Supplementary Materials

Feature Selection

We designed 80 candidate features that were potentially useful in identifying driver missense mutations (Supplementary Table 1). Features were selected by ranking them according to their mutual information with class labels of interest. We assessed the mutual information of each feature and class labels with a modified version of a method described in (1), which corrects for the overestimation of mutual information in small samples (2) with a permutation analysis. The method consists of these steps:

- if a feature is not categorical, its values are discretized into five equal-width bins;
- the mutual information between the discretized feature values and the class labels is computed;
- class labels are permuted 500 times to yield 500 estimates of ``random'' mutual information between feature values and permuted labels;
- the sample mean of the 500 estimates of random mutual information is subtracted from the initial estimate of mutual information.

The feature selection method was implemented in a custom perl script, available upon request from the authors.

After ranking the candidate features, we considered three alternative approaches for feature selection:

1. use all candidate features, regardless of their mutual information with class labels;
2. use the top k candidate features, ranked by mutual information;
3. use all candidate features with mutual information at least 0.001 bits.

We tried these approaches (including several values of k), and chose to use all candidate features with mutual information of at least 0.001 bits. This choice yielded the best results in the Random Forest performance on the training partition (data not shown). Features 45, 58 and 6 were discarded because they were essentially identical to other features with respect to their information content (Supplementary Table 1).

Information Theory

Our feature selection protocol relies on information theoretic measures to estimate the predictive value of a set of candidate features. Information theory has its roots in communications engineering, and information is related to the number of bits required to transmit a message, and to how surprising the message is to a receiver. A message consists of a series of events, represented by a random variable X, with probability density function $p(X)$ that takes on different values $X_1, X_2, X_3 \ldots X_n$, each with probability $p(X_i)$. As we

receive each event, we get some information that is reciprocally related to $p(X_i)$, so that the least-probable, most-surprising events convey the most information. Information is represented on a log scale as

$$\log_2 \frac{1}{p(X_i)} = -\log_2 p(X_i) . \qquad \text{Supp Eq 1}$$

The expected value of the information in a series of random events is called *entropy* (3), a sum of the events' information content, weighted by the probability of each event.

$$H(X) = -\sum_i p(X_i) \log_2 \, p(X_i) \qquad \text{Supp Eq 2}$$

A series of events about which we are very uncertain has high entropy, while a series of events about which uncertainty is low has low entropy.

Mutual information can be interpreted as the gain of information about a random variable *X* due to additional information from a second random variable *Y* or conversely, how the entropy of *X* is decreased by knowledge of *Y*.

$$I(X,Y) = H(X) - H(X \mid Y) \qquad \text{Supp Eq 3}$$

Substituting Supp Eq 2 into Supp Eq 3 yields (after some algebra)

$$I(X,Y) = \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)} \qquad \text{Supp Eq 4}$$

If we let random variable *X* represent a predictive feature and random variable *Y* represent class labels, we can use mutual information to measure how much information we gain about class labels (such as driver and passenger mutations) from knowledge of a feature. In this setting, we do not know the probability density functions *p(X), p(Y), or p(X,Y)* and we build contingency tables for each *(X,Y)* pair to obtain empirical estimates of the densities. Continuous-valued features are assigned to five discrete categories (bins) *via* a histogram analysis (Supplementary Figure 1). For example, the "17-way exon conservation" feature measures the average PhastCons score (4) of DNA bases in an exon, based on Multiz alignments of 17 vertebrate species (5, 6). In our training set of driver and synthetic passenger mutations, these values ranged from 0.47 to 0.9 and were binned into five categories (Supplementary Figure 1a). Each mutation is thus associated with a class label (1=driver, 2=passenger) and a feature category (1,2,3,4 or 5) (Supplementary Figure 1b). We then estimate *p(X,Y)* for each combination of class and feature category, and the marginal probabilities *p(X)* and *p(Y),* by their frequencies in the training set (Supplementary Figure 1c).

If a particular class and feature category always occurred together, the feature could be used to predict the class with certainty. In this case, the mutual information between feature and class is at a maximum. For example, if "17-way exon conservation" is always low (categories 1,2 or 3) for passengers and high (categories 4 or 5) for drivers, *p(X,Y)* must always be 0 or *p(X)*. An example of a contingency table for this scenario is shown in Supplementary Figure 1d. The maximum mutual information between a feature and class labels can then be computed from Supp Eq 4 as:

$$I(X,Y) = \sum_{x \in X, y \in Y} p(x) \log_2 \frac{1}{p(y)} \qquad \text{Supp Eq 5}$$

Synthetically Generated Mutations

The synthetic passenger mutations were generated using the steps shown in SYNTHETIC_MUTATE_TRANSCRIPT (Supplementary Figure 2). Specifically, each transcript sequence was converted into a sequence of di-nucleotide dependent contexts (C in CpG, C in TpC, G in CpG, G in GpA, other C, other G, A, T) (pseudocode shown in GET_CONTEXT (Supplementary Figure 2). Next, given a base position and a context, we randomly generated a synthetic passenger mutation by sampling from a multinomial distribution that depends on context and tumor type (Supplementary Table 2) pseudocode in SYNTHETIC_MUTATE (Supplementary Figure 2). Because of the way in which the steps are ordered, a C in both a CpG and a TpC di-nucleotide is considered to be in the CpG context. Likewise, if a G is in both a CpG and a GpA di-nucleotide, it is considered to be in the CpG context. 9590 synthetic passengers were generated and partitioned as follows: 4500 for feature selection, 4500 for classifier training, 590 for classifier testing of TP53 and EGFR mutations.

Missing Values

Any feature for which a value could not be computed was estimated using the K-nearest neighbors (KNN) algorithm, implemented in the EVM package of R statistical software (7). Missing values occurred because:

- the tumor sequencing yielded mutations in some transcripts whose identifiers (from RefSeq (8) or Ensembl (9) could not be mapped to UniProtKB (10), required for UniProtKB Annotation features;
- window-based amino acid residue sequence composition features could not handle residues at the beginning and end of a sequence.

We used the nearest 10% of the data to determine the value of a missing feature. Supplementary Figure 3 shows an overview of the CHASM feature building process.

ROC and PR Curves

In this setting, the receiver operating characteristic (ROC) curve (11, 12) is composed of points that represent the trade-off between sensitivity (fraction of drivers correctly classified) and 1 - specificity (1 - fraction of passengers correctly classified) over a monotone increasing threshold value. The precision-recall curve is similarly composed of points that represent the trade-off between precision (fraction of true drivers out of all predicted drivers) and recall (another name for sensitivity). Both are standard measures for assessing classifier performance. The precision recall curve is particularly useful in a setting where there is an imbalance between the size of the two classes of interest (13). The area under these curves (AUC) can be used as a summary statistic for classifier performance (11, 12).

