Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations

Hannah Carter,[1] Sining Chen,[2,3] Leyla Isik,[1] Svitlana Tyekucheva,[3] Victor E. Velculescu,[4] Kenneth W. Kinzler,[4] Bert Vogelstein,[4] and Rachel Karchin[1]

[1]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University; [2]Department of Environmental Health Sciences and Department of Biostatistics, Johns Hopkins School of Public Health; and [3]Department of Oncology and [4]Ludwig Center for Cancer Genetics and Therapeutics and Howard Hughes Medical Institute, Johns Hopkins Kimmel Cancer Center, Baltimore, Maryland

Abstract

Large-scale sequencing of cancer genomes has uncovered thousands of DNA alterations, but the functional relevance of the majority of these mutations to tumorigenesis is unknown. We have developed a computational method, called Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM), to identify and prioritize those missense mutations most likely to generate functional changes that enhance tumor cell proliferation. The method has high sensitivity and specificity when discriminating between known driver missense mutations and randomly generated missense mutations (area under receiver operating characteristic curve, >0.91; area under Precision-Recall curve, >0.79). CHASM substantially outperformed previously described missense mutation function prediction methods at discriminating known oncogenic mutations in *P53* and the tyrosine kinase epidermal growth factor receptor. We applied the method to 607 missense mutations found in a recent glioblastoma multiforme sequencing study. Based on a model that assumed the glioblastoma multiforme mutations are a mixture of drivers and passengers, we estimate that 8% of these mutations are drivers, causally contributing to tumorigenesis. [Cancer Res 2009;69(16):6660–7]

Introduction

Today we face a bottleneck between large-scale acquisition of genomic information discovered through medical resequencing projects and the application of this information to improved understanding of human disease. Projects to systematically resequence tumor genomes have discovered thousands of genes that were not previously linked to tumorigenesis but are somatically mutated in a relatively small fraction of tumors and may be important for tumor initiation or progression (1–6). Many of these somatic changes are likely to be "passengers" (1) that have no functional effects but were already present in the cell that gave rise to the tumor or were acquired during subsequent tumor growth. Only a small fraction of the genetic alterations in a tumor are expected to drive tumor evolution by giving cells a selective advantage over their neighbors.

Determining which mutations are drivers and which are passengers is one of the most pressing challenges in cancer genetics. Although genes that are mutated very frequently ("mountains") can be confidently classified as driver genes, most genes discovered thus far are mutated in a relatively small fraction of tumors ("hills"). The examination of large numbers of tumors can provide helpful information for classification of drivers versus passengers, but the ability of sequencing alone to provide definitive results is limited by the marked variation in mutation frequency among individual tumors and individual genes. Moreover, it has been clearly shown that genes that are mutated in only a small fraction (<1%) of tumors can still act as drivers (6). Thus, methods that can classify mutations as either drivers or passengers on the basis of data that is independent of mutation frequency are clearly needed. Such methods include functional studies in model organisms or in cultured cells, using gene knockout, siRNA, or overexpression approaches. These methods are extraordinarily useful for elucidating the function of individual mutated genes but are not well suited to the analysis of the hundreds of gene candidates that arise from every large scale cancer genome project.

Here, we describe a novel high-throughput computational prediction method to identify the mutations most likely to be drivers. We chose to focus on missense mutations as they account for the majority of somatic mutations found in the exons of tumor-derived DNA (6), and because their functional significance is more difficult to infer than that of nonsense or frameshift mutations.

Previous work in this area has resulted in several innovative ways to characterize the differences between driver and passenger missense mutations. Driver mutations may have characteristics similar to those causing Mendelian disease when inherited in the germ line (7) and may be identifiable by constraints on tolerated amino acid residues at the mutated positions (3, 7–9). In contrast, passenger mutations may have characteristics more similar to those of nonsynonymous single nucleotide polymorphisms (nsSNP) with high minor allele frequencies (MAF; refs. 3, 7). Based on these similarities, supervised machine learning methods have been used to predict which missense mutations are drivers (3, 7). The CAN-Predict method trains a Random Forest (10) to discriminate between mutations from the COSMIC cancer somatic mutation database (11) and nsSNPs with high MAFs (3). A method specific to protein kinases (7) trains a support vector machine (SVM; ref. 12) to discriminate between known disease kinase nsSNPs and common kinase nsSNPs. Although not specifically designed for this problem, bioinformatics methods, such as PolyPhen and SIFT (9, 13) have also been applied to identify pathogenic, tumor-derived mutations in genes of interest (6). These methods attempt to discriminate driver from passenger mutations by considering properties such as evolutionary conservation, compatibility of the mutant amino acid residue with the wild-type or with equivalently positioned residues in homologous proteins, the predicted protein

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Rachel Karchin, Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218. Phone: 410-516-5578; Fax: 410-516-5294; E-mail: Karchin@jhu.edu.

doi:10.1158/0008-5472.CAN-09-1133

local environment (7), and enrichment of the protein structural domain in which mutations occur with respect to biological processes thought to be critical for cancer (3).

We hypothesized that although existing computational methods could detect differences between somatic missense mutations observed in cancers and high MAF nsSNPs in the germline, these differences might be less relevant to the discrimination between driver and passenger mutations that occur somatically in tumors. Although high MAF nsSNPs and passenger mutations have properties in common, they also have differences. Passenger mutations may or may not have a functional impact on proteins; by definition, they are neutral with respect to cancer cell fitness. In contrast, high MAF nsSNPs have become fixed in the human genome and must be functionally neutral or have a mild functional impact with respect to normal cell fitness. We reasoned that we could train a classifier with improved specificity by representing passenger missense mutations not by high MAF nsSNPs, as done previously, but rather by *in silico* simulations using mutation profiles that reflected tumor type as well as mutation context.

