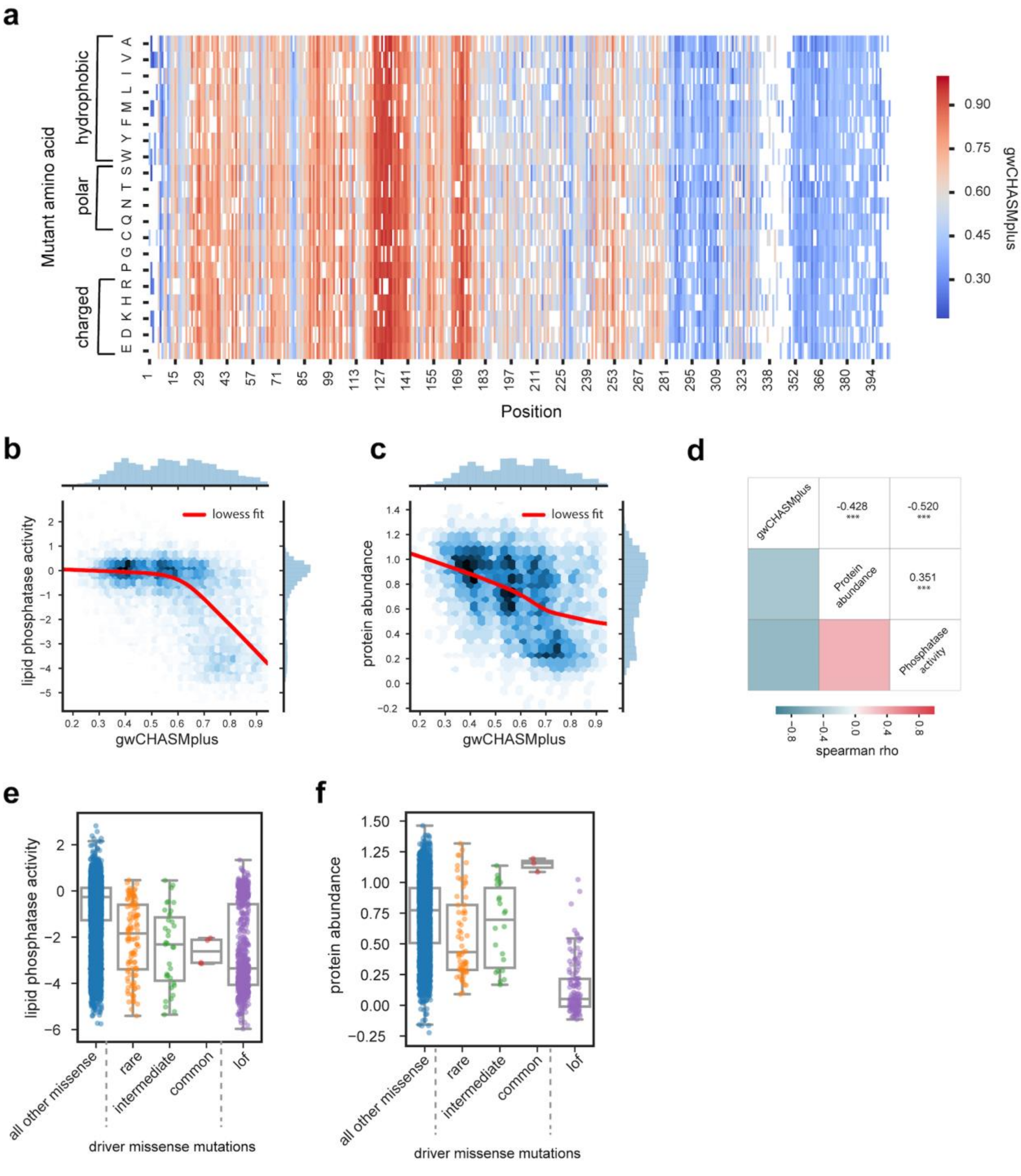


Figure 5. CHASMplus predictions correlate with saturation mutagenesis experiments in PTEN. a) Heatmap displaying gene-weighted CHASMplus scores (gwCHASMplus) across all possible missense mutations in PTEN assessed as high-confidence by a prior saturation mutagenesis experiment (Mighell et al., 2018). Correlation of gwCHASMplus with a saturation mutagenesis experiments measuring PTEN **b)** lipid phosphatase activity by (Mighell et al., 2018)

and **c)** protein abundance by (Matreyek et al., 2018). **d)** Comparison of the spearman correlation of gwCHASMplus with the two PTEN saturation mutagenesis experiments, as well as the correlation between the two experiments. ***=p<0.001. Comparison of **e)** PTEN lipid phosphatase activity or **f)** protein abundance for driver missense mutations identified from the TCGA to other missense mutations and loss-of-function mutations in (Mighell et al., 2018). Driver missense mutations are stratified by their maximum frequency in the TCGA cohort (common: >5% of cancer samples, intermediate: 1-5%, and rare: <1%).