Minimum Error Point

If we consider all possible thresholds at which to evaluate classifier error, there will be at least one threshold value, where the number of misclassified drivers plus the number of misclassified passengers is at a minimum. This threshold is called the minimum error point. To compare CHASM to classifiers that have a fixed threshold, such as the consensus between SIFT and PolyPhen, we computed the precision and recall of CHASM at its minimum error point (Supplementary Tables 4).

Controlling the False Discovery Rate (FDR)

When testing these hypotheses simultaneously, we wish to keep the rate of false discoveries low in the list of mutants we predict to be drivers. We use the Benjamini-Hochberg procedure (14) to control the FDR at a reasonably conservative and often-used level of 0.2. Benjamini-Hochberg offers a linear step-up algorithm for controlling FDR at level q. Let p_i denote the p-values of m mutations. The p-values are ranked in increasing order, such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. Let r be the largest i such that $p_{(i)} \leq \frac{i \cdot q}{m}$. Then, we reject all p-values that are less than $p_{(r)}$.

Estimating the Fraction of Drivers

We developed the following approach to estimate the fraction of true drivers in our set of GBM missense mutations. The GBM mutations are a mixture of drivers and passengers. Therefore, the density distribution of their scores f_G can be written as a mixture of two score density distributions: $f_D(s)$ for driver scores and $f_P(s)$ for passenger scores, with mixing parameter λ, that is the proportion of drivers

$$f_G = \lambda f_D(s) + (1-\lambda) f_P(s) \qquad \text{Supp Eq 6}$$

This model is similar to the model described in (15), in that $f_G(.)$ is the "full distribution" of scores, $f_P(.)$ can be regarded as the "null distribution", and $f_D(.)$ as the "alternative distribution". The driver proportion λ is equivalent to one minus the "null proportion".

We used kernel density estimation (16) of the GBM CHASM scores to obtain $f_G(.)$. To estimate $f_D(.)$ and $f_P(.)$, first, we used our trained Random Forest to compute scores for a held-out partition of Drivers and Passengers that were not included in its training set; then we used kernel density estimation on these scores.

When the null distribution is known, it has been suggested that the null proportion can be estimated by studying the full score (or p-value) distribution in an interval consisting of mostly nulls (17, 18). In our case, since the drivers used for estimating $f_D(.)$ are known, while passengers for $f_P(.)$ are synthesized, $f_D(.)$ is better characterized than is $f_P(.)$. Therefore we base the driver proportion estimation on an interval of $f_G(.)$ that we know with more confidence consists of mostly drivers. For this we choose the interval (0, 0.5). We then find proportion of drivers λ by finding λ*, the value that minimizes the distance between the observed f_G and the mixture of observed f_D and f_P in this interval:

$$\lambda^* = \arg\min_{\lambda} \left\langle f_G, \lambda f_D + (1-\lambda) f_P \right\rangle_{(0,0.5)} \qquad \text{Supp Eq 7}$$

where the distance metric between two densities $\langle f_1, f_2 \rangle$ is defined as the total squared difference between the two densities, which gives

$$\lambda^* = \arg\min_{\lambda} \int_0^{0.5} \left((\lambda f_D(u) + (1-\lambda) f_P(u) - f_G(u) \right)^2 du \qquad \text{Supp Eq 8}$$

We numerically solve for λ*, using R statistical software.

Filtering of Synthetically Generated Passenger Mutations

FDR control and mixture model estimation of the proportion of driver mutations in the GBM samples require that the passenger class is uncontaminated, because the null hypothesis needs to be true for each member of the class. Our method of constructing synthetic passengers is likely to have produced a few driver mutations in the passenger dataset, thereby contaminating our class labels. Therefore, for FDR control and mixture modeling, we filtered the synthetically generated passengers to exclude any gene transcript that has been associated with oncogenic impact, using the complete list of known cancer-associated and potentially cancer-associated genes from the Atlas of Genetics and Cytogenetics in Oncology and Haematology (http://atlasgeneticsoncology.org/Indexbyalpha/idxa_A.html). This procedure reduced our list of synthetic passengers by 41%. None of the synthetic passengers that were used for the empirical null were in the Random Forest training set. We did not apply this filtering procedure to the synthetic passengers used for initial feature selection or to those used for Random Forest training, because it likely also eliminates many true passengers. The Random Forest is known to be robust to label

contamination and benefits from having a larger and more comprehensive feature selection and training set (19).

Support Vector Machine

We compared the Random Forest classifier to a support vector machine classifier on our training set. To evaluate SVM performance, we used five-fold cross validation and constructed ROC and PR curves and area under the curve (AUC) summary statistics (Supplementary Figure 4). The SVM performed better when we used class weights (data not shown), which compensate for the class imbalance in our training set (the Random Forest performance did not improve with class weights). As described previously (20), the SVM class weights (w_D for drivers, w_P for synthetic passengers) were set to w_D =1.56 and w_P=0.44 to down-weight the majority class (passengers) and up-weight the minority class (drivers), according to the proportions that they are found in the training set and to ensure that the original sum of class weights was unchanged (in the "unweighted" case, each class has a default weight of 1.0), so that $w_D + w_P = 2$ and $0.78\ w_D = 0.22\ w_P$.

Comparison with Other Missense Mutant Function Prediction Methods

We submitted all driver mutations and synthetically generated passenger mutations used in our classifier training set to the PolyPhen server in batch *via* two perl scripts, *pph-submit.pl* and *pph-retrieve.pl*, provided to us by Ivan Adzhubey. The same mutations were submitted to the CanPredict server (http://www.cgl.ucsf.edu/Research/genentech/canpredict/), with assistance from Josh Kaminker. SIFT scores for the mutations were generated with a locally installed copy of SIFT3.0, downloaded from the SIFT website (http:// blocks.fhcrc.org/sift), and the most current version of the nr protein sequence database as of February 12, 2009. To obtain scores for the subset of these mutations that occur in protein kinases, all mutations were converted to KinBase gene identifiers (http://kinase.com/Kinbase/). For these mutations, Ali Torkamani provided blinded scores, based on the kinase SVM method (21). Finally, we looked at the intersection of mutations predicted to be functional by both SIFT and PolyPhen (SIFT/PolyPhen Consensus).

