Evaluating the evaluation of cancer driver genes

Collin J. Tokheim[a,b], Nickolas Papadopoulos[c,d], Kenneth W. Kinzler[c,d], Bert Vogelstein[c,d,1], and Rachel Karchin[a,b,e,1]

[a]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218; [b]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218; [c]Ludwig Center, Johns Hopkins Medical Institutions, Baltimore, MD 21231; [d]Howard Hughes Medical Institute, Johns Hopkins Medical Institutions, Baltimore, MD 21231; and [e]Cancer Biology Program, Department of Oncology, Johns Hopkins Medical Institutions, Baltimore, MD 21231

Contributed by Bert Vogelstein, October 17, 2016 (sent for review May 31, 2016; reviewed by Kyle Covington, Elaine R. Mardis, and Peter J. Park)

Sequencing has identified millions of somatic mutations in human cancers, but distinguishing cancer driver genes remains a major challenge. Numerous methods have been developed to identify driver genes, but evaluation of the performance of these methods is hindered by the lack of a gold standard, that is, bona fide driver gene mutations. Here, we establish an evaluation framework that can be applied to driver gene prediction methods. We used this framework to compare the performance of eight such methods. One of these methods, described here, incorporated a machine-learning–based ratiometric approach. We show that the driver genes predicted by each of the eight methods vary widely. Moreover, the *P* values reported by several of the methods were inconsistent with the uniform values expected, thus calling into question the assumptions that were used to generate them. Finally, we evaluated the potential effects of unexplained variability in mutation rates on false-positive driver gene predictions. Our analysis points to the strengths and weaknesses of each of the currently available methods and offers guidance for improving them in the future.

cancer genomics | DNA sequencing | driver genes | cancer mutations | computational method evaluation

The search for genetic drivers of cancer has rapidly progressed with systematic exome-sequencing studies (1). A major goal of these studies is to identify signals of positive selection and distinguish them from passenger mutations. A cancer driver gene is defined as one whose mutations increase net cell growth under the specific microenvironmental conditions that exist in the cell in vivo. The total number of driver genes is unknown, but we assume that is considerably less than 19,000. At present, the only way to assess the evidence for a gene being a driver gene in vivo in humans is through evaluation of the mutations present in clonal expansions of tumor cells. Because passenger gene mutations will also be fixed in any cell that expands due to a driver gene mutation, statistical evaluation in patient cohorts is required to distinguish the two. The first exomic analyses attempted to identify candidate driver genes as those having more mutations than expected from some presumed background somatic mutation rate, corrected for base context, gene size, and other variables (2, 3). Subsequent work has considerably refined the variables involved in determining whether a gene is more mutated in cancers than expected by chance. This has led to a variety of "significantly mutated gene" methods that adjust for covariates such as replication timing and gene expression as well as including more sophisticated metrics of mutational base contexts (4, 5).

An alternative approach to finding cancer drivers employs ratiometric methods. Rather than attempting to determine whether the observed mutation rate of a gene in cancers is higher than expected by chance, these methods simply assess the composition of mutations normalized by the total mutations in a gene. The ratiometric 20/20 rule (6) evaluates the proportion of inactivating mutation and recurrent missense mutations in a gene of interest. Other ratiometric approaches use mutation functional impact bias (7), mutational clustering patterns (8, 9), or mutation composition patterns (9). Here, we describe a machine-learning–based, ratiometric method (20/20+) that formalizes and extends the original 20/20 rule and enables automated integration of multiple features of positive selection.

Rigorous and unbiased evaluation is necessary to inform users about the comparative utility of prediction methods, including the method described herein. In many investigative domains, there is a generally accepted gold standard against which predictions can be benchmarked. However, only a limited number of genes have been fully vetted as cancer drivers. In previous work, driver prediction has been benchmarked by significant overlap with the Cancer Gene Census (CGC) (10), which is a manually curated list of likely but not necessarily validated driver genes (7, 8) or by agreement with a consensus gene list of drivers predicted by multiple methods (11). To our knowledge, a systematic framework for the evaluation of somatic mutations that can be generally applied has not been previously developed.

Significance

Modern large-scale sequencing of human cancers seeks to comprehensively discover mutated genes that confer a selective advantage to cancer cells. Key to this effort has been development of computational algorithms to find genes that drive cancer based on their patterns of mutation in large patient cohorts. Because there is no generally accepted gold standard of driver genes, it has been difficult to quantitatively compare these methods. We present a machine-learning–based method for driver gene prediction and a protocol to evaluate and compare prediction methods. Our results suggest that most current methods do not adequately account for heterogeneity in the number of mutations expected by chance and consequently yield many false-positive calls, particularly in cancers with high mutation rate.

Author contributions: C.J.T., N.P., K.W.K., B.V., and R.K. designed research; C.J.T. performed research; C.J.T. contributed new reagents/analytic tools; C.J.T. analyzed data; and C.J.T., B.V., and R.K. wrote the paper.