Materials and Methods

Feature selection. We used a Random Forest classifier (10, 14) that was trained on 49 predictive features (Supplementary Table S1). Feature selection was done with a protocol based on mutual information (Supplementary Materials and Methods: Feature Selection and Information Theory; Supplementary Fig. S1). Mutual information is a generalized version of correlation that does not make assumptions about linear relationships between two variables of interest (15). Features with missing values were estimated with a k-nearest neighbors algorithm (Supplementary Materials and Methods: Missing Values).

Driver mutation data set. We selected 2,488 missense mutations previously identified as playing a functional role in oncogenic transformation from breast, colorectal, and pancreatic tumor resequencing studies (2, 4–6) and the COSMIC database (11).

Synthetically generated passenger mutation data set. The synthetic passenger mutations were generated by sampling from eight multinomial distributions that depend on dinucleotide context and tumor type (Supplementary Materials and Methods: Synthetically Generated Mutations; Supplementary Table S2; Supplementary Fig. S2).

Classifier training. The Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM) method is based on a Random Forest classifier (10, 14) trained to discriminate between driver missense mutations and synthetically generated passenger missense mutations. The classifier is implemented using PARF,[5] a Fortran 95 adaptation of Leo Breiman's original Random Forest software.[6] Before training, all features were standardized with the Z score method using the scale command in R statistical software (16). To avoid overfitting, we divided our known driver mutations and synthetic passenger mutations into two partitions, one for feature selection and one for classifier training.

This Random Forest is an ensemble of "decision trees," specifically classification and regression trees (17), each of which uses a hierarchical set of rules to decide whether a mutation is a driver or a passenger. The rules are based on our input features and the final score yielded for each mutation is the fraction of trees that voted for the passenger class. We used a forest with 500 trees, and default parameters (mtry = 7). The Random Forest algorithm is robust to class label contamination and performs well with high dimensional data sets (10, 14).

Classifier assessment. We assessed Random Forest classifier performance by two threshold-independent measures— receiver operating characteristic (ROC) and Precision-recall (PR) curves (Supplementary Materials and Methods: ROC and Precision-recall Curves and Minimum Error Point). We considered both the training set out-of-bag error (10) and the error on two held-out validation sets of known oncogenic mutations in *P53* and epidermal growth factor receptor (*EGFR*). The out-of-bag error estimate is produced while the Random Forest is being trained and is a viable replacement for error estimates by cross-validation (18). We compared the Random Forest with a SVM classifier (assessed with 5-fold cross-validation; Supplementary Materials and Methods: Support Vector Machine; ref. 12) and with the performance of several state-of-the-art missense mutation function prediction methods.

Probabilistic interpretation of random forest classification scores in tumor-derived glioblastoma multiforme mutations. We used the trained Random Forest to compute a classification score for each of 607 glioblastoma multiforme (GBM) missense mutations reported by Parsons and colleagues (4). However, these scores are not probabilities and the statistical behavior of the algorithm has not been well-characterized (10). Therefore, it is not evident where to set a trusted score cutoff for purposes of identifying driver mutations. To do this, we first interpret the scores in the framework of statistical hypothesis testing. For each of the 607 GBM mutants, we test the null hypothesis: the mutant is not functionally related to the growth of the tumor (passenger), versus the alternative hypothesis that it is (driver). We obtain a *P* value for a mutation by comparing its score to the null distribution, which consists of the scores of a filtered set of synthetic passengers that were held out from Random Forest training (Supplementary Materials and Methods: Filtering of Synthetically Generated Passenger Mutations), using the Benjamini-Hochberg algorithm to correct for multiple testing (Supplementary Materials and Methods: Controlling the False Discovery Rate; ref. 19).

GBM mutations. We assessed 607 GBM mutations from 21 patient samples (4). Five of the mutations described by Parsons and colleagues (4) were dropped because they occurred in gene transcripts that are no longer supported by the RefSeq database (20). Three mutations were dropped because they were found in gene transcripts that were larger than 14,000 codons. For gene transcripts of this size, we were unable to generate protein multiple sequence alignments because of their high computational expense. Finally, one of the GBM tumor samples was from a patient with a hypermutator phenotype who had been treated with radiation and temozolomide. Because this sample had 17 times as many alterations as the other GBM samples and a radically different mutation spectrum (4), these mutations were excluded from our analysis.

Estimation of fraction of drivers in GBM. We assumed that the GBM mutations are a mixture of drivers and passengers and wanted to estimate the proportion of drivers in the mixture. The probability distribution of the GBM CHASM scores should then be similar to the CHASM score distribution of a mixture of known driver and synthetic passenger mutations (21). We numerically find the mixing proportion, which minimizes the distance between these two score distributions (Supplementary Materials and Methods: Estimating the Fraction of Drivers).

Comparison with other methods. For comparison purposes, we assessed the performance of several published methods that were possibly useful for driver mutation prediction both on our training set and the two held-out validation sets of *P53* and *EGFR* mutations. The tested methods were as follows: PolyPhen (22), SIFT (9), CanPredict (3), and KinaseSVM (7). We also assessed a consensus prediction, based on agreement between SIFT and PolyPhen (Supplementary Materials and Methods: Comparison with Other Missense Mutation Function Prediction Methods).

Wherever possible, we assessed the performance of these methods using a numerical score, rather than a categorical prediction, so that we could construct threshold-independent ROC and PR curves. We computed precision and recall statistics (Eq 4) when only categorical predictions were available (CanPredict and the PolyPhen/SIFT consensus).

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad \text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

where *TP* is the number of drivers correctly classified, *FP* is the number of synthetic passengers misclassified, and *FN* is the number of drivers

[5] http://www.irb.hr/en/cir/projects/info/parf/
[6] http://www.math.usu.edu/~adele/forests/cc_home.htm

misclassified. We compared the performance of these methods to CHASM's performance on its own training set, based on out-of-bag scores, and also to CHASM's performance when all *P53* and *EGFR* mutations were held out of its training and feature selection sets. We also compared Random Forest performance with performance of a SVM (12), another state-of-the-art machine learning classifier, using the same training sets and predictive features. The SVM was trained using the e1071 package in R statistical software and assessed using 5-fold cross-validation and constructing ROC and PR curves.