PolyPhen

PolyPhen classifies missense mutants as “Probably Damaging”, “Possibly Damaging” or “Benign” and also provides a continuous measure of a mutation’s functional impact, the PSIC score (22). We assessed PolyPhen with threshold-independent ROC and PR curves, based on the PSIC score (Figures 2,3,4 in the manuscript, Supplementary figures 6,7,8). We also computed the Precision and Recall of the PSIC score at the threshold where the total number of errors was at a minimum (the minimum error point). PolyPhen was

able to classify ~50% of our training set mutations and ~60% of the test sets used to evaluate TP53 and EGFR mutations (Supplementary Tables 4,5).

SIFT

SIFT provides a score that ranges between 0 and 1 to report the probability that a missense mutation will be tolerated. Although the default threshold for an intolerant mutation is <= 0.05, we assessed SIFT with threshold-independent ROC and PR curves (Figures 2,3,4 in the manuscript, Supplementary figures 6,7,8). Only mutations with median score < 3.25 (a measure of sufficient diversity within the SIFT-constructed multiple sequence alignment) were considered, as recommended by the authors of SIFT (23). SIFT was able to classify approximately ~42% of our training set mutations and ~60% of the test sets used to evaluate TP53 and EGFR mutations (Supplementary Tables 4,5).

SIFT/PolyPhen consensus

We used the default SIFT threshold (<=0.05) to define a consensus between SIFT and PolyPhen. A mutation is classified as a driver by the SIFT/PolyPhen consensus if it has SIFT score <= 0.05 and a PolyPhen designation of "Possibly Damaging" or "Probably Damaging"). The SIFT/PolyPhen consensus was able to classify 22% of our training set mutations and 42-51% of the test sets used to evaluate TP53 and EGFR mutations (Supplementary Table 5).

CanPredict

CanPredict classifies a missense mutation as "likely-cancer", "likely-not-cancer" or "not-determined". The method combines several predictors: SIFT score, LogRE score (24) and GOSS score (25). These predictors are used as features to train a Random Forest classifier, using a training set of missense mutations found in the COSMIC database and >20% minor allele frequency SNPs. CanPredict was able to classify 45% of our training set mutations and 52-57% of the test sets used to evaluate TP53 and EGFR mutations. Precision and Recall of CanPredict are shown in Supplementary Table 5. These estimates may be overly optimistic, as the driver mutations in these datasets are primarily those documented in the COSMIC database, and may be included in the CanPredict training set.

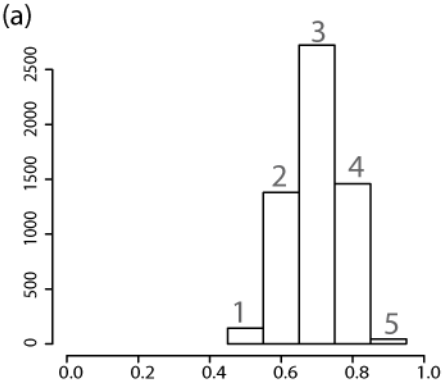
KinaseSVM

We also assessed a method for classifying driver and passenger missense mutations (21) that is specialized for protein kinases. We refer here to the method as KinaseSVM, because it uses a support vector machine trained on a wide variety of disease SNPs occurring in protein kinases (26). We obtained scores for mutations in our training set that occur in kinases, courtesy of Ali Torkamani. It was not possible to obtain scores for all of the kinases in the training set, because of difficulties mapping some mutations onto KinBase identifiers (http://kinase.com/kinbase/). For some of these mappings, the KinBase version of the sequence differed from our gene transcripts and we were not able to pinpoint the residue position where the mutation occurred. We were able to obtain KinaseSVM scores for 218 driver mutations and 123 synthetic passenger mutations from our training set (Supplementary Figure 5), yielding area under the ROC curve of

0.71 and area under the PR curve of 0.81. Although KinaseSVM was able to classify 112 out of 133 of the EGFR mutations in our test set, it was only able to classify 29 out of 590 synthetic passenger mutations in the test set, and was not applicable to TP53 (Supplementary Table 5), which was not sufficient to produce meaningful ROC and PR curves.

Supplementary Figure 1: (a) The continuous-valued feature "17-way exon conservation" measures the average PhastCons score (4) of DNA bases in an exon. A histogram analysis partitions the values of this feature, in our training set of driver and synthetic passenger mutations, into five categories. (b) A toy list of mutations in the training set. Each mutation is shown with a transcript or protein identifier, its reference and variant amino acid residue and position in protein sequence. Identifiers shown are from the databases UniProtKB (initial letter P), RefSeq (initial letters NP) and CCDS (8, 10). Each mutation is associated with a class label and a feature category. (c) A contingency table is used to estimate the joint probability density function *p(X,Y)* of a feature X (17-way exon conservation) and class labels Y (driver=1 and passenger=2) and the marginal probability density functions *p(X)* and *p(Y)*. (d) An example of a contingency table for a data set where "17-way exon conservation" is a perfect predictor of mutation class. The feature is always low (categories 1,2 or 3) for passengers and high (categories 4 or 5) for drivers. Thus *p(X,Y)* is always 0 or *p(X)*. In this scenario, mutual information between feature and class labels is at a maximum.

(a)



(b)

Mutation	Class	Feature
P10721_N566D	1	4
P01111_G12V	1	4
P08581_M1268T	1	5
NP_001667_H308Q	2	2
CCDS11845.1_E31K	2	1
CCDS12816.1_L123R	2	3

(c) p(x,y)

Y \ X	1	2	3	4	5	TOTAL	
1	0.0006	0.0305	0.0962	0.1755	0.0526	0.3553	p(y)
2	0.0308	0.1516	0.2445	0.1838	0.0339	0.6447	
TOTAL	0.0312	0.1821	0.3409	0.3595	0.0864	1.0000	
	p(x)						