Reviewers: K.C., Human Genome Sequencing Center, Baylor College of Medicine; E.R.M., Washington University School of Medicine; and P.J.P., Harvard Medical School.

Conflict of interest statement: B.V. is a founder of PapGene and Personal Genome Diagnostics and a member of the Scientific Advisory Boards of Morphotek, Syxmex-Inostics, and Exelixis GP. The first four of these companies, as well as other companies, have licensed technologies from Johns Hopkins University, on which B.V. is an inventor. These licenses and relationships are associated with equity or royalty payments to B.V. The terms of these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies. K.W.K. is a founder of PapGene and Personal Genome Diagnostics and a member of the Scientific Advisory Boards of Morphotek and Syxmex-Inostics. These companies, as well as other companies, have licensed technologies from Johns Hopkins University, on which K.W.K. is an inventor. These licenses and relationships are associated with equity or royalty payments to K.W.K. The terms of these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies. N.P. is a founder of PapGene and Personal Genome Diagnostics. These companies, as well as other companies, have licensed technologies from Johns Hopkins University, on which N.P. is an inventor. These licenses and relationships are associated with equity or royalty payments to N.P. The terms of these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: bertvog@gmail.com or karchin@jhu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1616440113/-/DCSupplemental.

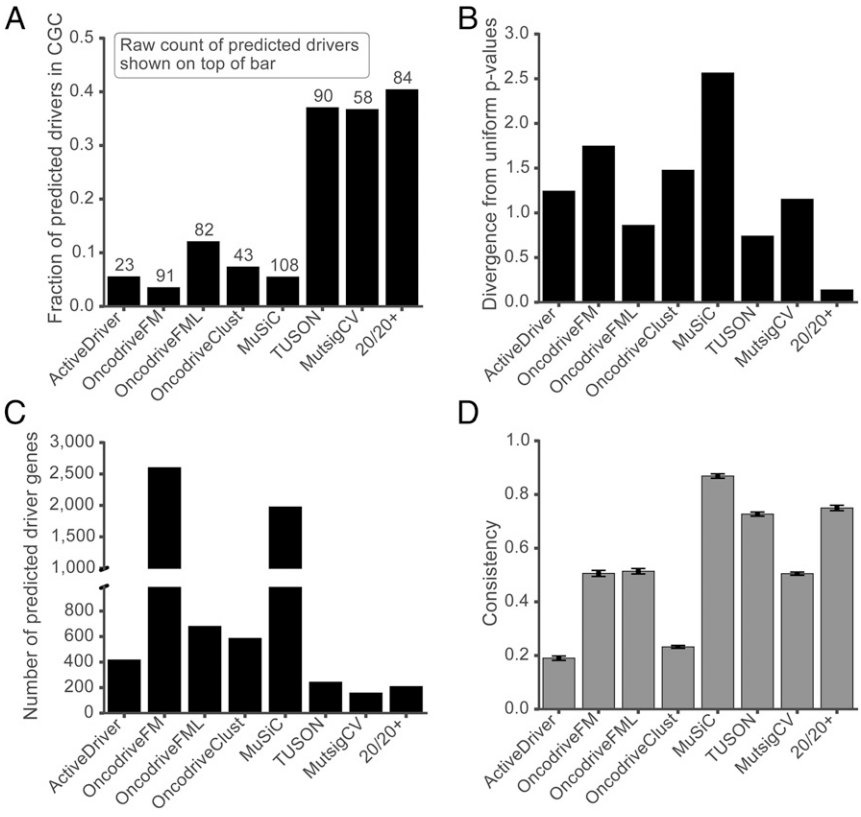

Fig. 1. Outputs of eight driver prediction methods run through the evaluation protocol. (*A*) Fraction of predicted driver genes ($q \leq 0.1$) that are found in the Cancer Gene Census (CGC) (downloaded April 1, 2016). Raw count of predicted driver genes indicated on *Top* of each bar. (*B*) Divergence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

In this work, we present a framework for such evaluations. The framework has five components, some of which have been previously applied in isolation, but not as part of a unified system. We considered overlap with CGC, agreement between methods, comparison of observed vs. theoretical *P* values, number of significant genes predicted, and prediction consistency on independent partitions of the dataset. To implement this framework, we first collected 729,205 published somatic mutations from 34 cancer types (9, 12) (Fig. S1). These mutations were composed of single base substitutions and in-frame and out-of-frame insertions and deletions (indels) of less than 10 bp. We then compared various methods on the full pancancer set and on four selected cancer types with diverse mutation rates—pancreatic adenocarcinoma (PDAC), breast adenocarcinoma (BRAC), head and neck squamous cell carcinoma (HNSCC), and lung adenocarcinoma (LUAD).