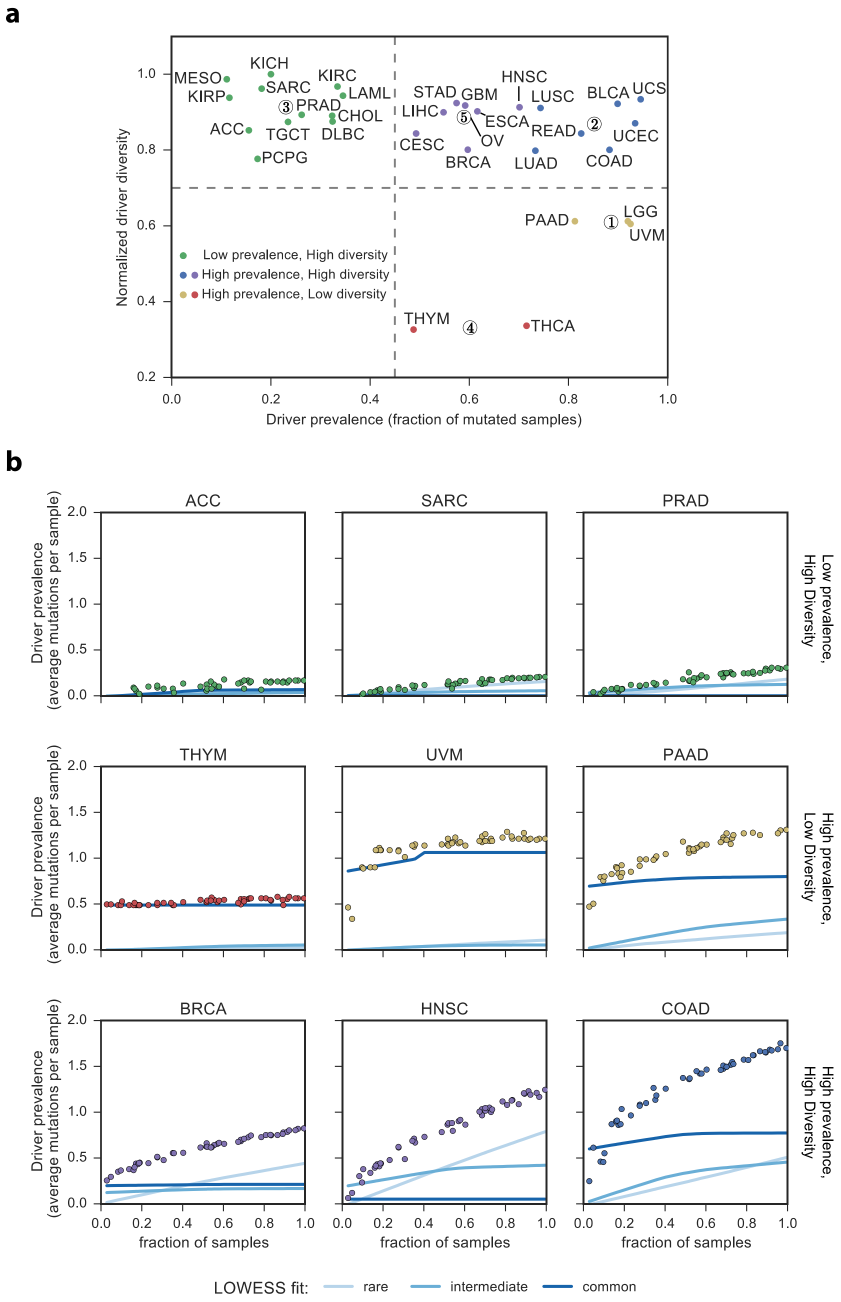


Figure 6. Characteristics and trajectory of driver discovery for missense mutations. a) Plot displaying normalized driver diversity and driver prevalence (fraction of samples mutated) for driver somatic missense mutations in 32 cancer types. K-means clustering identified 5 clusters with centroids shown as numerically designated