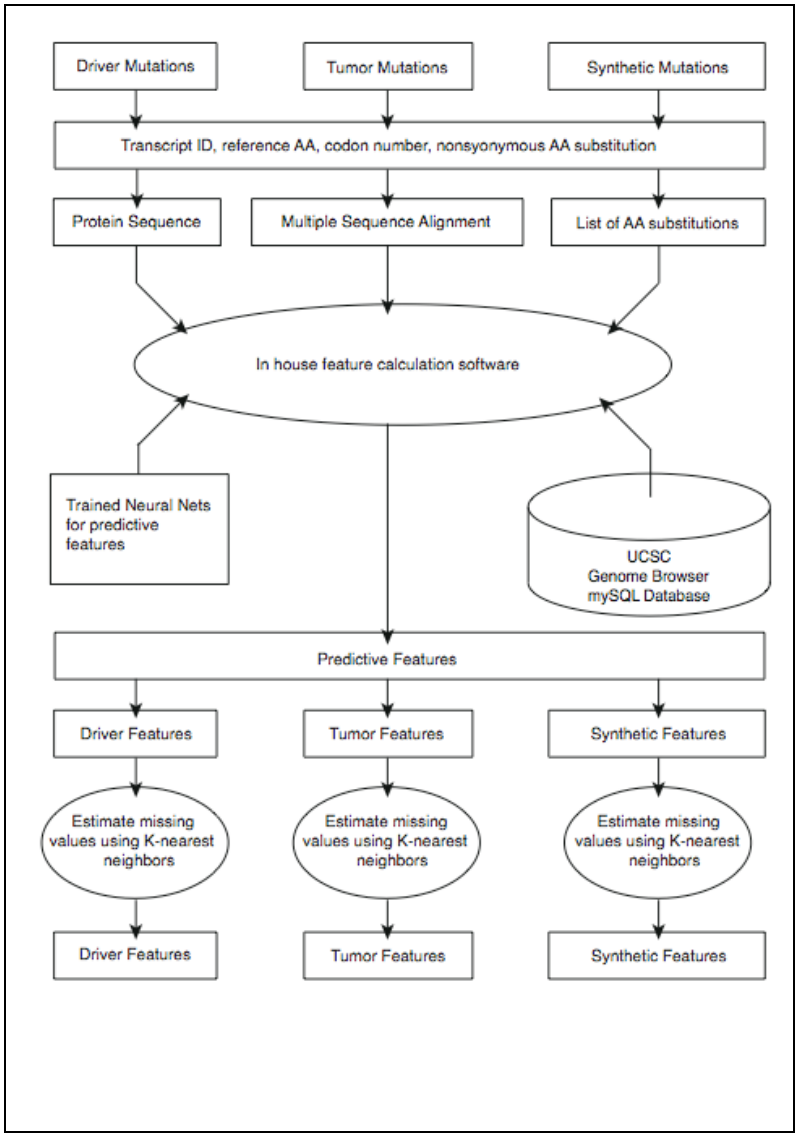
(d) p(x,y)

Y \ X	1	2	3	4	5	TOTAL	
1	0.0000	0.0000	0.0000	0.3595	0.0864	0.4459	p(y)
2	0.0312	0.1821	0.3409	0.0000	0.0000	0.5541	
TOTAL	0.0312	0.1821	0.3409	0.3595	0.0864	1.0000	
	p(x)						

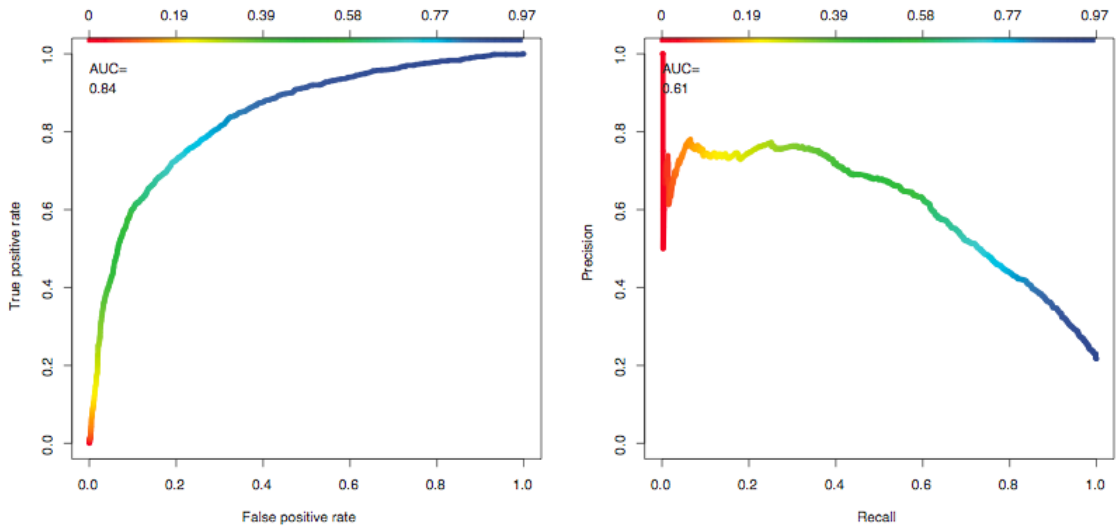
Supplementary Figure 2: *Pseudocode outlining the algorithm used to generate synthetic passenger mutations.* Mutations are generated in a tumor type and context specific fashion. Given tumor mutation data we determine the context in which the mutation occurred (Supplementary Table 2) and generate a new mutation sampled from the probability distribution, corresponding to that context at a randomly selected position in the same transcript.

```
SYNTHETIC_MUTATE_TRANSCRIPT(transcriptSequence)
realContextList ← NULL
syntheticMissenseMutationList ← NULL
for i = 1 to length(transcripSequence) do
  realContextList[i] ← GET_CONTEXT(transcriptSequence, i)
end for
for i = 1 to length(transcriptSequence) do
  if transcriptSequence[i] is a non-silent mutation then
    syntheticMissenseMutationList ← SYNTHETIC_MUTATE(i, realContextList)
  end if
end for
RETURN syntheticMissenseMutationList

GET_CONTEXT(transcriptSequence,i)
context ← NULL
if transcriptSequence[i] is C then
  if transcriptSequence[i + 1] is G then
    context ← CinCpG
  else if transcriptSequence[i − 1] is T then
    context ← CinTpC
  else
    context ← C
  end if
end if
if transcriptSequence[i]=G then
  if transcriptSequence[i − 1]=C then
    context ← GinCpG
  else if transcriptSequence[i + 1]=A then
    context ← GinGpA
  else
    context ← G
  end if
end if
if transcriptSequence[i]=A then
  context ← A
end if
if transcriptSequence[i]=T then
  context ← T
end if
RETURN context

SYNTHETIC_MUTATE(i, contextList)
mutantLocation ← randomSelect((contexts in contextList) == contextList[i])
baseChange ← randomSelect(contextList[i] multinomial)
syntheticMutant ← mutantLocation + baseChange
RETURN syntheticMutant
```