Results

Overlap of the Driver Genes Predicted by Each Method. Eight methods were evaluated: MutsigCV (12), ActiveDriver (13), MuSiC (5), OncodriveClust (8), OncodriveFM (7), OncodriveFML (14), Tumor Suppressor and Oncogenes (TUSON) (9), and 20/20+ (https://github.com/KarchinLab/2020plus). All data, on all cancers, was considered ("pancancer") in these comparisons. First, we assessed overlap of the predicted driver genes with the CGC. We considered only those CGC genes typed as somatic, missense, frameshift, nonsense or splice site, excluding translocations, large amplifications/deletions, and other mutation consequence types not addressed in our study, yielding a total of 188 CGC genes. Although the driver genes predicted by all methods were enriched for CGC genes, the predicted drivers by any individual method did not contain a majority of CGC genes (Dataset S1 and Fig. 1*A*). Three methods (20/20+, MutsigCV, and TUSON) had substantially higher fractions of predicted drivers in the CGC than the other methods. When we considered a subset of 99 CGC genes supported by functional studies (15), the results were very similar. The ranking of methods by fraction predicted was essentially the same as with the full CGC, with the three methods listed above having substantially higher fractions than the rest (Fig. S2).

Genes predicted by more than one method may be more likely to be drivers (11). For each method, we calculated the fraction of predicted drivers that were unique or predicted by at least one, two, or three other methods (Fig. S3 and Dataset S2). As shown in Fig. S3, there was little consensus in prediction of driver genes among the methods. The majority (59–80%) of genes identified by MuSiC, ActiveDriver, OncodriveClust, OncodriveFML, or OncodriveFM were not observed by any of the other seven methods. The fractions of genes identified by TUSON, 20/20+, and MutsigCV that were not identified as driver genes in at least one of the other seven methods was 14%, 19%, and 33%, respectively. Although it is likely that some of the uniquely predicted drivers are bona fide, we could not find convincing literature support for the top-ranked unique predictions of MuSiC, ActiveDriver, and the Oncodrive methods (Dataset S3). A consensus list of drivers predicted by TUSON, 20/20+, or MutsigCV appears in Table S1.

Observed vs. Expected *P* Values. Given the lack of agreement among these various methods, we compared *P* values reported by each method to those expected theoretically. Such comparisons are often used in statistics and can indicate invalid assumptions or inappropriate heuristics. Theoretically, the *P* value distribution should be approximately uniform after likely driver genes are removed (16). Therefore, we removed all genes predicted to be drivers by at least three methods after Benjamini–Hochberg multiple-testing correction ($q \leq 0.1$) and any remaining genes in the CGC. We assumed that the number of bona fide driver genes not removed by this procedure would be small enough to have minimal impact on the *P* value distribution. To quantify the differences between the observed *P* values and those expected from a uniform distribution, we developed a measure named

mean absolute $\log_2$ fold change (MLFC) (*SI Materials and Methods*). MLFC values near zero represent the smallest discrepancies and the closest agreement between observed and theoretical *P* values.

One method (20/20+) had an MLFC that was fivefold lower than the seven others (Fig. 1*B*). We also compared observed and theoretical *P* values with quantile–quantile plots, which provide a detailed view of *P* value behavior (Fig. S4*A*). 20/20+ *P* values had by far the best agreement with theoretical expectation across the entire range of supported values. In the critical range typically used to assess statistical significance ($P \leq 0.05$), OncodriveClust, OncodriveFM, OncodriveFML, ActiveDriver, and MuSiC substantially underestimated *P* values, whereas MutsigCV substantially overestimated them (Fig. S4*B*). For methods that combine multiple *P* values for each gene, failure to model correlation between *P* values may be responsible for this underestimation. The null *P* value distributions at the other end of the distribution (0.2–1.0) should also be uniform and in this case independent of the actual number of true driver genes.

Number of Predicted Driver Genes. The number of predicted driver genes ($q \leq 0.1$) ranged from 158 (MutsigCV) to 2,600 (OncodriveFM) (Fig. 1*C*). There were two obvious categories of methods with respect to predicted driver genes: MutSigCV, 20/20+, and TUSON predicted 158–243 genes (Table S1), whereas the remaining had over 400 driver genes.

Driver Gene Prediction Consistency. Statistical methods suffer from both systematic and random prediction errors. When no gold standard is available, it is difficult to estimate systematic error, but possible to estimate random error by measuring the variability of predictions. We tested the eight methods on 10 repetitions of a random two-way split of the all samples in our dataset, while maintaining the proportion of samples in each cancer type. An ideal method would produce the same list of driver genes, ranked by *P* value, for each half of the split. For a fair comparison, we considered that methods predicting many drivers would be less likely to have consistent rankings than those predicting only a few. Thus, we developed a measure named TopDrop consistency (TDC) (*SI Materials and Methods*) that examines the overlap between genes ranked at a defined depth (e.g., the top 100 genes) for each half of the random split. Examining TDC at a depth of 100 genes showed MuSiC, 20/20+, and TUSON to be the three with the highest consistency (Fig. 1*D*). Most methods decreased in consistency when the gene depth was varied between 20 and 300, but the ordering of the TDC scores among the eight methods remained relatively stable (Fig. S5).