circles. **b)** Prevalence of driver missense mutations identified by CHASMplus as a function of sample size. Lines represent LOWESS fit to different rarities of driver missense mutations. All TCGA cancer type acronyms are in the Methods.

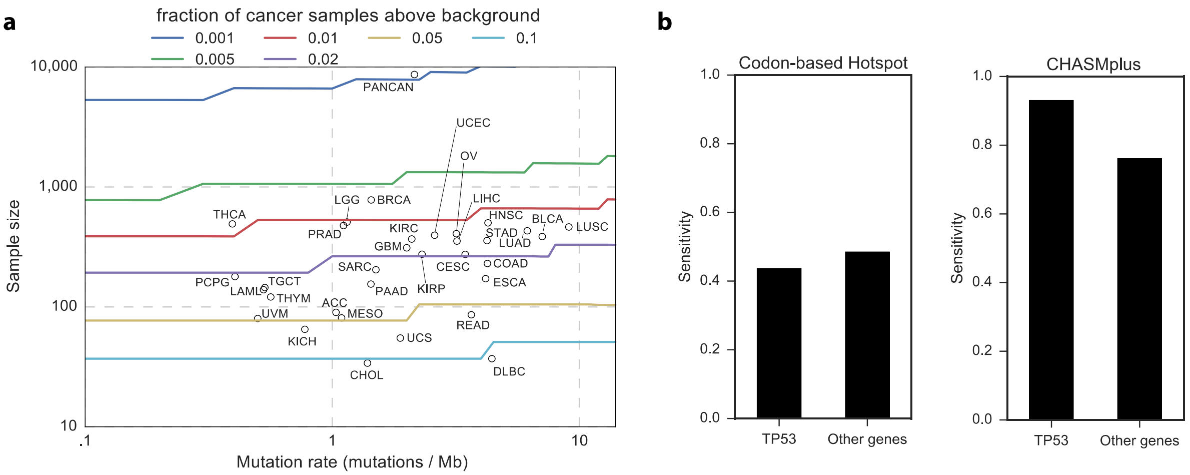
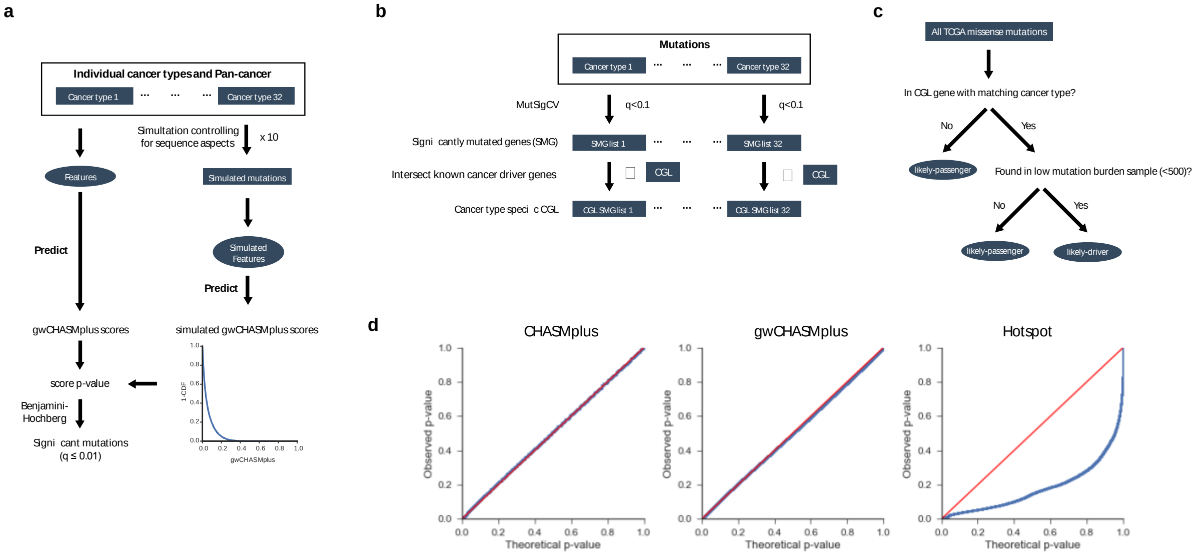
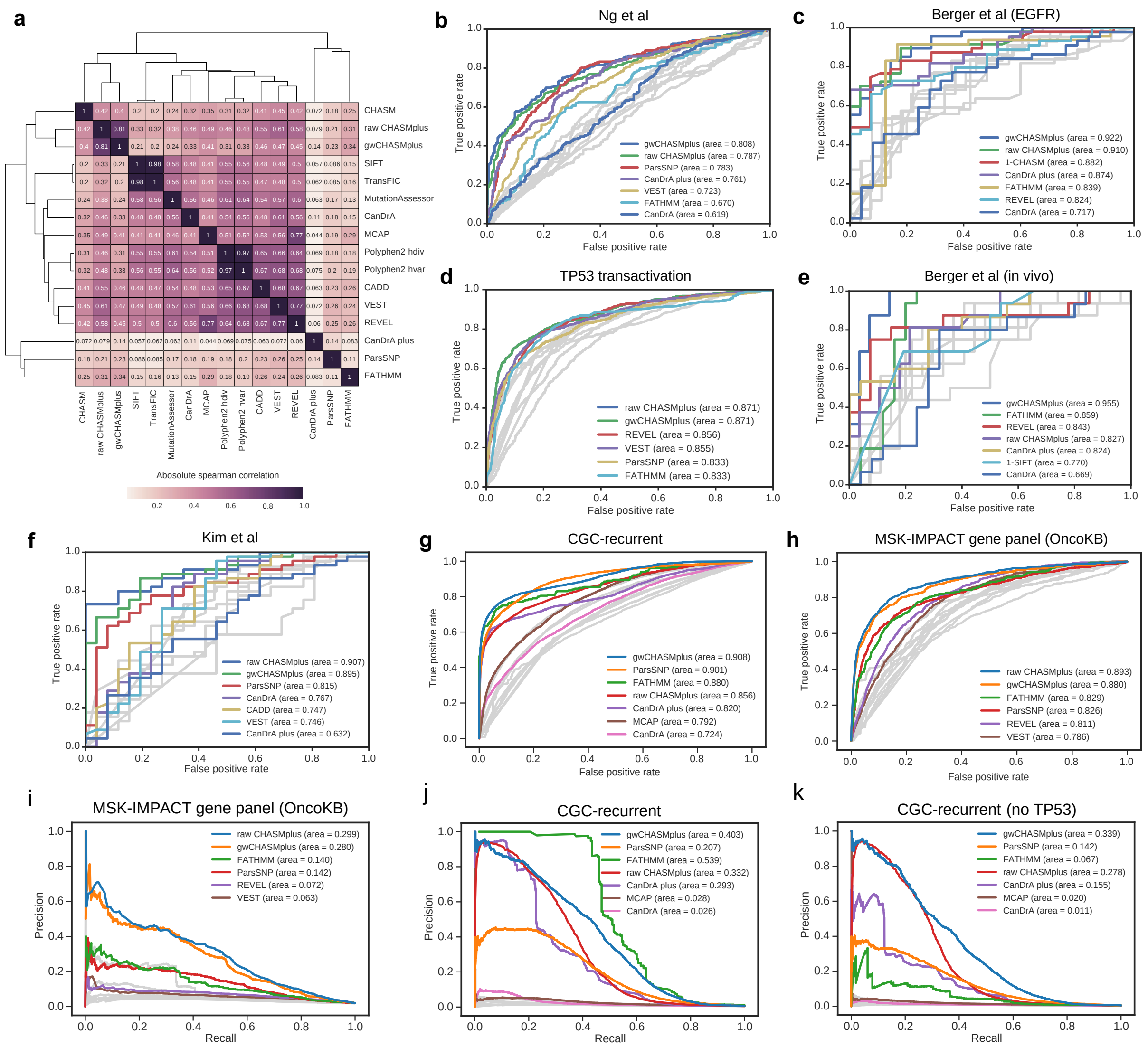