Results

Feature selection. To develop a new classifier, we first evaluated a large number of candidate predictive features and found that >50 features contained at least some information that seemed to be useful for discriminating between driver and passenger mutations. In particular, using a method that estimates mutual information between a predictive feature and class labels, we found that the majority of the candidate predictive features were weakly informative (Supplementary Table 3; ref. 23). In our training set (described in Materials and Methods), we calculated that a feature capable of correctly classifying a mutation as a passenger or driver would require 2.05 bits of information (Supplementary Materials and Methods: Information Theory). As our top-ranked feature had only 0.06 bits of information, we compensated by using 49 features (Supplementary Table S3; Supplementary Fig. S3). This is a much larger number of features than used in previous studies (3, 7). The sum of the information in each individual feature was 0.37 bits. However, the Random Forest works with all features jointly, which may yield much higher information content than the simple sum.

Some of our top-ranked features have not, to our knowledge, been used previously for missense mutant function prediction. These features include the average nucleotide-level conservation of the exon in which a mutation occurs in 17-way vertebrate Multiz alignments (24), estimated by PhastCons (25); SNP density (the number of SNPs in the exon where the mutation occurs, normalized by exon length); and frequency of missense change type in the COSMIC database of somatic variation in cancer (11).

Datasets used for training. As noted in the Introduction, the choice of training sets is critically important to the performance of any classifier. As drivers, we selected 2,488 missense mutations previously identified as playing a functional role in cancer, culled from the COSMIC database and recent large-scale resequencing studies (see Materials and Methods). The passenger data set was derived by a two-step process. First, we selected genes that were mutated at least once in four large-scale sequencing studies of colorectal, breast, brain, or pancreatic tumors (2, 4–6). Second, we generated synthetic passenger missense mutations in these genes *in silico*, using an algorithm that recapitulated the type of base substitutions found in brain tumors (mutation context). Note that we purposefully chose genes that were mutated as the substrate for the *in silico* generation of synthetic mutations. This increased the likelihood that the new classifier would detect *mutations* that were extraordinary rather than detect *genes* that were extraordinary (e.g., had very different codon compositions than the average). Our classifier would thus be able to detect differences between driver and passenger mutations even if the mutations were in the same gene.

Past classifiers have often used high MAF nsSNPs as the passenger data set rather than the synthetic passenger data set described above. To determine whether there were major differences between our new data set and high MAF nsSNPs, we compared them using principal component analysis applied to the top-ranked 21 predictive features (Supplementary Table S1). As shown in Fig. 1, a randomly selected set of 4,395 high MAF ns SNPs from the HapMap project were distributed differently than a set of 4,500 synthetic passengers. Interestingly, the synthetic passengers formed two distinct clusters in this analysis, along the dimension of principal component four, which is dominated by feature 72. The feature is a binary descriptor of regions in proteins that are functionally interesting, as annotated in the UniProtKB database (26). It seems that although a subset of the synthetically generated passenger mutations were located in annotated regions of functional interest, the MAF nsSNPs tended not to be located in these regions. This result is consistent with evolutionary selective pressure on MAF nsSNPs for functional neutrality. Other features with large magnitude coefficients in these principal components analysis components included predicted amino acid residue propensities for secondary structure, solvent accessibility, backbone flexibility, and additional protein-based functional annotations from UniProtKB.

Classifier construction. We then attempted to use these features and data sets to design a new classifier using two state-of-the-art machine learning methods, SVMs, and Random Forests. Although both methods were able to define good classifiers, the Random Forest proved superior (Supplementary Fig. S4) and was used for the remainder of the analyses. Details of the construction of the Random Forest–based classifier, henceforth termed CHASM, are described in Materials and Methods.

To test the performance of CHASM, we first assessed it with respect to its out-of-bag classification error on the training sets (equivalent to a cross-validation test (10)). For this purpose, ROC and PR curves were used, as these metrics consider classification errors at all possible score thresholds. Using area under the curve (AUC) as a performance summary statistic, where 1.0 indicates perfect classification, CHASM yielded AUCs of 0.91 and 0.79 for ROC and PR, respectively (Fig. 2).

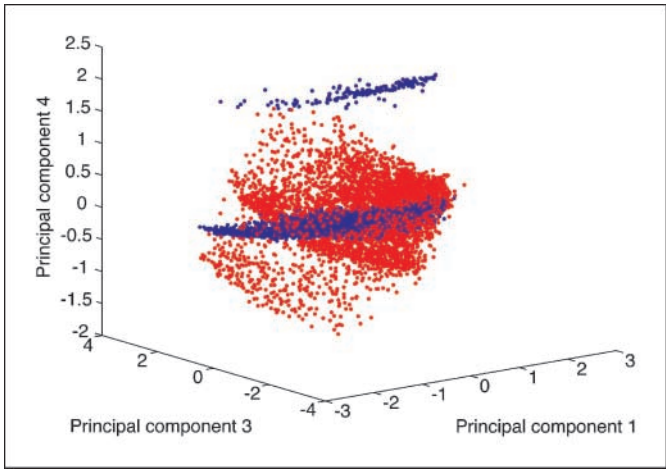

Figure 1. Principal components analysis of nsSNPs versus synthetic passenger mutations. Synthetic passenger mutations (*red*) and high MAF nsSNPs from the HapMap project (*blue*) have substantial overlap in the space defined by principal components one, three, and four, but there are regions in the space occupied only by high MAF nsSNPs and regions occupied only by synthetic passengers.