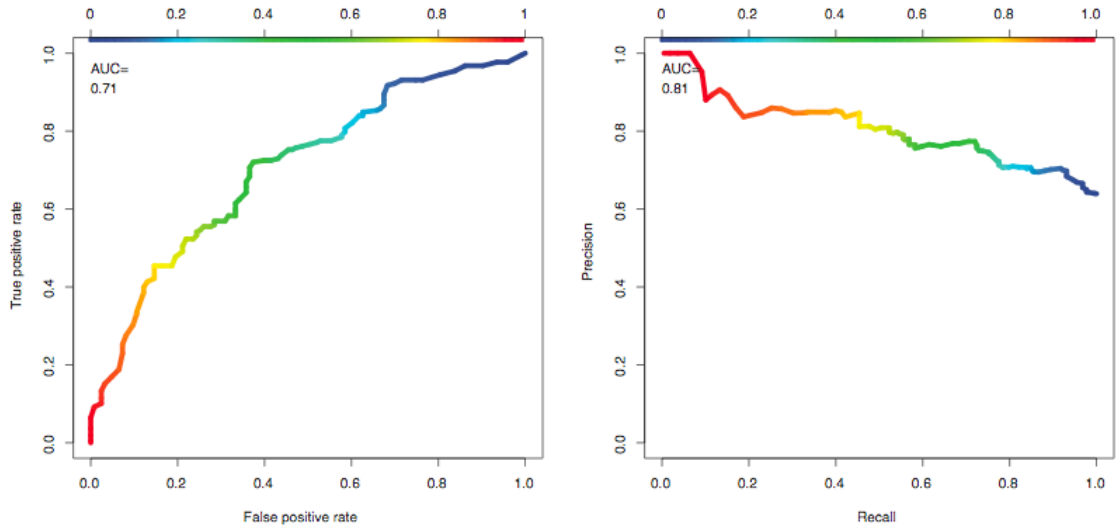
Supplementary Figure 3: *Flow Chart of CHASM's feature-building process.* Each mutation must be identified by a valid RefSeq ID, CCDS ID, ensembl ID or UniProt accession number. Protein sequence is retrieved for each transcript, and used to construct a multiple sequence alignment. Our in-house software uses this information, augmented by various amino acid substitution matrices, the UCSC Human Genome Browser and several trained neural networks to calculate our predictive features. Once features have been calculated, mutations are grouped according to their origin: known driver mutations, synthetic passenger mutations or unlabeled tumor mutations. After the data is segregated into these distinct sets, missing values are filled using the k-nearest neighbors algorithm.



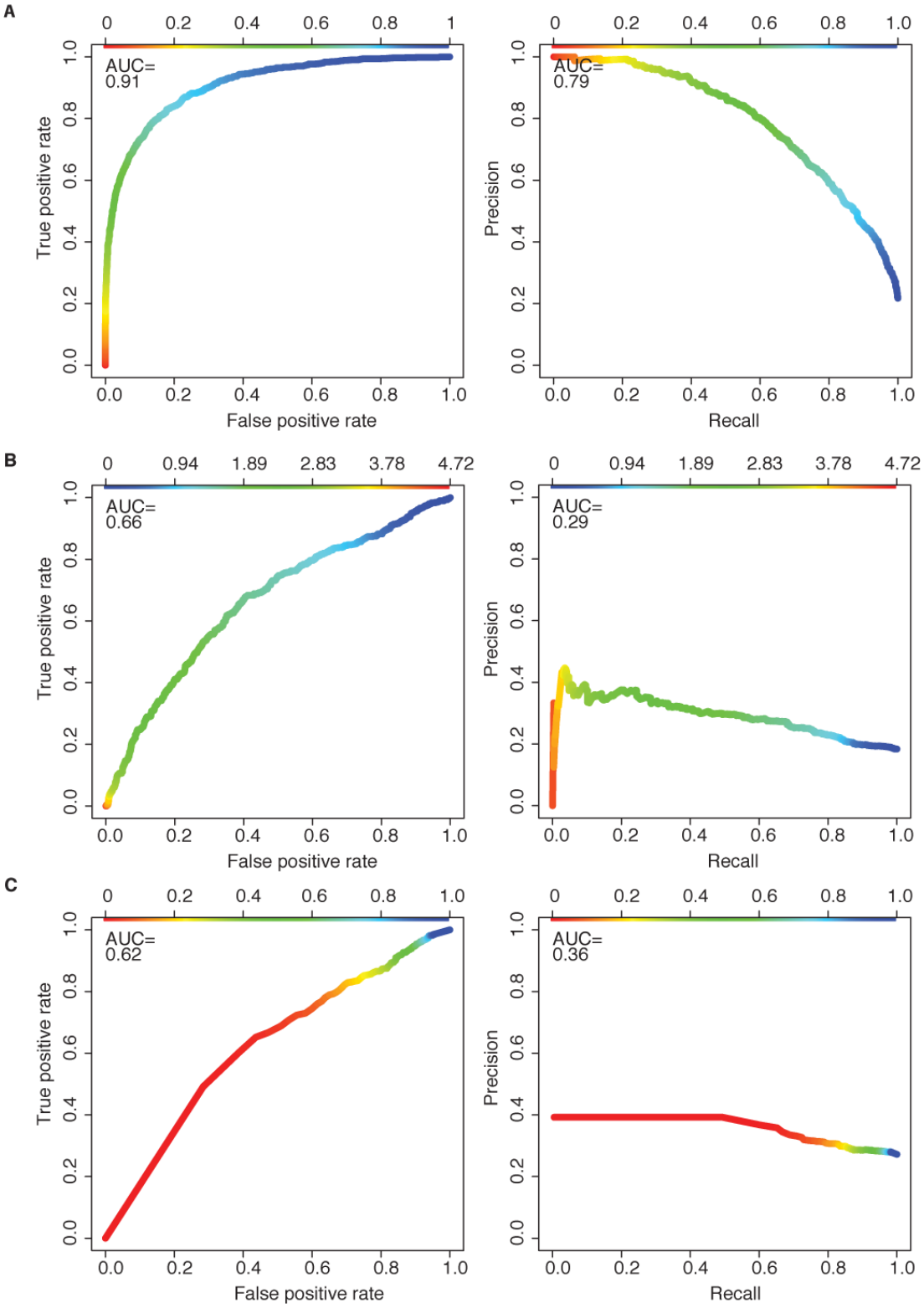
Supplementary Figure 4: *ROC and PR curves for a support vector machine trained on the predictive features used by CHASM.* Random Forest outperformed the support vector machine on our data and was selected for use in CHASM (Random Forest ROC AUC=0.91 and PR AUC=0.79 as compared to SVM ROC AUC=0.84 of and PR AUC=0.61.)