Figure 7. Limited statistical power of hotspot detection. a) Statistical power to detect significantly elevated number of non-silent mutations for individual codons as a function of sample size and mutation rate. Circles represent each cancer type from the TCGA and is placed according to sample size and median mutation rate. Curves are colored by the frequency of driver mutations (fraction of non-silent mutated cancer samples above the expected background mutation rate). If a circle is below a curve, then hotspot detection is not yet sufficiently powerful to detect driver mutations of that frequency. **b)** Bar graph comparing sensitivity to detect labeled oncogenic driver missense mutations from OncoKB between CHASMplus and a hotspot detection approach.

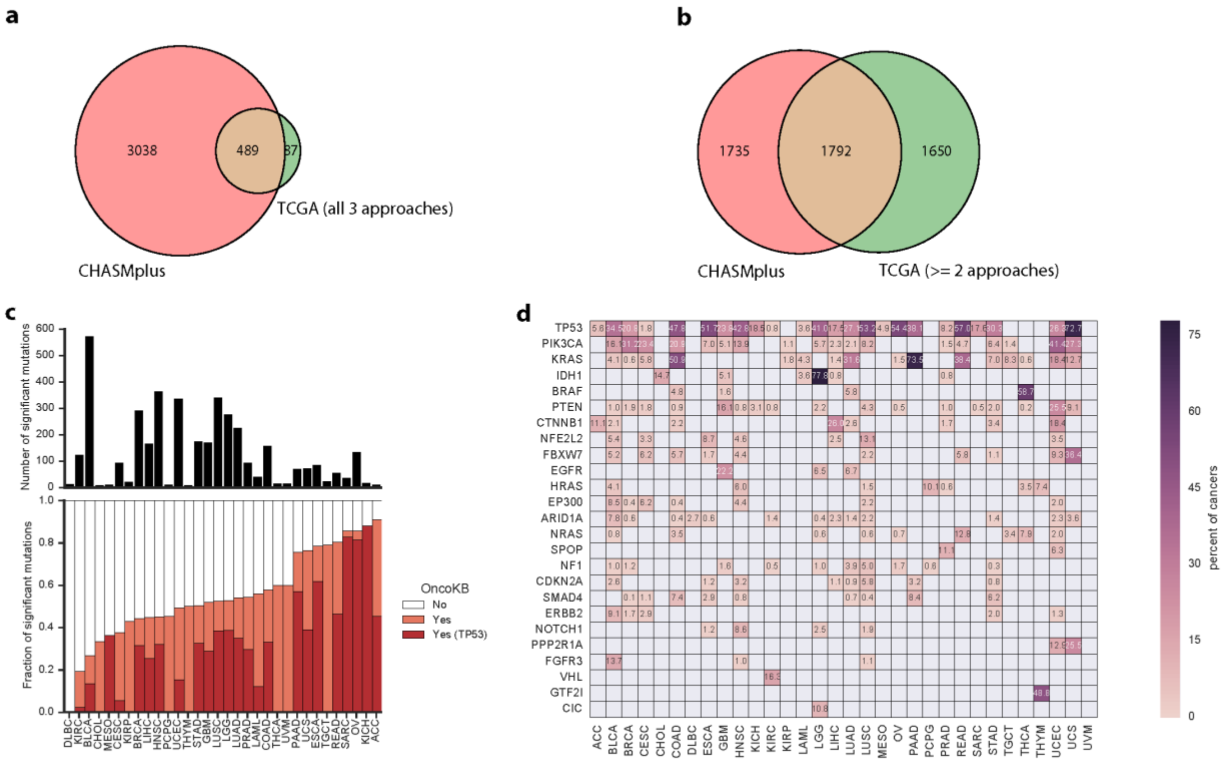
Supplemental Figure Legends



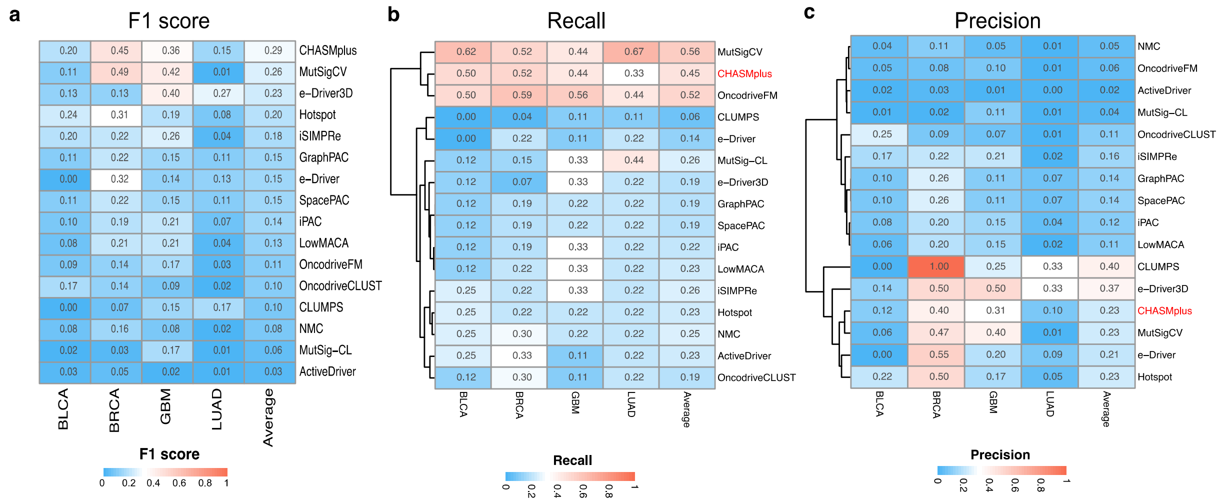
Supplementary Figure 1. Overview of CHASMplus. Related to Figure 1. a) Diagram of how CHASMplus identifies statistically significant driver somatic missense mutations in each of the 32 cancer types individually and in aggregate (pan-cancer). **b)** Diagram demonstrating how the cancer type specificity of Cancer Genome Landscape (CGL) genes were determined. **c)** Somatic missense mutations were labeled either as “likely-passenger” or “likely-driver” based on a semi-supervised approach using two steps: overlap with previously known genes from CGL in a cancer type specific manner and samples with low mutation burden. **d)** QQ plot of observed p-values for a method (blue line) compared to theoretically expected under the null hypothesis (red line). All mutations in genes found in the Cancer Gene Census were removed to eliminate possible driver mutations in this comparison. CHASMplus represents unweighted CHASMplus scores, gwCHASMplus represents gene weighted CHASMplus scores, and Hotspot is a previous codon-based mutation hotspot detection method(Chang et al., 2016).



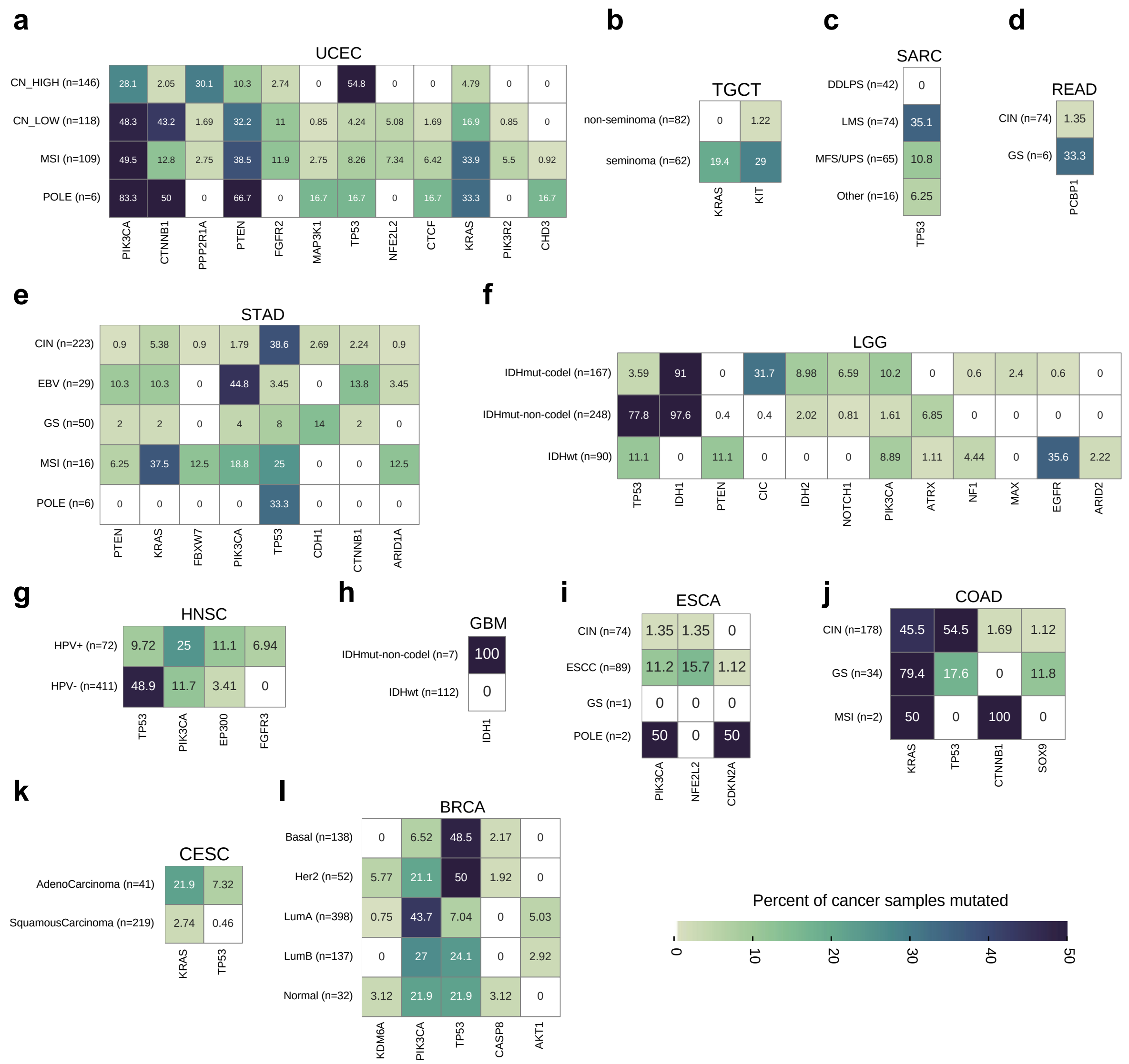
Supplementary Figure 2. Detailed mutation-level benchmark performance. Related to Figure 2. a) Heatmap of the absolute spearman correlation between methods on the TCGA mutation dataset. **b-h)** Receiver Operating Characteristic curves for, in order, the following benchmarks: Ng et al., Berger et al (EGFR resistance), TP53 transactivation (IARC database), Berget et al. (in vivo tumorplex assay), Kim et al., recurrent mutations in the Cancer Gene Census, and gene panel using OncoKB. Area under the curve is shown in parenthesis. Top 5 methods are labeled, but if the method has two version of scores then both are shown. Precision-recall curve for imbalanced benchmarks: **i)** MSK-IMPACT gene panel using OncoKB and **j)** recurrent mutations in the Cancer Gene Census (CGC-recurrent). **k)** To identify potential overfitting by methods, we repeated the Precision-Recall curve with all TP53 mutations removed. Differential performance between panels b and c is a reasonable indicator of overfitting for a method, suggesting FATHMM and CanDrA plus may have overfit to TP53. Area under the curve is shown in parenthesis. Top 5 methods are labeled according to auROC performance, but if the method has two version of scores then both are shown.