Overall Performance. In Table 1, we summarize the performance of each method according to the criteria described above on the pancancer mutation data. The overall protocol is shown as a flowchart in Fig. S6. We assume that a preferable method would predict a higher fraction of driver genes that overlap with the CGC, that overlap with at least one other method, that have the least deviation from expected null *P* values, and that have the highest consistency. Each method is accordingly ranked by these four criteria and the average rank is shown. The top ranked methods are 20/20+, TUSON, OncodriveFML, and MutsigCV.

Evaluation of Specific Cancer Types. To evaluate whether methods performed differently on specific cancer types, rather than on the pancancer dataset, we repeated our evaluations using four specific cancer types. We considered two moderate mutation rate cancers (PDACs, with median of 0.7 mutations per MB, BRACs, with median of 1 mutation per MB), and two high-mutation-rate cancers (head and neck squamous carcinomas, with median of 3.2 mutations per MB, and LUADs, with median of 6.7 mutations per MB). These mutation rates were based on the same dataset used throughout this study and described above. As with the pancancer types, the 20/20+ method had the least discrepancy between observed and theoretical *P* values (Fig. S7*A*). For the TDC evaluation, we used a rank depth of 10 genes rather than the 100 used for pancancer, as it is likely that this number is closer to the number of driver genes in a single cancer type. The most consistent methods were 20/20+ and MuSiC (Fig. S7*C*), and the least consistent methods were ActiveDriver and OncodriveClust. Driver genes in HNSCC were the most consistently predicted overall.

The number of cancer-specific predicted driver genes varied widely (Fig. S7*B*). PDAC had the fewest predicted drivers ($q \leq 0.1$), ranging from 3 (TUSON) to 49 (OncodriveFM), whereas LUAD had the most, ranging from 9 (TUSON) to 922 (OncodriveFM). ActiveDriver, OncodriveFM, and OncodriveClust predicted hundreds of cancer type-specific drivers, whereas OncodriveFML, MuSiC, 20/20+, and MutsigCV tended to predict no more than 40 in any of the four cancer types.

Variability in Background Mutation Rate. Because only a small fraction of the total somatic mutations in any common solid tumor affects driver genes, the remaining mutations can be considered passengers. These reflect the mutation "background," that is, all mutations that occurred during the divisions of the cells that eventually formed the tumor, from embryogenesis until the tumor was surgically removed (17). The total number of mutations (drivers plus passengers) is therefore only slightly larger than the number of passenger mutations, and, for simplicity, we refer to this number as the background mutation rate. The median background mutation rate for cancer types in our pancancer set varied over two orders of magnitude (Fig. S8), with individual samples varying over an even larger range. Mutation rates vary among individual genes and are influenced by nucleotide context, environmental factors, gene expression, chromatin state, replication timing, DNA repair activity, strand, and

Table 1. Performance of eight evaluated cancer driver gene prediction methods on the pancancer dataset of small somatic mutations

Method	No. significant genes	CGC overlap	Method consensus	*P* value deviation	Consistency	CGC rank	Consensus rank	*P* value rank	Consistency rank	Average rank
2020+	208	0.40	0.808	0.139	0.749	1	2	1	2	1.5
TUSON	243	0.37	0.856	0.740	0.727	2	1	2	3	2
OncodriveFML	679	0.12	0.408	0.860	0.514	4	4	3	4	3.75
MutsigCV	158	0.37	0.671	1.153	0.505	3	3	4	6	4
OncodriveClust	586	0.07	0.336	1.477	0.232	5	5	6	7	5.75
MuSiC	1,975	0.05	0.199	2.564	0.869	7	8	8	1	6
ActiveDriver	417	0.06	0.242	1.243	0.19	6	7	5	8	6.5
OncodriveFM	2,600	0.04	0.315	1.747	0.506	8	6	7	5	6.5

The number of significant genes is reported using a threshold of $q \leq 0.1$. Each method is ranked by overlap fraction with the CGC, with the other methods listed (consensus), MLFC of *P* values (deviation from null uniform distribution), and consistency in the TopDrop 100 test.