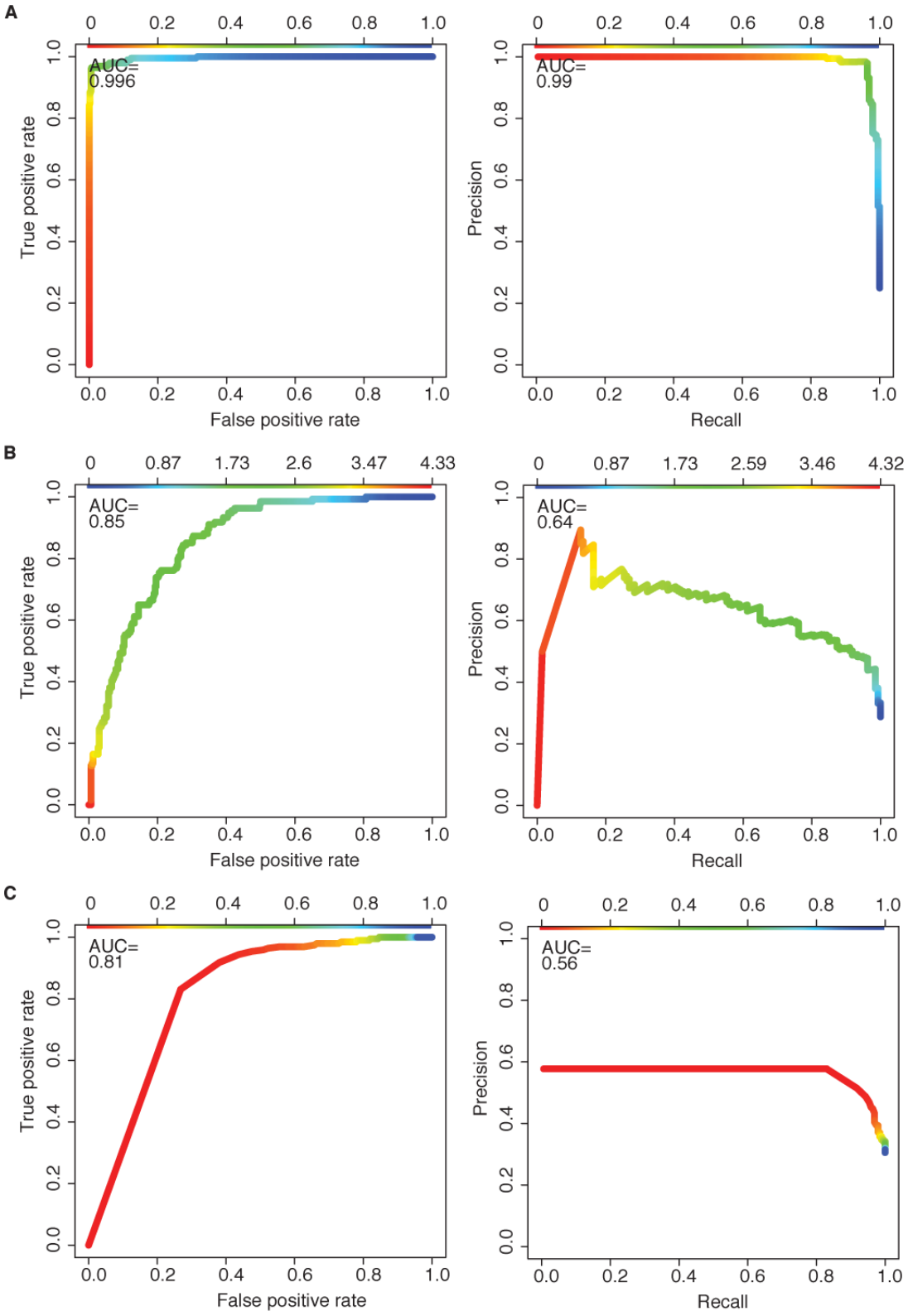
Supplementary Figure 5: *ROC and PR curves calculated from KinaseSVM scores on the CHASM training set.* We were able to obtain KinaseSVM scores for 218 (out of 1248) of the training drivers and 123 (out of 4500) of the training passengers. We estimate the area under the curve (AUC) for both the ROC curve (AUC = 0.71) and PR curve (AUC = 0.81) constructed based o these scores.



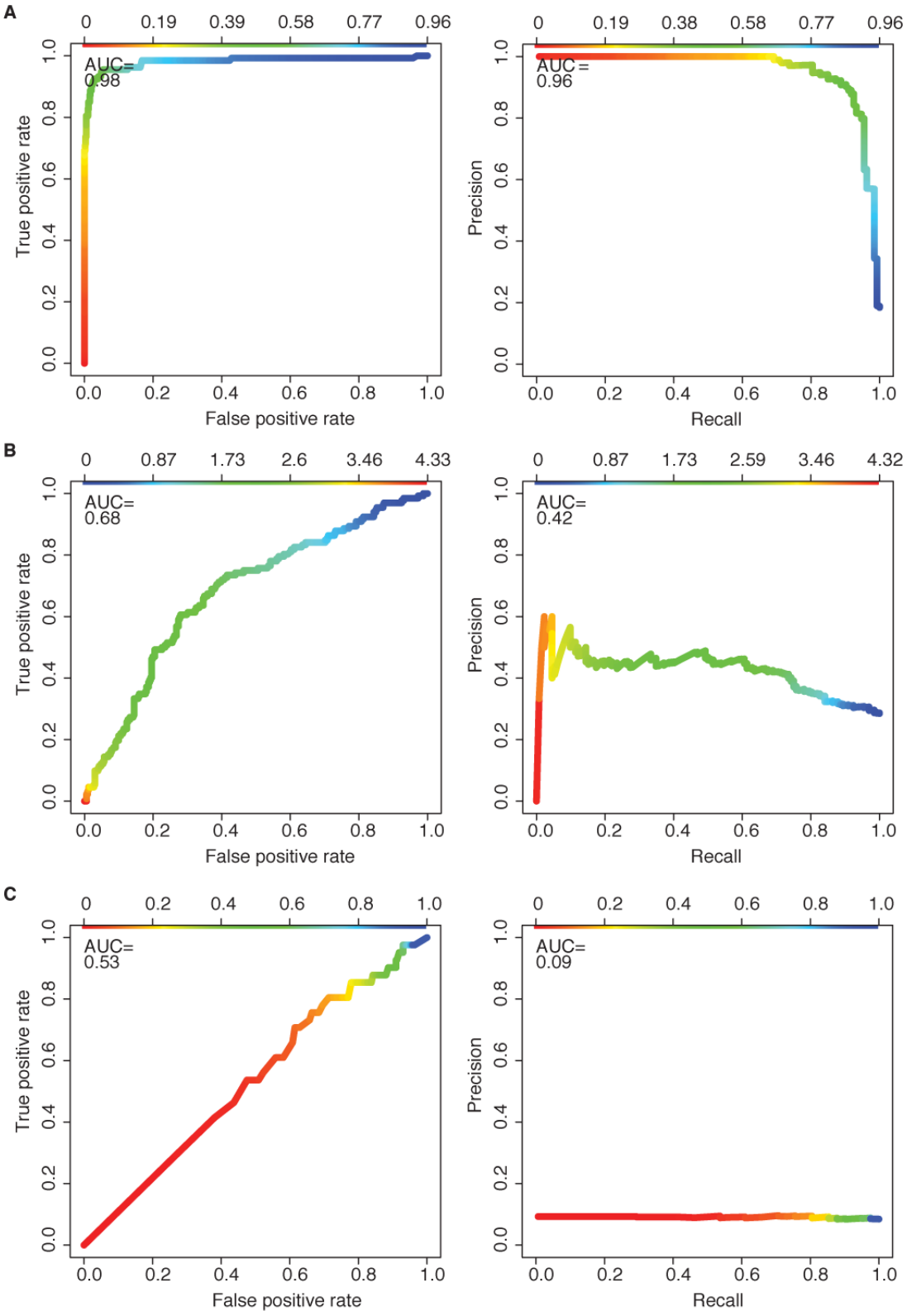
Supplementary Figure 6: *ROC and PR curves calculated for A) CHASM, B) PolyPhen PSIC, and C) SIFT on the training set mutations*. CHASM training out-of-bag scores were used to generate the ROC and PR curves in A). The color keys at the top of the images indicate the score thresholds used to calculate each point on the curve.