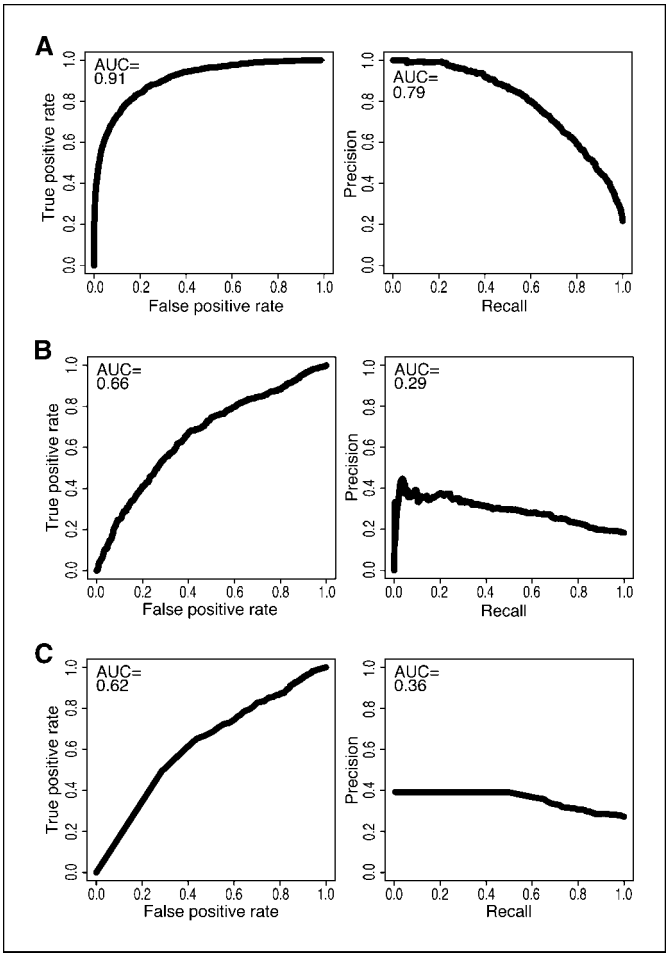

Figure 2. ROC and PR curves calculated for (*A*) CHASM, (*B*) PolyPhen PSIC, and (*C*) SIFT on the training set mutations. CHASM training out-of-bag scores were used to generate the ROC and PR curves in *A*. A color version is available as Supplementary Fig. S6.

This performance was then compared with that of other methods, including PolyPhen's PSIC score, SIFT, CanPredict, KinaseSVM (Supplementary Fig. S5), and a SIFT-PolyPhen consensus. The fraction of mutations that could be evaluated by these alternative methods (coverage) was considerably lower than that of CHASM (Supplementary Materials and Methods; Supplementary Table S4). Moreover, even the best performing of the alternative methods was inferior to CHASM in specificity, sensitivity, and precision (Supplementary Table S4). These differences translated to much lower AUCs for ROC and PR (Fig. 2).

As another test of the performance of CHASM, *P53* or *EGFR* mutations were held out of the mutation data set used for training, and then these known driver mutations were assigned scores by CHASM and the other algorithms. To evaluate both the sensitivity and specificity of each method, we also held out 590 synthetic passenger mutations. If we consider the fraction of misclassified mutations at the minimum error point, the CHASM classifier had high sensitivity and specificity for both the *P53* and *EGFR* test sets (Supplementary Table S4). The performance of CHASM was considerably better, both in terms of sensitivity and specificity, than previously described classifiers (Supplementary Table S4). These differences are graphically illustrated in the AUCs presented in Figs. 3 and 4 (Further detail is provided in Supplementary Table S5.).

For a practical estimate of the CHASM performance, we calculated *P* values for each of the held out *P53* and *EGFR* mutations, then controlled the false discovery rate (FDR) to 0.2 using the Benjamini-Hochberg procedure. We found that 195 of the 196 experimentally observed *P53* mutations and 131 of the 133 experimentally observed *EGFR* mutations were predicted to be drivers by CHASM. In comparison, a maximum of 188 of the 196 experimentally observed *P53* mutations and 101 of the 133 experimentally observed *EGFR* mutations were predicted to be drivers by PolyPhen or SIFT.

Analyses of GBM. The CHASM Random Forest classifier was then used to score 607 missense mutations in glioblastoma multiforme (GBM) described by Parsons and colleagues (4). The driver data set used to train the Random Forest was the same as that described above except that all of the missense mutations actually observed in GBMs were excluded. The raw CHASM scores of the mutations, representing the fraction of trees in the forest that voted for classifying the mutation as passenger, ranged from 0 to 1 (Fig. 5). For each of these missense mutants, we tested the null hypothesis that the mutant was a passenger. A *P* value was calculated for each mutant by comparing its CHASM