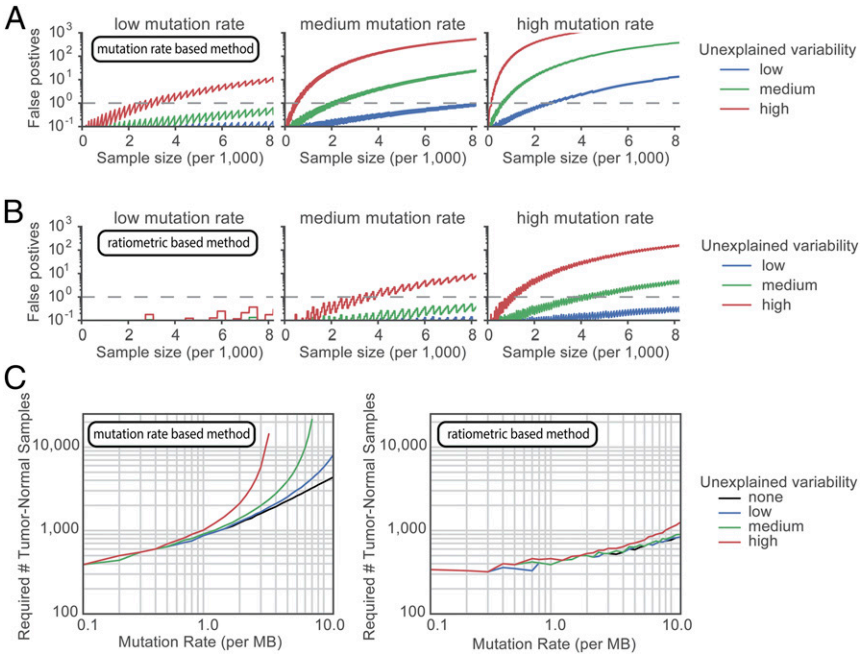

Fig. 2. Models of mutation rate-based and ratiometric-based methods suggest decrease in false positives and increased power with ratiometric approach. (*A*) Expected false positives for a mutation rate-based predictor that identifies genes with increased mutation rate over background. (*B*) Expected false positives for a ratiometric predictor that identifies genes with increased inactivating mutation fraction over background. For both *A* and *B*, we assume there is unexplained variability in either background mutation rate or inactivating mutation fraction that is not accounted for in driver gene prediction. False positives are shown as a function of sample size (up to 8,000 paired tumor–normal samples) for low (0.5 mutations per MB), medium (3.0 mutations per MB), and high (10.0 mutations per MB) background mutation rates and low (blue), medium (green), and high (red) unexplained variability (CVs of 0.05, 0.1, and 0.2, respectively). The dashed line indicates one expected false positive. For the mutation rate-based method, the number of false positives increases to undesirable levels for high mutation rates, particularly when there is high unexplained variability. (*C*) Sample size required for near-comprehensive detection of intermediate-effect driver genes (90% detection and 2% effect size/increase with respect to background). Results are shown for scenarios with no unexplained variability (black), low (blue), medium (green), and high (red) unexplained variability (CVs of 0.0, 0.05, 0.1, and 0.2, respectively). The number of required samples for the mutation rate-based method becomes very large for moderate-to-high mutation rates and levels of unexplained variability, but it is considerably lower for the ratiometric method. MB, megabase. The jagged behavior of the curve in *A* and *B* is due to the discrete nature of our data.

perhaps by a variety of factors that have yet to be discovered (4, 18, 19).

We analyzed the possible impact of unexplained variability in background mutation rate on expected false-positive driver gene predictions. First, we applied a binomial model previously used for driver gene detection power analysis (12). The model assumes a gene-specific background mutation rate μ, which is set to a relatively high value, corresponding to genes in the 90th percentile of mutation rate. We used the binomial to set a critical value for driver gene prediction, that is, the number of mutations required for a gene to be considered significantly different from the background. Next, we modeled the situation where the genes actually had mutation rates that varied around μ, using a beta-binomial model. We estimated the false positives expected under the binomial, after a highly conservative multiple-testing correction (Bonferroni). The number of mutations required to meet or exceed the binomial critical value was compared with that of the beta-binomial, for different background mutation rates, for levels of variability [beta-binomial coefficients of variation (CVs)], and for sample size ranging up to 8,000 (Fig. 2*A*). Levels of variability defined by CVs (CV = 0.05, 0.1, and 0.2) were chosen to approximate low, medium, and high unexplained variation around the mean. As the number of samples increased, so did the number of expected false positives. At the low end of background mutation rates (0.5 mutations per MB), the expected false positives remained low, even when 8,000 samples were evaluated, regardless of the level of variability. At an intermediate background mutation rate of 3.0 mutations per MB and with high unexplained variability, ∼1,000 false positives were expected from 8,000 samples. At a high background mutation rate (10.0 mutations per MB), both medium and high unexplained variability produced many thousand expected false positives.

We reasoned that unexplained variability might also have an impact on power calculations to estimate how many samples must be sequenced to find the majority of cancer driver genes. To this end, we repeated previous calculations performed with a binomial power model, in which the required sample size was estimated to be 600–5,000 per cancer type (12). The original model was parameterized to detect intermediate frequency driver genes, having 2–20% mutation rates above background per sample, with background defined by genes in the 90th percentile of background mutation rates. First, we calculated the sample size required to detect 90% of these drivers, given exome-wide backgrounds of 0.1–10 mutations per MB, and a conservative estimate of 2% effect size (*SI Materials and Methods*, *Unexplained Variability Affects Power and False Positives*). Next, we calculated the sample size required if the gene mutation rate varied around the original estimate, using a beta-binomial model with different CVs (CV = 0.05, 0.1, 0.2). The binomial power model was in accord with previous estimates. However, when unexplained variability was taken into account, the number of required samples increased sharply, particularly for higher background mutation rates (Fig. 2*C*, *Left*).