Supplementary Figure 7: *ROC and PR curves calculated for A) CHASM, B) PolyPhen PSIC, and C) SIFT on TP53 and synthetic passenger mutations held out of the CHASM training set.* The color keys at the top of the images indicate the score thresholds used to calculate each point on the curve.



Supplementary Figure 8: *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.* The color keys at the top of the images indicate the score thresholds used to calculate each point along the curve.



Supplementary Tables

Supplementary Table 1: Candidate predictive features that were initially evaluated for use in the CHASM classifier. The most informative features that are currently being used in CHASM are highlighted in yellow in Supplementary Table 3.

	Feature	Description
1	Net residue charge change	The change in formal charge resulting from the mutation. Histidine is assumed protonated (formal charge of +1)
2	Net residue volume change	The change in residue volume resulting from the mutation (27).
3	Net residue hydrophobicity change	The change in hydrophobicity resulting from the substitution (28).
4	Positional Hidden Markov model (HMM) conservation score	This feature is calculated based on the degree of conservation of the residue estimated from a multiple sequence alignment built with SAM-T2K software (29), using the protein in which the mutation occurred as the seed sequence (30). The SAM-T2K alignments are large, superfamily-level alignments that include distantly related homologs (as well as close homologs and orthologs) of the protein of interest.
5	Entropy of HMM alignment	The Shannon entropy calculated for the column of the SAM-T2K multiple sequence alignment, corresponding to the location of the mutation (31).
6	Relative entropy of HMM alignment	Difference in Shannon entropy calculated for the column of the SAM-T2K multiple sequence alignment (corresponding to the location of the mutation) and that of a background distribution of amino acid residues computed from a large sample of multiple sequence alignments (31).
7	Compatibility score for amino acid substitution in the column of a multiple sequence alignment of orthologs	These multiple sequence alignments are calculated using groups of orthologous proteins from the OMA database (32), which are aligned with T-Coffee software(33). The compatibility score for the mutation in the column of interest is computed from a large sample of multiple sequence alignments (31).
8	Grantham Score	The Grantham substitution score for the wild type to mutant transition (34).
9-11	Predicted residue solvent accessibility	These features consist of the probability of the wild type residue being buried, intermediate or exposed as predicted by a neural network trained with Predict-2nd software (29) on a set of 1763 proteins with high-resolution X-ray crystal structures sharing less than 30% homology (35).
12-14	Predicted contribution to protein stability	These features consist of the probability that the wild type residue contributes to overall protein stability in a manner that is highly stabilizing, average or destabilizing, as predicted by a neural network trained with Predict-2nd software (29) on a set of 1763 proteins with less than 30% homology. Stability estimates for the neural network training data were calculated using the FoldX force field (36, 37).
15-17	Predicted flexibility (Bfactor)	These features consist of the probability that the wild

		type residue backbone is stiff, intermediate or flexible as predicted by a neural network trained with Predict-2nd software (29) on a set of 1763 proteins with less than 30% homology. Flexibilities for the neural net training data were estimated based on normalized temperature factors, computed using the method of (38) from the X-ray crystal structure files.
18-20	Predicted secondary structure	These features consist of the probability that the secondary structure of the region in which the wild type residue exists is helix, loop or strand as predicted by a neural net trained with Predict-2nd software (29) on a set of 1763 proteins with crystal structures and with less than 30% homology.
21	Change in hydrophobicity	Change in residue hydrophobicity due to the wild type to mutant transition.
22	Change in volume	Change in residue volume due to the wild type to mutant transition.
23	Change in charge	Change in residue charge due to the wild type to mutant transition. Histidine is assumed neutral.
24	Change in polarity	Change in residue polarity due to the wildtype to mutant transition calculated in (34)
25	EX substitution score	Amino acid substitution score from the EX matrix (37).
26	PAM250 substitution score	Amino acid substitution score from the PAM250 matrix (39).
27	BLOSUM 62 substitution score	Amino acid substitution score from the BLOSUM 62 matrix (40).
28	MJ substitution score	Amino acid substitution score from the Miyazawa-Jernigan contact energy matrix (37, 41).
29	HGMD2003 mutation count	Number of times that the wild type to mutant substitution occurs in the Human Gene Mutation Database, 2003 version (25, 30, 31).
30	VB mutation count	Amino acid substitution score from the VB (Venkatarajan and Braun) matrix (37, 42).
31-33	Probability of seeing the wild type residue in the first, middle, or last position of an amino acid triple	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB* (10).
34-36	Probability of seeing the mutant residue in the first, middle, or last position of an amino acid triple	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB* (10).
37-39	Difference in probability of seeing the wildtype vs. the mutant reside in the first, middle, or last position of an amino acid triple	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB* (10).
40	Background probability of wildtype residue in UniProtKB* human proteins	Estimated as frequency of amino acid residue type occurrence.
41	Background probability of mutant residue in UniProtKB* human proteins	Estimated as frequency of amino acid residue type occurrence.
42	Probability of seeing the wild type at the center of a window of 5 amino acid residues	Calculated by a Markov chain of amino acid quintuples in human proteins found in UniProtKB* (10).
43	Probability of seeing the mutant at the center of a window of 5 amino acid residues	Calculated by a Markov chain of amino acid quintuples in human proteins found in UniProtKB* (10).