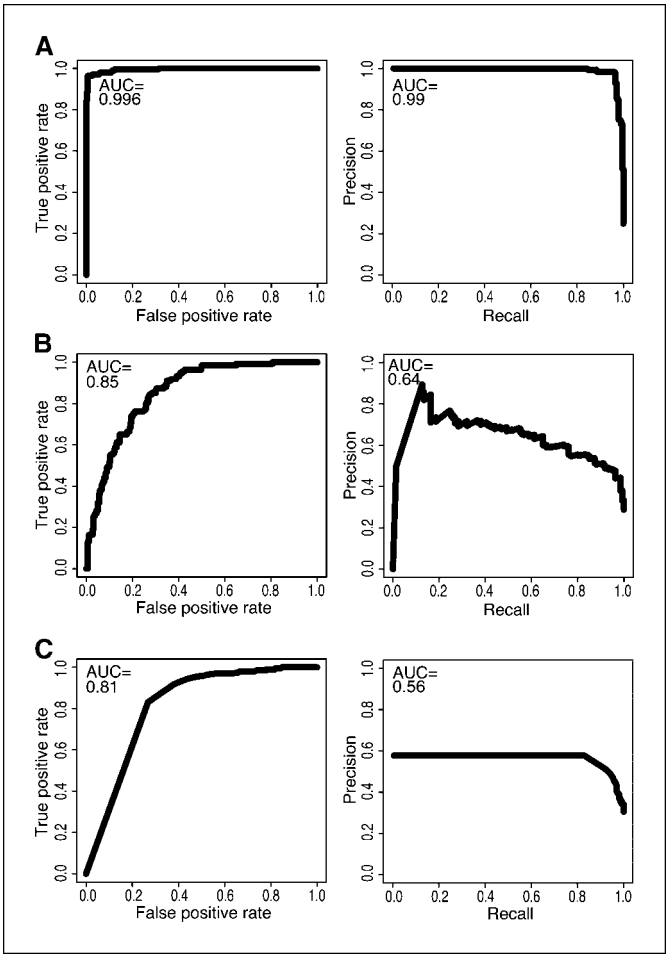

Figure 3. ROC and PR curves calculated for (*A*) CHASM, (*B*) PolyPhen PSIC, and (*C*) SIFT on *P53* and synthetic passenger mutations held out of the CHASM training set. A color version is available as Supplementary Fig. S7.

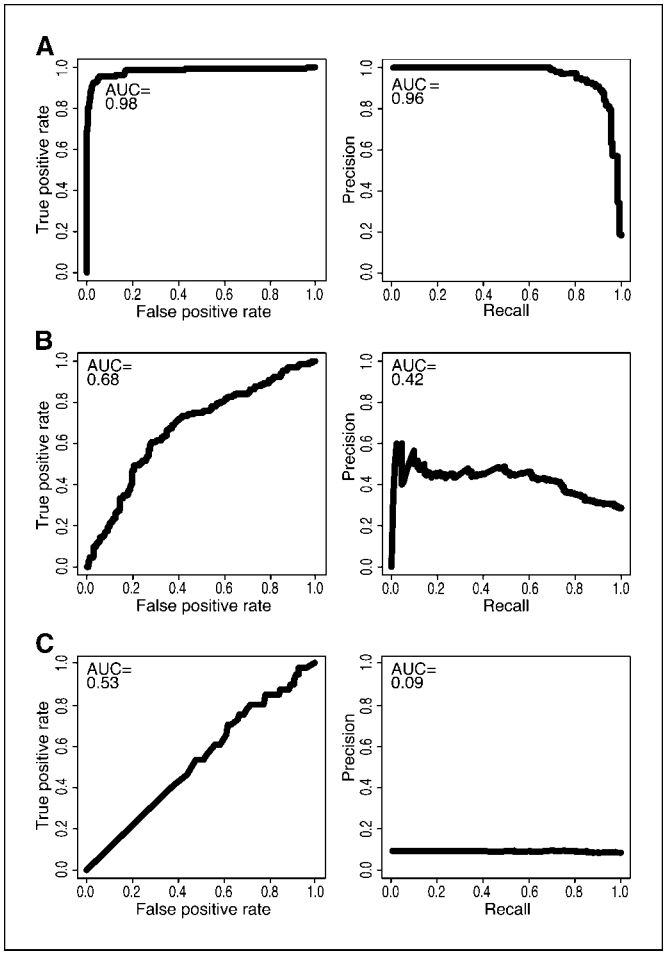

Figure 4. ROC and PR curves calculated for (*A*) CHASM, (*B*) PolyPhen PSIC, and (*C*) SIFT on *EGFR* and synthetic passenger mutations held out of the CHASM training set. A color version is available as Supplementary Fig. S8.

score to the score distribution of a filtered set of synthetic passengers (see Materials and Methods for details). The Benjamini-Hochberg procedure was used to control the FDR at the desired level of 0.2 (19).

At this FDR level, CHASM classified 24 of the 607 GBM mutations as drivers (Table 1). Importantly, CHASM successfully identified 11 mutations that were likely to be drivers based on previous experimental data. These 11 mutations included nine in *P53* or *PTEN*, well-known tumor suppressor genes, one in *PIK3CA*, a well-known oncogene and one in *IDH1*, a gene recently discovered to be altered in many brain tumors (27). In addition to these, 11 CHASM identified 13 others that otherwise would not have been suspected of playing a major role in GBM tumorigenesis (Table 1). Intriguingly, these mutations included those in genes that are likely to be involved in critical signaling pathways, such as the protein kinases *STK39* and *RIPK4*, the protein phosphatase *PTPRM*, and the insulin-signaling mediator *PHIP*.

Finally, to estimate the proportion of driver missense mutations in the GBM mutation set, we minimized the difference between the distributions of the CHASM scores of the GBM mutations and the CHASM scores of a mixture of known driver and synthetic passenger mutations (see Materials and Methods for details). We thereby estimated that 49 of the 607 missense mutations identified in GBM, or 8%, were drivers.

Discussion

Computational methods to predict the impact of mutations discovered in tumor resequencing are still under development. Although initial work focused on identification of driver *genes* rather than driver *mutations* (1, 5), it has recently been suggested that the occurrence of some missense mutations in oncogenes or tumor suppressor genes are actually passengers (7), motivating the need for a higher resolution approach that identifies individual mutations as drivers. In light of the large number of mutations that are being discovered in current large-scale cancer gene sequencing efforts, and the impossibility of assessing this large number through experimental functional studies, bioinformatic approaches to classify and prioritize mutations for further analysis are essential for progress.

Confronted with this problem, some researchers have tried to apply methods that were developed to predict the impact of germline missense variants. We found that these methods have good sensitivity in recognizing recurrent driver missense mutations in *P53* and *EGFR*, but poor specificity (Supplementary Table S4; Figs. 3 and 4). This result implies that there may be differences between the distinguishing characteristics of neutral mutations in the cancer genome versus the germline genome. Application of methods developed for the latter problem to the former problem yielded less than optimal results. In contrast, the CHASM classifier, specifically developed to detect somatic rather than germline driver mutations, had substantially improved sensitivity, specificity, and precision over previously described methods.