Variability in Ratiometric Features. Ratiometric features are expected to have significantly less variability among cancer types than background mutation rates. Fig. S8 shows the variability of the median ratio of nonsilent to silent mutations for cancer types in our pancancer set. The variability of this ratiometric feature among tumor types is miniscule compared with that of mutation rates in the same tumor types. However, ratiometric features might also be sensitive to unexplained variability in their background distributions. Therefore, we calculated the expected false positives and statistical power of a simple ratiometric feature, the fraction of mutations in a gene having a specific nonsilent mutation consequence type. A slightly modified version of the calculations used for background mutation rate was applied (*SI Materials and Methods*, *Unexplained Variability Affects Power and False Positives*). We observed improved false-positive control (Fig. 2*B*), and reduction in the number of required samples (Fig. 2*C*, *Right*), particularly for high-mutation-rate cancers.

Discussion

A major goal of the huge public investment in large-scale cancer sequencing has been to find driver genes. Robust computational prediction of drivers from small numbers of somatic variants is

critical to this mission, and it is essential that the best methods for this purpose be identified. Although many such methods have been proposed (*SI Materials and Methods*, *Evaluated Driver Gene Prediction Methods*), it has been difficult to evaluate them because there is no gold standard to use as a benchmark. Here, we developed an evaluation framework for driver gene prediction methods that does not require a gold standard. The framework includes a large set of small somatic mutations from a wide range of cancer types and five evaluation metrics. It can be used to systematically evaluate new prediction methods and compare them to existing methods. The results would be more informative to users of these methods than current ad hoc approaches.

To apply the framework to a new method, a ranked list of predicted driver genes can be generated from the pancancer mutation dataset (Dataset S4 and Figs. S1 and S6), including a P value and a Benjamini–Hochberg corrected q value for each gene. The choice of a threshold $q \leq 0.1$ to define driver genes worked well in our evaluations but can be adjusted if so desired. The same threshold should be used for fair comparison of different methods. If a driver prediction tool does not produce P values, a raw score threshold that represents the desired false-discovery rate could be selected.

By calculating the overlap fraction of predicted drivers with both the CGC and the eight methods evaluated here, it is possible to quickly determine whether a new method is on the right track. Two baselines of good performance are a method's ability to recapitulate many of the well-studied cancer genes in CGC and ability to identify a core set of genes that are predicted as drivers by several other methods. The methods with strongest support by these criteria were 20/20+, TUSON, and MutsigCV. Roughly 40% of predicted drivers by these methods were in CGC, contrasted with roughly 10% of predicted drivers by the remaining methods. They also had substantially more overlap with other methods and predicted the smallest fraction of unique genes, those having no overlap with other methods (Table 1 and Fig. S3). Although detecting some unique genes is desirable for purposes of discovering novel drivers, if the fraction of unique predictions is much greater than one-half, a method may be prone to false positives. Comparing the gene P value distribution of a method of interest with theoretically expected P values and with the total number of predicted driver genes may help make sense of such a result.

The TDC metric (TDC 100 and TDC 10) evaluates the stability of the top k genes in a ranked list, when a method is applied repeatedly to matched random partitions of a dataset. Each method produces its own ranked list, and therefore a potentially different set of top k genes. For the pancancer dataset, we considered the top 100 genes predicted by each method, and for cancer-specific datasets, we considered the top 10 genes predicted by each method. For pancancer, the most consistent methods had a high average fraction of their top 100 genes consistently ranked between different partitions: MuSiC (87 of their top 100 genes), 20/20+ (75 of their top 100 genes), and TUSON (73 of their top 100 genes) (Fig. 1*D* and Table 1). These three methods also consistently ranked an average of 8 or 9 of their top 10 genes in cancer-specific datasets of BRAC, LUAD, and HNSCC (except for TUSON on LUAD) (Fig. S7). We suggest that the ability to reproduce approximately three-quarters of the same top-ranked genes is a reasonable baseline standard.

Most commonly used driver gene prediction methods produce a P value for each gene. These probabilistic scores have utility for end users, because they provide a means to separate driver and passenger genes by a threshold selected according to a user's tolerance for false discoveries. If properly calibrated, P values can help eliminate arbitrary decisions about where to set a score threshold when one does not know in advance how exactly many driver genes are expected. However, poorly distributed P values can inflate false positives or reduce sensitivity, and they are an indication that a tool may be making inappropriate assumptions. Based on the assumption that the total number of cancer driver genes is small compared with the total number of human genes (6), it is reasonable to assume that P values assigned to all genes should be approximately uniform after a core set of well-established driver genes have been dropped.