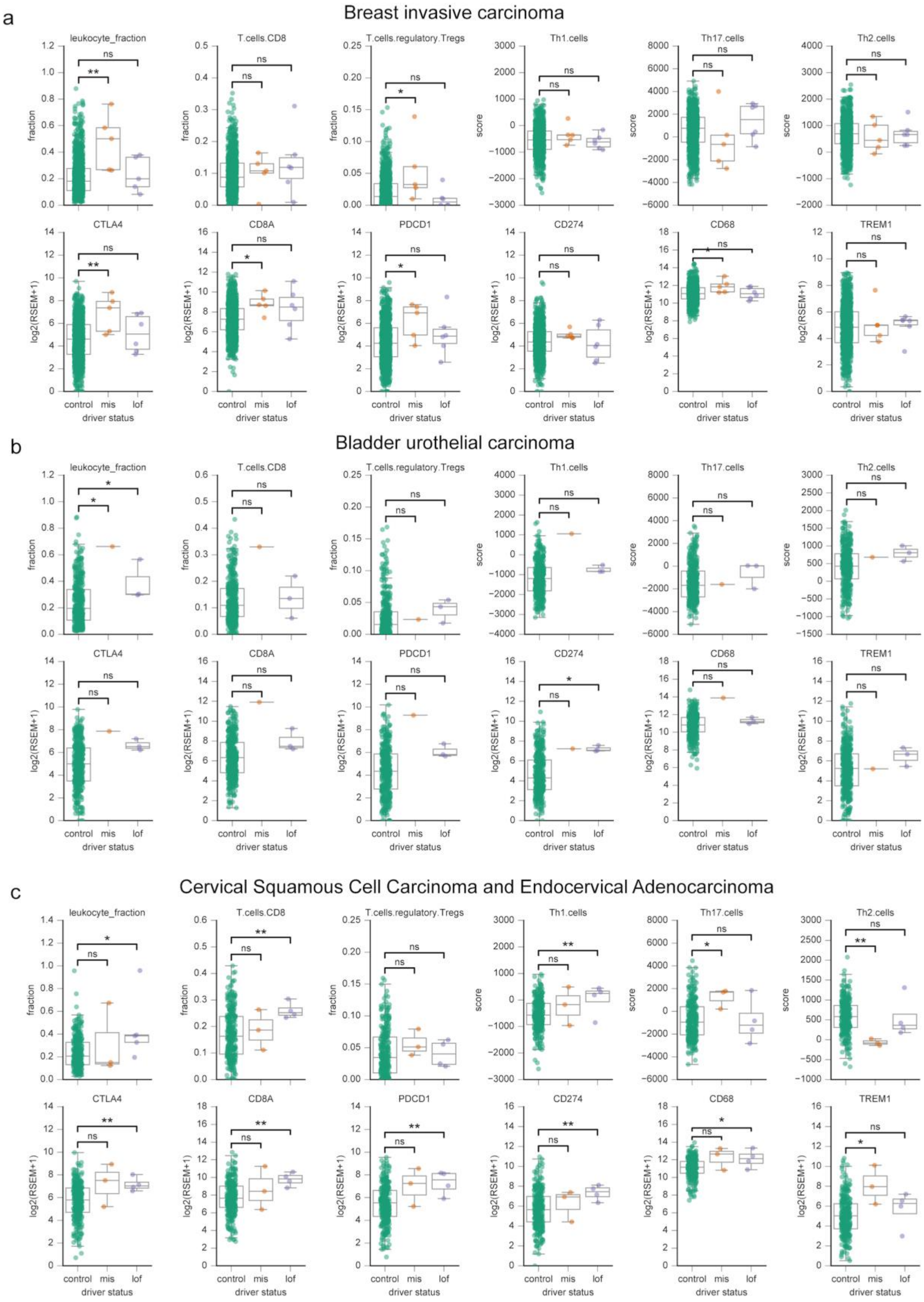
Supplementary Figure 3. Comparative analysis of driver missense mutations found by CHASMplus. Related to Figure 3. A prior TCGA analysis (Bailey et al., 2018) performed a pan-cancer analysis to identify driver missense mutations by combining three approaches, clustering of mutations in protein structure and two types of machine learning predictions. Mutations supported by more approaches had higher experimental validation rates. **a)** Nearly all consensus mutations (identified by all three approaches) in the prior study are found by pan-cancer predictions by CHASMplus. **b)** A substantial proportion of mutations identified by at least two approaches in the prior study are also identified by CHASMplus. **c)** Bar graphs showing the number of unique driver somatic missense mutations (top) and the proportion previously known in OncoKB, a literature curated database (bottom). **d)** 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.