Overall, our results highlight the importance of "null model" selection in designing a predictive algorithm to identify driver mutations in cancer resequencing data. Within the context of a prediction method, the null model incorporates assumptions about what driver missense mutations do not look like. It is used explicitly in supervised learning methods such as CAN-predict, Kinase SVM, and our previous version of CHASM (2, 4). It is also

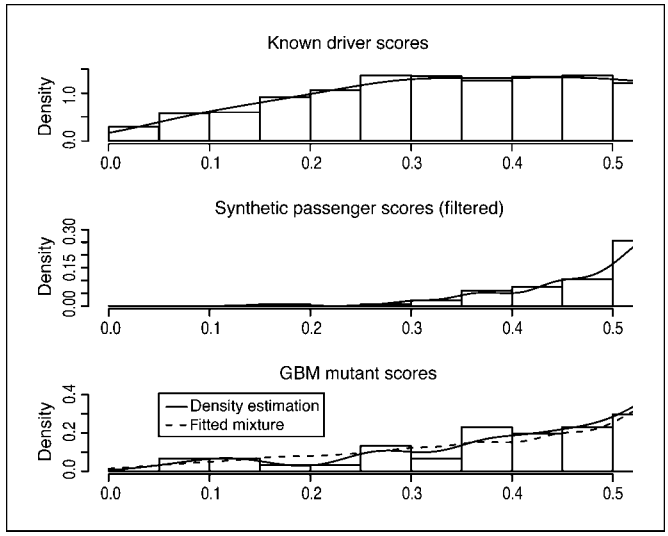

Figure 5. Histograms of CHASM scores for driver mutations and passenger mutations held out from the training set, and 607 mutations experimentally identified in GBM. Estimated kernel density for each set of scores (*solid line*) and fitted mixture of the driver and passenger score densities (*dashed line*) are shown superimposed on the histograms.

Table 1. Driver mutations predicted by CHASM at FDR of 0.2, shown with their associated Random Forest scores and *P* values

Hugo Gene Symbol	Mutation	CHASM score	*P*	Protein function	Cancer association
P53	C176F	0.054	0.0004	Regulates various cellular processes including cell cycle, proliferation and, apoptosis (32)	*P53* is a tumor suppressor and is compromised in almost all human cancers (32)
	R273H	0.128	0.0004		
	G245S	0.098	0.0004		
	G245D	0.112	0.0004		
	R273C	0.156	0.0004		
	R248W	0.242	0.0008		
	V197E	0.264	0.0008		
	R282W	0.266	0.0008		
STK39/SPAK	I208T	0.268	0.0008	A serine/threonine kinase that regulates the *p38* MAP kinase pathway (33)	*STK39/SPAK* has been implicated in the regulation of prostate cell proliferation through androgen response (33)
ST8SIA4	R168S	0.286	0.0011	ST8SI4 is an enzyme necessary for the synthesis of polysialic acid, which is present on the embryonic neural cell adhesion molecule (N-CAM). N-CAM plays an important role in neuronal plasticity (34)	E-cadherin–mediated cell-cell adhesion is repressed by polysialylated N-CAM in pancreatic tumor cells (34)
F2RL1/PAR2	C226S	0.302	0.0011	Acts as a receptor for trypsin and trypsin-like enzymes. F2RL1 is coupled to G proteins that stimulate phosphoinositide hydrolysis. This protein has also been suggested to play a role in the regulation of vascular tone (35)	*F2RL1/PAR2* signaling may contribute to angiogensis and tumor growth (35)
IDH1	R132S	0.324	0.0019	IDH1 catalyzes the oxidative carboxylation of isocitrate to α-ketoglutarate, resulting in the production of NADPH (4)	*IDH1* mutations occur frequently in brain tumors and have been causally implicated in glioma progression (27)
ABL2	P487L	0.336	0.0019	Regulates cytoskeleton in cellular division, differentiation, and adhesion through phosphorylation of proteins controlling cytoskeleton dynamics (36)	*ABL2* may inhibit glioma cell migration and cause cytoskeletal collapse through inactivation of RhoA (36)
PHIP	D246G	0.36	0.0030	Interacts with IRS-1 and may mediate downstream insulin signaling (37)	IRS-1 interacts with many oncogenes and is important for their ability to transform the cell (38)
PHF2	A199T	0.366	0.0030	Contains a PHD finger domain and may play a role in chromatin structure modification (39)	*PHF2* is frequently altered in breast cancer (40)
PIK3CA	G1049S	0.386	0.0042	Phosphorylates the second messengers PtdIns, PtdIns4P and PtdIns(4, 5)P2 with a preference for PtdIns(4, 5)P2 (41)	*PIK3CA* regulates cell cycle progression and cell survival through AKT and is frequently altered in glioblastomas (41)
PTEN	G132S	0.376	0.0042	Antagonizes the PI3K-AKT/PKB signaling pathway by dephosphorylating phosphoinositides and thereby modulating cell cycle progression and cell survival (42)	*PTEN* is a tumor suppressor in the PIK3CA/AKT pathway and is altered in 60% of glioblastomas (42)
ABCC3	D1505Y	0.376	0.0042	ABC proteins transport various molecules across cellular membranes. *ABCC3* is a member of the MRP subfamily of ABC transporters implicated in multidrug resistance (43)	Small-cell lung cancer patients with aberrations in *ABCC3* show significantly decreased progression-free survival (43)

(Continued on the following page)

Table 1. Driver mutations predicted by CHASM at FDR of 0.2, shown with their associated Random Forest scores and *P* values (Cont'd)