The MLFC metric proposed here quantifies the deviation between these theoretically expected P values and the observed P values generated by a method. Lower values of MLFC are desired; 20/20+ had the lowest MLFC of methods evaluated, both for pancancer and for all four cancer types evaluated (Fig. 1*B*, Table 1, and Fig. S7). The highest MLFC values were for MuSiC, OncodriveClust, and ActiveDriver. More detailed P value behavior can be seen in quantile–quantile plots that compare the observed P values with a theoretical uniform distribution (Fig. S4 *A* and *B*). The plots show that MuSiC, ActiveDriver, OncodriveClust, OncodriveFM, and OncodriveFML P values are underestimated (lower than theoretically expected) in the critical range (P values from 0 to 0.05), whereas the P values of TUSON and MutsigCV are overestimated. The low end of the P value range is critical because it corresponds to the threshold used to call driver genes ($q = 0.1$ in this work). If P values are underestimated in this range, too many genes will be called as drivers. In fact, the methods that underestimate P values predict the largest number of drivers and have the highest fraction of uniquely predicted drivers (Fig. 1*C*, Table 1, and Fig. S3).

These results suggest that the true number of driver genes represented in the pancancer dataset is closer to the number predicted by 20/20+ (208) and TUSON (243), which have the lowest MLFC. If a new method is tested and reports a substantially larger number of drivers than these two methods, the MLFC and quantile–quantile plots should be carefully checked for discrepancies with theoretically expected P values. Once a sufficient number of tumors have been sequenced, analysis of an individual tumor type can provide more relevant information about driver genes in that tumor type than can a pancancer analysis. For example, if a driver gene is only important for a particular tumor type or subtype, then the mutations observed in other tumor types contribute only noise. The reliability of the analytic method is particularly important when evaluating driver genes in this situation because the numbers of tumors and mutations will be relatively small.

The MFLC also has substantial implications for the accuracy of driver gene prediction methods. The relatively high MFLC of several methods brings into question the validity of the assumptions or analytic methods used in their construction. We believe that the most likely problem is with the assumptions rather than the analytic methods, which all appear to be well thought-out. In addition, the most likely problem with the assumptions is that there is unexplained variability in the background mutation rates. This variability may be tumor type specific or even patient or tumor specific. In an effort to understand the potential effects of such unexplained variability on cancer gene driver predictions, we modeled the situation encountered when various numbers of mutations were available for evaluation. The results were striking in that high levels of variability produced huge numbers of false positives when background mutations rates were high. Thus, cancers with high background mutation rates (such as those associated with environmental carcinogens) are the most problematic for driver prediction methods. This analysis also demonstrated how unexplained variability in gene mutation rates might confound power calculations. Estimates of near-saturation discovery of drivers using ~5,000 samples in high-mutation-rate cancers, such as melanoma (12), may be overly optimistic, and such discovery may require more resources than currently projected. Identifying the optimum driver gene prediction methods will be an important part of this effort, and we hope that the

evaluation protocol described here will help to test and improve those methods.

Materials and Methods

Data Collection. The pancancer dataset consists of 729,205 small somatic variants encompassing 7,916 distinct samples from 34 specific cancer types by merging data in published whole-exome or whole-genome sequencing studies used by TUSON (elledgelab.med.harvard.edu/wp-content/uploads/2013/11/Mutation_Dataset.txt.zip) (9) and Mutsig (www.tumorportal.org/load/data/per_ttype_mafs/PanCan.maf) (12) and removing duplicate samples in both studies. Any studies that did not report silent mutations were removed. Data in refs. 9 and 12 originated from The Cancer Genome Atlas, International Cancer Genome Consortium, the Catalogue of Somatic Mutations in Cancer database (20), and dbGAP (21). We did not see evidence of batch effects by data source in the number of variants per tumor type, single-nucleotide mutation spectra, or specific mutation consequence types. We further applied quality control to this data by filtering out hypermutated samples (>1,000 intragenic small somatic variants) (6), and regions prone to mutation calling artifacts [any sequencing read mappability warning cataloged in the University of California, Santa Cruz (UCSC), Genome Browser (22)]. The cleaned pancancer dataset is at karchinlab.org/data/Protocol/pancan-mutation-set-from-Tokheim-2016.txt.gz. The CRAVAT webserver (version 3.0) (23) was used to automatically retrieve the mappability warning codes. Gene names were standardized to HUGO Gene Nomenclature Committee through converting previous symbols and synonyms to the accepted gene name (downloaded January 29, 2015: ftp://ftp.ebi.ac.uk/pub/databases/genenames/locus_groups/protein-coding_gene.txt.gz). Four cancer-specific mutation sets were extracted from the pancancer set: PDAC, BRAC, LUAD, and HNSCC.