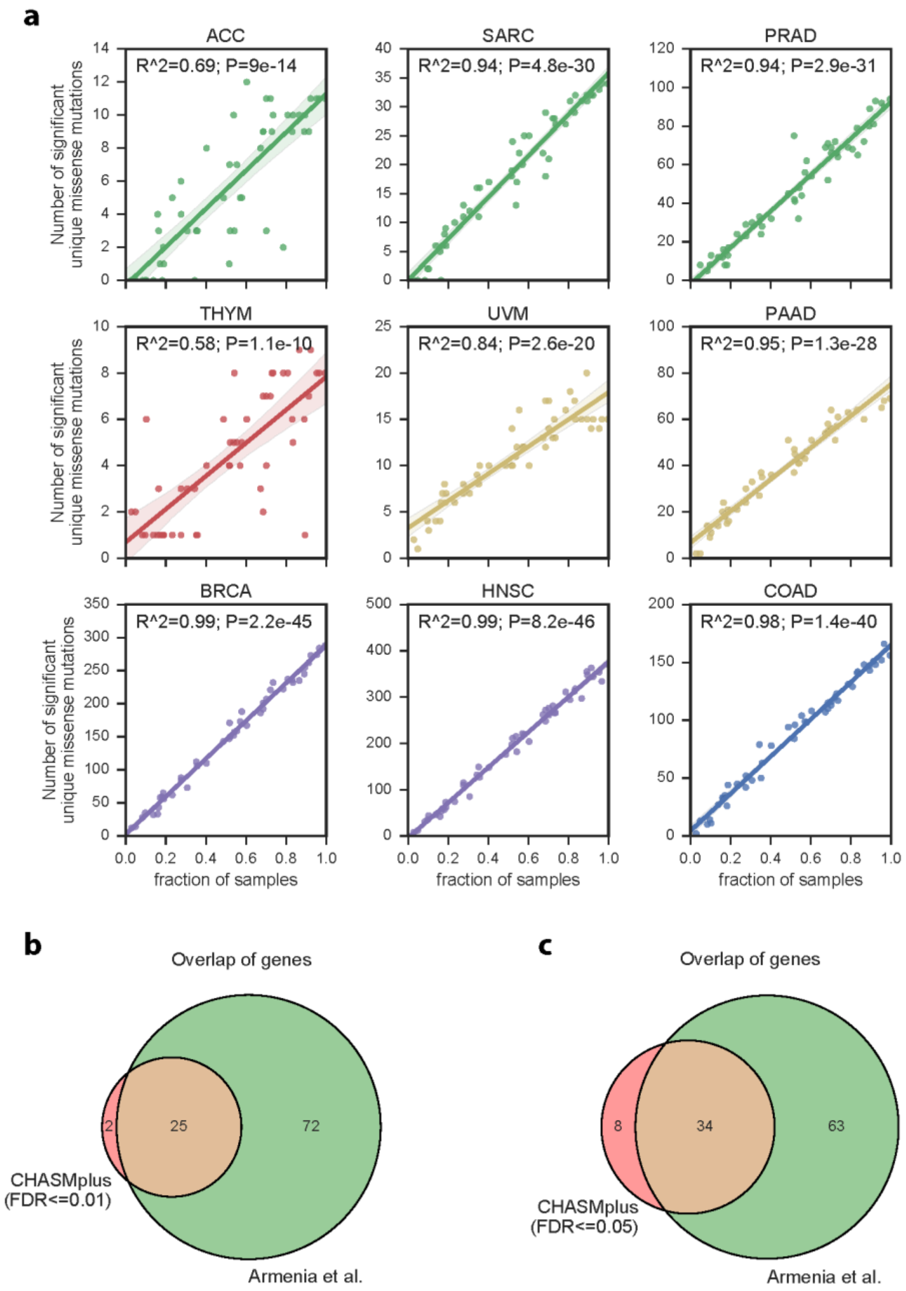
Supplementary Figure 4. Cancer type-specific benchmark on identifying cancer driver genes. Related to Figure 3. a) Heatmap showing performance (F1 score) on a Porta-pardo et al. benchmark identifying genes found in the Cancer Gene Census. The overall performance on four cancer types (BLCA, BRCA, GBM, and LUAD) is measured by the average F1 score (right column). Heatmap of **c)** recall and **d)** precision of methods in detecting genes with matching tissue type association in the cancer gene census. Overall performance is shown with average values in the right column. TCGA cancer type acronyms are listed in Methods.



Supplementary Figure 5. Subtype enrichment for driver missense mutations at the gene-level. Related to Figure 3. All genes with a statistically significant preferential enrichment for driver missense mutations in one or more cancer subtypes are shown in the form of a heatmap (q<0.1, chi-square test). Heatmaps are formatted as follows: the cancer type is noted above the heatmap, the y-axis represents cancer subtypes, and the x-axis represents genes. The percentage of samples containing driver missense mutations is indicated in each heatmap cell. **a-l)** Heatmap results, in order, for UCEC, TGCT, SARC, READ, STAD, LGG, HNSC, GBM, ESCA, COAD, CESC, and BRCA. Cancer subtype information was obtained from (Sanchez-Vega et al., 2018).


a
Breast invasive carcinoma
leukocyte_fraction
T.cells.CD8
T.cells.regulatory.Tregs
Th1.cells
Th17.cells
Th2.cells
CTLA4
CD8A
PDCD1
CD274
CD68
TREM1
fraction
score
log2(RSEM+1)
ns
*
**
control
mis
lof
driver status
b
Bladder urothelial carcinoma
leukocyte_fraction
T.cells.CD8
T.cells.regulatory.Tregs
Th1.cells
Th17.cells
Th2.cells
CTLA4
CD8A
PDCD1
CD274
CD68
TREM1
fraction
score
log2(RSEM+1)
control
mis
lof
driver status
c
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
leukocyte_fraction
T.cells.CD8
T.cells.regulatory.Tregs
Th1.cells
Th17.cells
Th2.cells
CTLA4
CD8A
PDCD1
CD274
CD68
TREM1
fraction
score
log2(RSEM+1)
control
mis
lof
driver status

Supplementary Figure 6. Correlation of putative driver mutations in CASP8 with immune phenotypes. Related to Figure 4. The CASP8 mutation status of tumor samples was compared to immune phenotypes in three cancer types: **a)** breast invasive carcinoma (BRCA), **b)** bladder urothelial carcinoma (BLCA), and **c)** cervical squamous cell carcinoma and endocervical adenocarcinoma. Control samples have no CASP8 mutations, mis indicates samples with driver missense mutations identified by CHASMplus, and lof is likely loss-of-function variants. Top row, immune cell/phenotype response inferred from DNA methylation or gene expression from (Thorsson et al., 2018). Bottom row, gene expression values from RNASeq for several important immune-related genes reported in (Thorsson et al., 2018). All TCGA cancer type acronyms are listed in Methods. *=p<0.05, **=p<0.01, and ***=p<0.001.