Hugo Gene Symbol	Mutation	CHASM score	*P*	Protein function	Cancer association
RIPK4	P222Q	0.374	0.0042	A serine/threonine protein kinase that interacts with protein kinase C-δ and can increase nuclear factor-κB (NF-κB) activity. This protein is necessary for keratinocyte differentiation (44)	Regulates NF-κB, a transcription factor implicated in the initiation and progression of cancer (45)
FLJ10276/BSDC1	K172E	0.4	0.0053	Uncharacterized protein containing a BSD domain. May act as a transcription factor (46)	Unknown
SLC30A9/HUEL	G321D	0.424	0.0060	*SLC30A9* may be a housekeeping gene involved in cellular replication, DNA synthesis, and/or transcriptional regulation (47)	*SLC30A9* is located in a region of chromosome 4 that is frequently deleted in carcinomas (47)
CYP2C19	P382L	0.428	0.0064	CYP2C19 is a cytochrome P450 enzyme that metabolizes a number of therapeutic agents including the anticonvulsant drug S-mephenytoin, omeprazole, proguanil, certain barbiturates, diazepam, propranolol, citalopram, and imipramine (48)	Altered CYP2C19 mediated drug metabolism could effect the tumor's response to therapy (48)
LBP	E363K	0.428	0.0064	LBP binds bacterial lipopolysaccharides, and transfers them to the CD14 receptor (49)	CD14 is upstream of both the NF-κB and MAP kinase signaling pathways, both of which are often deregulated in cancer (49)
PTPRM	M1220V	0.434	0.0072	PTPRM is implicated in cell-cell contact formation through homophillic interaction and seems to play a role in signal transduction in response to cell density (50)	*PTPRM* my play a role in cell-cell contact signaling to regulate cell growth (50)

NOTE: This list includes 11 mutations likely to be drivers based on previous experimental data and 13 others that otherwise would not have been suspected of playing a major role in GBM tumorigenesis, but which are found in genes that are likely to be involved in critical signaling pathways.

used implicitly in methods such as SIFT and PolyPhen because their utility has been assessed with a validation or benchmark set as a false-positive control. SIFT has used experimental results of functional assays in bacterial and viral proteins as a control; PolyPhen has used species divergence data from amino acid substitutions found in equivalent positions in alignments of protein orthologs. We suggest that these null models of functional neutrality do not optimally represent the passenger missense mutations found in tumors.

Although existing methods for missense mutant function prediction in cancer have provided tools to prioritize candidate driver mutations, we have developed a quantitative approach to identify candidate drivers by controlling the FDR. To our knowledge, this is the first application of FDR to the classification of missense mutations, providing a statistically meaningful threshold for discovery.

We estimate that the proportion of drivers among all GBM missense mutations in our data set is ~8%, with 5.4% occurring outside of known gene mountains. Note that the actual number of drivers in the mutation data set of Parsons and colleagues (4) is likely to be higher, as CHASM only considers missense mutations. Many of the tumor suppressor gene alterations that drive tumorigenesis are nonsense mutations, frameshifts, or large deletions.

Our method is high-throughput and can be easily adapted to any tumor type of interest, given a sufficient sample size to compute context-based DNA mutation rates. It also represents an advance over previous classifiers in that most mutations can be scored (coverage; Supplementary Table S4). Because the method focuses on properties of individual mutations, rather than the frequency at which mutations appear in a gene, it can potentially detect driver mutations that are present at low frequencies. These mutations may disregulate pathways that are potential new drug targets. A recent example is the isocitrate dehydrogenase (IDH1) R132 mutation, discovered in GBM resequencing (4). In the initial screen by Parsons and colleagues (4), this mutation was originally found in only a small proportion of GBMs, so its role as a driver was questionable. CHASM, however, shows that the mutation has a high likelihood of being a driver when present in a tumor. Subsequent studies revealed that the mutation was present in a high fraction of an uncommon GBM subtype as well as other brain tumor types (4, 27–30). Functional studies suggest that mutant IDH1 dominantly inhibits production of α-ketogluterate, which is required by enzymes that degrade HIF-1α, thus hyperactivating the HIF-1 pathway and promoting tumor angiogenesis. Drugs designed to be α-ketogluterate mimics might thus be useful for GBM patients with the *IDH1* mutation (31). We hope CHASM will provide a useful tool to guide follow-up experiments based on the

results of the many cancer genome projects now being performed or planned.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Received 3/27/09; revised 6/11/09; accepted 6/18/09; published OnlineFirst 8/4/09.

Grant support: DOD NDSEG fellowship; NIH/National Cancer Institute grant CA135877; Susan G. Komen foundation grant KG080137; the Virginia and D.K. Ludwig Fund for Cancer Research, NIH grants CA121113, CA43460, CA57345, and CA62924; and National Cancer Institute/SAIC contract 28XS268.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We would like to thank Giovanni Parmigiani and Mark Diekhans for valuable discussions and Ivan Adzubey, Josh Kaminker, and Ali Torkamani for help scoring the mutations with PolyPhen, CanPredict, and KinaseSVM.