Evaluation Metrics. The MLFC is a metric of discrepancy between an observed *P* value distribution reported by a method and a theoretical uniform null distribution. We define $P(i)$ = ith smallest P value, $q(i) = i/n$, and the MLFC = $(1/n)\sum_{i=1}^{n}|\log_2(P(i)/q(i))|$, where P represents the observed P value, q is the corresponding expected P value from a uniform distribution, n is the total number of genes, and MLFC is the average difference of observed and theoretical P values. Values of MLFC near zero indicate smaller discrepancies, and therefore better statistical modeling of the passenger gene null distribution. To evaluate TUSON, which reports both an oncogene (OG) and tumor suppressor gene (TSG) P value, we calculated the average MLFC score between the two.

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

We expect that all methods will lose statistical power and have greater random sampling error when they are predicting on a dataset that has been split in half. Therefore, we chose to allow genes to fall twice as far down the list in the "B" half of the split, to better distinguish random effects and methods with intrinsically low consistency.

20/20+: A Method for Driver Gene Prediction. Our goal was to generalize the 20/20 rule (Fig. S9) by replacing the decision tree and fixed thresholds with a more flexible learning framework. We selected a Random Forest (ensemble of decision trees) and used the set of OGs and TSGs identified by the original 20/20 rule as a training set. We designed a set of 24 predictive features described in Dataset S5 and Fig. S10. 20/20+ uses a three-class Random Forest (24, 25) machine-learning algorithm from the *randomForest* R package to predict whether a gene is an OG, TSG, or passenger gene. Each gene was scored as the fraction of trees that voted for OG, TSG, or passenger gene. A driver score for each was calculated as the sum of the OG and TSG scores. To assess statistical significance, we computed a P value for each score by using Monte Carlo simulations, and subsequently corrected for multiple hypothesis testing using the Benjamini–Hochberg method. See *SI Materials and Methods* for details.

A paper that evaluates various cancer driver methods by independent criteria (26) was published during the review of our manuscript.

ACKNOWLEDGMENTS. Thanks to Dr. Daniel Naiman for review of the statistical analysis. This research was funded by National Cancer Institute (NCI) Grant F31CA200266 (to C.J.T.); NCI Grants 5U01CA180956-03 and 1U24CA204817-01 (to R.K.); and The Virginia and D. K. Ludwig Fund for Cancer Research, Lustgarten Foundation for Pancreatic Cancer Research, The Sol Goldman Center for Pancreatic Cancer Research, and NCI Grant P50-CA62924 (to B.V.).

1. Watson IR, Takahashi K, Futreal PA, Chin L (2013) Emerging patterns of somatic mutations in cancer. *Nat Rev Genet* 14(10):703–718.
2. Sjöblom T, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314(5797):268–274.
3. Parmigiani G, et al. (2009) Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics* 93(1):17–21.
4. Lawrence MS, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214–218.
5. Dees ND, et al. (2012) MuSiC: Identifying mutational significance in cancer genomes. *Genome Res* 22(8):1589–1598.
6. Vogelstein B, et al. (2013) Cancer genome landscapes. *Science* 339(6127):1546–1558.
7. Gonzalez-Perez A, Lopez-Bigas N (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res* 40(21):e169.
8. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N (2013) OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* 29(18):2238–2244.
9. Davoli T, et al. (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* 155(4):948–962.
10. Futreal PA, et al. (2004) A census of human cancer genes. *Nat Rev Cancer* 4(3):177–183.
11. Tamborero D, et al. (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep* 3:2650.
12. Lawrence MS, et al. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505(7484):495–501.
13. Reimand J, Bader GD (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* 9:637.
14. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N (2016) OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* 17(1):128.
15. Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, Bose R (2015) Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics* 31(22):3561–3568.
16. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100(16):9440–9445.
17. Tomasetti C, Vogelstein B, Parmigiani G (2013) Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci USA* 110(6):1999–2004.
18. Schuster-Böckler B, Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 488(7412):504–507.
19. Supek F, Lehner B (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* 521(7550):81–84.
20. Forbes SA, et al. (2011) COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 39(Database issue):D945–D950.
21. NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44(D1):D7–D19.
22. Rosenbloom KR, et al. (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* 43(Database issue):D670–D681.
23. Douville C, et al. (2013) CRAVAT: Cancer-related analysis of variants toolkit. *Bioinformatics* 29(5):647–648.
24. Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. *Neural Comput* 9(7):1545–1588.
25. Breiman L (2001) Random Forests. *Mach Learn* 45(1):5–32.
26. Hofree M, et al. (2016) Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat Commun* 7:12096.
27. Cheng F, Zhao J, Zhao Z (2016) Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief Bioinform* 17(4):642–656.
28. Chernick MR, Liu CY (2012) The saw-toothed behavior of power versus sample size and software solutions. *Am Stat* 56(2):149–155.
29. Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 5(3):1752–1779.
30. Wong WC, et al. (2011) CHASM and SNVBox: Toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* 27(15):2147–2148.