Supplementary Figure 7. Subsampling analysis of unique driver somatic missense mutations by CHASMplus and comparison to a larger study. Related to Figure 6. a) The number of driver somatic missense mutations identified as significant by CHASMplus (q<=0.01) as a function of sample size. CHASMplus was ran on random subsets of various sizes (fraction of samples) of the full data. To ascertain whether the trajectory suggested by subsampling analysis was consistent with what would happen with a study with greater number of tumor samples, we compared predictions to a larger prostate cancer study by Armenia and colleagues

(Armenia et al., 2018). Venn diagram of the significantly mutated genes as reported from Armenia et al. compared with genes containing a significant missense mutation predicted by **b)** CHASMplus (FDR<=0.01) or **c)** CHASMplus (FDR<=0.05).

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

For additional information regarding the data, please contact Rachel Karchin: karchin@jhu.edu.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The Cancer Genome Atlas (TCGA) collected both tumor and non-tumor biospecimens from human samples with informed consent under authorization of local institutional review boards (https://cancergenome.nih.gov/abouttcga/policies/informedconsent).

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 from whole-exome sequencing (v0.2.8, https://synapse.org/MC3)(Ellrott et al., 2018). We analyzed 32 cancer types with abbreviations for the cancer types are listed below. 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.

Cancer types in the TCGA are abbreviated as follows: Acute Myeloid Leukemia (LAML, n=139), Adrenocortical carcinoma (ACC, n=90), Bladder Urothelial Carcinoma (BLCA, n=386), Brain Lower Grade Glioma (LGG, n=510), Breast invasive carcinoma (BRCA, n=779), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, n=274), Cholangiocarcinoma (CHOL, n=34), Colon adenocarcinoma (COAD, n=230), Esophageal carcinoma (ESCA, n=172), Glioblastoma multiforme (GBM, n=311), Head and Neck squamous cell carcinoma (HNSC, n=502), Kidney Chromophobe (KICH, n=65), Kidney renal clear cell carcinoma (KIRC, n=368), Kidney renal papillary cell carcinoma (KIRP, n=275), Liver hepatocellular carcinoma (LIHC, n=354), Lung adenocarcinoma (LUAD, n=431), Lung squamous cell carcinoma (LUSC, n=464), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC, n=37), Mesothelioma (MESO, n=81), Ovarian serous cystadenocarcinoma (OV, n=408), Pancreatic adenocarcinoma (PAAD, n=155), Pheochromocytoma and Paraganglioma (PCPG, n=179), Prostate adenocarcinoma (PRAD, n=477), Rectum adenocarcinoma (READ, n=86), Sarcoma (SARC, n=204), Stomach adenocarcinoma (STAD, n=357), Testicular Germ Cell Tumors (TGCT, n=145), Thymoma (THYM, n=121), Thyroid carcinoma (THCA, n=492), Uterine Carcinosarcoma (UCS, n=55), Uterine Corpus Endometrial Carcinoma (UCEC, n=396), Uveal Melanoma (UVM, n=80), and Pan-cancer (all cancer types) (PANCAN, n=8657).

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 S1). Additional new features include features 1-10. Original features used in CHASM were obtained from an updated SNVBox MySQL database (features 11-95)(Wong et al., 2011).

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(Vogelstein et al., 2013); 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(Lawrence et al., 2014). 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 623,992 likely-passenger missense mutations, for which we found sufficient annotation to compute our selected features. Skin cutaneous melanoma mutations were not included in training due to the systematically high mutation burden for this cancer type, however, predictions for melanoma are included in Table S7.

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(Tokheim et al., 2016b). 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(Amit and Geman, 1997; Breiman, 2001), 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(Hulse et al., 2007). 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(Capriotti and Altman, 2011), 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(Tokheim et al., 2016b) 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(Altmann et al., 2010). 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(Carter et al., 2013), CADD(Kircher et al., 2014), FATHMM cancer(Shihab et al., 2013), SIFT(Ng and Henikoff, 2001), MutationAssessor(Reva et al., 2011), REVEL(Ioannidis et al., 2016), MCAP(Jagadeesh et al., 2016), ParsSNP(Kumar et al., 2016), CHASM(Carter et al., 2009), Polyphen2(Adzhubei et al., 2010), transFIC(Gonzalez-Perez et al., 2012) and CanDrA(Mao et al., 2013)). 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 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)(Chang et al., 2016), 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(Futreal et al., 2004). 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_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) (Forbes et al., 2017). 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(Forbes et al., 2017).

MSK impact gene panel and OncoKB

We obtained all missense mutations from the MSK-impact gene panel of 414 cancer-related genes(Zehir et al., 2017). Mutations were annotated against OncoKB (downloaded 4/3/2017), if the oncogenicity annotation was available for an individual mutation(Chakravarty et al., 2017). 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 (Kim et al., 2016) used a competitive screen of mutations in mice to assess the oncogenicity of mutations. The study selected mutations 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 (Petitjean et al., 2007). 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(Ng et al., 2018). 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).

Performance analysis based on area under the Precision-Recall curve

An alternative performance metric to the area under the Receiver Operating Characteristics curve (auROC) for a binary classification task is the area under the Precision-Recall curve (auPR). Like auROC, auPR summarizes the performance over all possible score thresholds from a method (Davis and Goadrich, 2006). However, auPR is preferable when there is substantial class imbalance, i.e., when the positive class of interest (in our case, cancer drivers) is the substantial minority (Saito and Rehmsmeier, 2015). The maximum auPR is 1.0 but the baseline score for a random predictor changes depending the skew of the class distribution (Saito and Rehmsmeier, 2015). Specifically, the expected auPR performance of a random baseline predictor is as follows:

$$baseline\ auPR = \frac{P}{P+N} \qquad (Equation\ 1)$$

where P is the number of samples from the positive class of interest and N is the number from the other, negative class. In contrast to the auROC, auPR will give systematically high values for all methods in benchmarks that contain a majority of positive class examples (Eq. 1). Consequently, Precision-Recall curves and auPR are a poor means of comparison for several benchmarks that we used, with auROC being a better alternative.

We therefore performed auPR analysis on the two benchmarks with a substantially under represented positive class: MSK-impact gene panel (OncoKB) and CGC-recurrent (see methods). Like auROC, the auPR from the MSK-impact gene panel benchmark indicated CHASMplus had higher performance than other methods (Figure S2i). At first glance, the auPR on the CGC-recurrent benchmark for FATHMM seemed to be higher than for CHASMplus (Figure S2j). However, the performance of FATHMM and CanDrA plus dropped substantially if TP53 mutations are not included in the benchmarks (Figure S2k). This suggests the two methods may have overfit to TP53 and seemingly do not generalize as well to other genes. CHASMplus, on other hand, maintains a high auPR, which is twice as high as the next best method.