References

1. Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153–8.
2. Jones S, Zhang Z, Parsons DW, et al. Core signaling pathways in human pancreatic cancer revealed by tumor genome analysis. Science 2008;321:1801–6.
3. Kaminker JS, Zhang Y, Waugh A, et al. Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res 2007;67:465–73.
4. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of glioblastoma multiforme. Science 2008;321:1807–12.
5. Sjoblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–74.
6. Wood LD, Parsons DW, Jones S, et al. The genomic landscapes of human breast and colorectal cancers. Science 2007;318:1108–13.
7. Torkamani A, Schork NJ. Prediction of cancer driver mutations in protein kinases. Cancer Res 2008;68:1675–82.
8. Barnholtz-Sloan J, Sloan AE, Land S, Kupsky W, Monteiro ANA. Somatic alterations in brain tumors. Oncol Rep 2008;20:203–10.
9. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863–74.
10. Breiman L. Random forest. Machine Learning 2001; 45:5–32.
11. Forbes S, Clements J, Dawson E, et al. Cosmic 2005. Br J Cancer 2006;94:318–22.
12. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995.
13. Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet 2001;10:591–7.
14. Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Comput 1997;9: 1545–88.
15. Cover T, Thomas J. Elements of information theory. 1st ed: Wiley and Sons; 1991.
16. R Core Development Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2008.
17. Breiman L. Classification and regression trees Regression trees The Wadsworth statistics/probability series: Wadsworth International Group; 1984.
18. Bylander T. Estimating generalization error on two-class datasets using out-of-bag estimates. Machine Learning 2002;48:287–97.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 1995;57:289–300.
20. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2008;36 S1:D13–21.
21. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 2001; 96:1151–60.
22. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30:3894–900.
23. Karchin R, Kelly L, Sali A. Improving functional annotation of non-synonomous SNPs with information theory. Pac Symp Biocomput 2005;10:397–408.
24. Blanchette M, Kent WJ, Riemer C, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 2004;14:708–15.
25. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034–50.
26. Wu CH, Apweiler R, Bairoch A, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. 2006.
27. Yan H, Parsons DW, Jin G, et al. IDH1 and IDH2 Mutations in Gliomas. N Engl J Med 2009;360:765.
28. Balss J, Meyer J, Mueller W, Korshunov A, Hartmann C, von Deimling A. Analysis of the IDH1 codon 132 mutation in brain tumors. Acta Neuropathol (Berl) 2008; 116:597–602.
29. Watanabe T, Nobusawa S, Kleihues P, Ohgaki H. IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. Am J Pathol 2009;174:1149.
30. Bleeker FE, Lamba S, Leenstra S, et al. IDH1 mutations at residue p. R132 (IDH1R132) occur frequently in high-grade gliomas but not in other solid tumors Communicated by Richard Wooster. Human Mutation 2009;30:7–11.
31. Zhao S, Lin Y, Xu W, et al. Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1α. Science 2009;324:261–5.
32. Whibley C, Pharoah PDP, Hollstein M. p53 polymorphisms: cancer implications. Nat Rev Cancer 2009;9: 95–107.
33. Qi H, Labrie Y, Grenier J, Fournier A, Fillion C, Labrie C. Androgens induce expression of SPAK, a STE20/SPS1-related kinase, in LNCaP human prostate cancer cells. Mol Cell Endocrinol 2001;182:181–92.
34. Schreiber SC, Giehl K, Kastilan C, et al. Polysialylated NCAM represses E-cadherin-mediated cell-cell adhesion in pancreatic tumor cells. Gastroenterology 2008;134: 1555–66.
35. Ruf W, Mueller BM. Thrombin generation and the pathogenesis of cancer. Semin Thromb Hemost 2006;32 Suppl 1:61-8.
36. Shimizu A, Mammoto A, Italiano JE, Jr., et al. ABL2/ARG tyrosine kinase mediates SEMA3F-induced RhoA inactivation and in cytoskeleton collapse human glioma cells. J Biol Chem 2008;283:27230–8.
37. Kaburagi Y, Okochi H, Satoh S, et al. Role of IRS and PHIP on insulin-induced tyrosine phosphorylation and distribution of IRS proteins. Cell Struct Funct 2007;32: 69–78.
38. Dearth RK, Cui X, Kim HJ, Hadsell DL, Lee AV. Oncogenic transformation by the signaling adaptor proteins insulin receptor substrate (IRS)-1 and IRS-2. Cell Cycle 2007;6:705.
39. Sinha S, Singh R, Alam N, Roy A, Roychoudhury S, Panda C. Alterations in candidate genes PHF2, FANCC, PTCH1 and XPA at chromosomal 9q22.3 region: Pathological significance in early- and late-onset breast carcinoma. Molecular Cancer 2008;7:84.
40. Hasenpusch-Theil K. PHF2, a novel PHD finger gene located on human chromosome 9q22. Mamm Genome 1999;10:294–8.
41. Kita D, Yonekawa Y, Weller M, Ohgaki H. PIK3CA alterations in primary (*de novo*) and secondary glioblastomas. Acta Neuropathol (Berl) 2007;113:295–302.
42. Koul D. PTEN signaling pathways in glioblastoma. Cancer Biol Ther 2008;7:1321–5.
43. Muller PJ, Dally H, Klappenecker CN, et al. Polymorphisms in ABCG2, ABCC3 and CNT1 genes and their possible impact on chemotherapy outcome of lung cancer patients. Int J Cancer 2009;124:1669–74.
44. Moran ST, Haider K, Ow Y, Milton P, Chen L, Pillai S. Protein kinase C-associated kinase can activate NFκB in both a kinase-dependent and a kinase-independent manner. J Biol Chem 2003;278:21526–33.
45. Basseres DS, Baldwin AS. Nuclear factor-κ-B and inhibitor of κ-B kinase pathways in oncogenic initiation and progression. Oncogene 2006;25:6817–30.
46. Doerks T, Huber S, Buchner E, Bork P. BSD: a novel domain in transcription factors and synapse-associated proteins. Trends Biochem Sci 2002;27: 168–70.
47. Sim DLC, Yeo WM, Chow VTK. The novel human HUEL (C4orf1) protein shares homology with the DNA-binding domain of the XPA DNA repair protein and displays nuclear translocation in a cell cycle-dependent manner. Int J Biochem Cell Biol 2002;34: 487–504.
48. Rodriguez-Antona C, Ingelman-Sundberg M. Cytochrome P450 pharmacogenetics and cancer. Oncogene 2006;25:1679–91.
49. Triantafilou M, Triantafilou K. Lipopolysaccharide recognition: CD14, TLRs and the LPS-activation cluster. Trends Immunol 2002;23:301–4.
50. Anders L, Mertins P, Lammich S, et al. Furin-, ADAM 10-, and γ-Secretase-mediated cleavage of a receptor tyrosine phosphatase and regulation of β-Catenin's transcriptional activity. Mol Cell Biol 2006; 26:3917–34.