Evaluation 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(Porta-Pardo et al., 2017). Genes were labeled by their designations in the Cancer Gene Census as a cancer driver gene for a specific cancer type. 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(Chang et al., 2016). 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(Porta-Pardo et al., 2017) (Figure S4a). 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) (Figure S4b-c). 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(Richards et al., 2015). 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.

CASP8 mutations and immune phenotypes

All values for leukocyte fraction, type of immune response (CD8 T cell, regulatory T cell, Th1 response, Th2 response, and Th17 response), and immune-related gene expression for tumor samples were obtained from (Thorsson et al., 2018). Leukocyte fraction is an estimated proportion of cells in the tumor sample that are leukocytes, as inferred from DNA methylation

(Thorsson et al., 2018). The fraction consisting of CD8 T cells or regulatory T cells was estimated using the method CIBERSORT (Newman et al., 2015). Th1, Th2, and Th17 response scores are computed from RNA-Seq gene expression using single sample Gene Set Enrichment Analysis (ssGSEA) (Hanzelmann et al., 2013). Gene expression for CTLA4, CD8A, PDCD1, CD274, TRIM1, and CD68 are quantitated from RNA-Seq using RSEM (score version 2)(Li and Dewey, 2011).

To examine the likely effect of CASP8 mutations, tumor samples were divided into control samples (no CASP8 mutations), samples with driver missense mutations predicted by CHASMplus (q<=0.01), and samples with truncating/likely loss-of-function mutations. A two-sided Mann-Whitney U test was used to determine whether CASP8 mutated samples were significantly different from control for the above immune-related phenotypes.

Comparison with PTEN saturation mutagenesis

CHASMplus was compared to two previous saturation mutagenesis studies of PTEN, examining lipid phosphatase activity (Mighell et al., 2018) and intracellular PTEN protein abundance (Matreyek et al., 2018). We first compiled a list of 6,564 missense mutations designated as high confidence for lipid phosphatase activity by Mighell and colleagues. We then merged the smaller data set of protein abundance, resulting in 3,540 missense mutations matching Mighell et al. gwCHASMplus scores were then computed for each missense mutation available from the experiments. Correlation of gwCHASMplus with lipid phosphatase activity and protein abundance was carried out using Locally Weighted Scatterplot Smoothing (LOWESS) and spearman rank correlation. Driver missense mutations identified by CHASMplus from the pan-cancer analysis were compared to all other missense mutations observed in PTEN using a two-sided Mann-Whitney U test.

Clustering of cancer types

We clustered TCGA cancer types according to two features, prevalence (fraction of samples mutated) 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.

Beyond biological differences intrinsic to each cancer type, technical difficulties in mutation calling could possibly explain the above clustering patterns. To evaluate this possibility, we correlated the mean Variant Allele Fraction (VAF) for mutations in tumor samples in each cancer type with a variety of metrics summarizing our results. VAF acts as a combined indicator of mutation sub-clonality and normal tissue contamination within the tumor sample, both of which lower the capability to detect mutations. We found no significant correlation between mean VAF for cancer types with: average number of predicted driver mutations per sample (Pearson r=0.26, p=0.14, correlation test), fraction of samples with predicted driver mutations (Pearson r=0.2, p=0.28, correlation test), unique number of significant mutations (Pearson r=0.04, p=0.82, correlation test), and normalized driver diversity (Pearson r=0.33, p=0.07, correlation test).

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.

Analysis of the tail of driver discovery for Prostate Adenocarcinoma

Here, we performed a detailed analysis of 1,013 prostate adenocarcinoma samples from the study of Joshua Armenia and colleagues(Armenia et al., 2018). Mutation calls were directly used as reported from the study. This expanded data set allowed us to assess whether the expected trajectory of discovery for prostate adenocarcinoma (see sub-sampling procedure in methods) from The Cancer Genome Atlas (TCGA) (n=477) matches what is observed from more than 1,000 samples. Additionally, since the Armenia *et al.* study only performed gene-level analysis, our analysis adds value to understanding of cancer drivers at the level of individual missense mutations in prostate adenocarcinoma (Table S6).

We found that the genes which contain significant CHASMplus missense mutations substantially overlap the significantly mutated genes found in Armenia *et al* (Figure S7b-c). We note that there are three reasons why Armenia et al. report a larger number of genes than CHASMplus: 1) they reflect genes driven by mutations other than missense mutations (e.g., nonsense, frameshift, splice site, etc.); 2) They selectively rescued previously known cancer genes seen in other types of cancers; and 3) They used a considerably laxer threshold at a false discovery rate (FDR) of 25%, while CHASMplus uses FDR of 1%.

In our TCGA analysis, we noted that the number of unique driver mutations in Prostate Adenocarcinoma linearly increased with sample size. Our original analysis identified 94 unique driver missense mutations at a false discovery rate of 1%. A strictly linear increase would expect 200 unique driver missense mutations for 1,013 samples (=94 mutations / 477 samples * 1,013 samples). Indeed, we find 203 unique driver missense mutations from the data in Armenia *et al.*, strikingly close to our expectations.

We next examined the prevalence of driver missense mutations with this expanded data set compared to that from the TCGA. To compare the two, we used the average number of driver missense mutations per sample. Like in the TCGA analysis, a 1% false discovery rate threshold was applied. We found the driver prevalence from Armenia *et al.* data was 0.40 mutations per sample compared with 0.31 mutations per sample from the TCGA data. As expected, driver prevalence increased gradually but with a rate showing diminishing returns. A strictly linear increase would have yielded 0.66 mutations per sample. Next, we looked at mutations that were marginally significant at a false discovery rate (FDR) of 5%. At 5% FDR, the driver prevalence was 0.48 mutations per sample. So, despite the diminishing returns for discovery, there were a reasonable prevalence of marginally significant driver missense mutations, suggesting that discovery is not completely saturated.

Limited power for mutation hotspot detection approaches

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 Figure S7a, 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 (see Statistical Power Analysis). For example, the TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) cohort has 274 samples and a background mutation rate of 3.5 mutations/Mb. This sample size is sufficient to detect driver mutations that occur in ~2% of the samples with 90% power.

At current TCGA sample sizes, 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 (Figure 7a). 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(Chang et al., 2016) (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 potential 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) (Figure 7b). 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 (Figure 1d). Importantly, our increased sensitivity did not come at the cost of low specificity, as evidenced by our p-value calibration (Figure S1d) and extensive ROC analysis across seven benchmarked datasets (Figure 2b), which measures a balance of sensitivity and specificity.

Statistical power methodology

We estimated the statistical power to find frequently mutated codons within driver genes by using a binomial model, as done previously(Tokheim et al., 2016b). 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(Lawrence et al., 2014), 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 (Tokheim et al., 2016b).