44-46	Frequency of missense change type in the Catalog of Somatic Mutations in Cancer (COSMIC) database	Frequency that missense change type (amino acid type X to amino acid type Y, e.g. ALANINE to GLYCINE) is seen in COSMIC. These frequencies were calculated during the week of August 14, 2008, using COSMIC release 38 (43) and normalized by the occurrences of the wild type residue in human proteins found in UniProtKB* (10), the occurrences of the wild type reside in cosmic or the number of times the change type is observed in the HapMap SNPs database (44).
47-55	Regional AA composition	The percentage of amino acids in a 15 residue window surrounding the mutation that fall into one of the following categories (P,C,G,DE,Q,H,KR,WYF,ILVM).
56	17way exon conservation	The conservation score for the entire exon calculated from a 17-species phylogenetic alignment using the UCSC Genome Browser (45). Scores are given for windows of nucleotides. We retrieve the scores for each region that overlaps the exon in which the base substitution occurred and calculated a weighted average of the conservation scores where the weight is the number of bases with a particular score.
57-59	SNP Density	The number of genetic variants, polymorphisms or verified HapMap SNPs (44) in the exon where the mutation is located
60-80	UniProt Annotations (fingerprints)	These features give annotations, curated from the literature, of general binding sites, general active sites, lipid, metal, carbohydrate, DNA, phosphate and calcium binding sites, disulfides, modified residues, propeptide residues, signal peptide residues, known mutagenic sites, transmembrane regions, compositionally biased regions, repeat regions, known motifs, and zinc fingers. The integer 1 indicates that a feature is present and the integer 0 indicates that it is absent at a mutated position

Supplementary Table. 2:

Synthetic Mutations were generated from eight multinomial distributions that depend on both tumor type and DNA context. The columns of the table show the eight contexts for each wild type DNA base and the rows show the (multinomial) probability distributions of base substitutions in GBM, based on (46).

Glioblastoma Multiforme (GBM)

	C in CpG	G in CpG	C in TpC	G in GpA	A	C	G	T
A	0.05	0.97	0.31	0.44	0.00	0.29	0.50	0.39
C	0.00	0.02	0.00	0.22	0.13	0.00	0.13	0.39
G	0.02	0.00	0.21	0.00	0.62	0.20	0.00	0.22
T	0.93	0.01	0.48	0.33	0.25	0.51	0.37	0.00

Supplementary Table 3. 80 candidate predictive features ranked according to their mutual information (in units of bits) with respect to driver and passenger classes. Detailed feature descriptions are in Supplementary Table 1. FP = fingerprint (a binary feature that takes on values of either 0 or 1).

Rank	Abbreviated Name	Feature	Mutual Information	Rank	Abbreviated Name	Feature	Mutual Information
1	17-Way Exon Conservation	56	0.0611	41	FP14 Signal Peptide Domain	64	0.00199
2	COSMIC subst frequency	45	0.0267	42	FP8 NTP Binding Domain	61	0.00197
3	FP30 PTM Enzyme Domain	80	0.026	43	Pred 2ndary Structure: Helix	18	0.00185
4	COSMIC	44	0.0258	44	FP13 Propeptide Domain	63	0.00172
5	PAM250 substitution score	26	0.0203	45	Pred 2ndary Structure: Strand	20	0.00134
6	JM substitution score	28	0.0202	46	FP27 Membrane Binding DM	77	0.00131
7	FP7 DNA Binding Domain	60	0.018	47	Difference in hydrophobicity	21	0.00126
8	VB substitution count	30	0.0178	48	Pred backbone flex: Low	15	0.00124
9	Positional HMM_Cons.	4	0.0168	49	Plastwt	38	0.00122
10	SNPDensity –all variants	57	0.0152	50	pdiff_last	33	0.0011
11	SNPDensity – validated only	58	0.0152	51	FP16 Domain contains variants	66	0.00106
12	Rel. Entropy of alignment	6	0.0152	52	Grantham substitution score	7	0.00104
13	Ex substitution score	25	0.0141	53	FP18 Domain has comp bias	68	0.000995
14	Entropy of alignment	5	0.0135	54	Region Composition H	52	0.000907
15	HGMD substitution count	29	0.0123	55	FP23 Protein-Protein Inter. DM	73	0.000784
16	BLOSUM substitution score	27	0.00872	56	Plastmut	39	0.000709
17	pdiff_middle	32	0.00723	57	FP15_Mutagen	65	0.000642
18	Background prob of WT res	40	0.00682	58	p5resmut	43	0.000478
19	Background prob of mut res	41	0.00527	59	FP26 Localization/Transport	76	0.000385
20	Pfirstmut	35	0.00495	60	Pred 2ndary structure: Loop	19	0.000371
21	Difference in polarity	24	0.0049	61	FP25 Transcription Factor Dom	75	0.000343
22	Pred solvent access:Intermed	10	0.0044	62	Region Composition KR	53	0.000283
23	Change in hydrophobicity	3	0.00433	63	FP29 PTM Recognition Dom.	79	0.000261
24	OMA alignment score	8	0.00376	64	Pred backbone flex: High	17	0.000194
25	Charge change (H neutral)	23	0.00332	65	Region Composition DE	50	0.000133
26	Pred backbhone flex: Med	16	0.00331	66	Region Composition Q	51	9.59E-05
27	COSMICvsHAPMAP	46	0.00331	67	FP20 Region Contains Motif	70	2.62E-05
28	Volume change	2	0.00307	68	SNPDensity hapmap only	59	0
29	Pred solvent access:Exposed	11	0.00292	69	FP9 CA Binding	62	0
30	Volume difference	22	0.00282	70	FP28 Chromatin Domain	78	0
31	Pred solvent access:Buried	9	0.00282	71	Charge change (H protonated)	1	-0.000187
32	FP24 RNA Binding	74	0.00253	72	FP19 Region Contains Repeats	69	-0.000345
33	FP22_REGION	72	0.00252	73	Region Composition C	48	-0.000359
34	p5reswt	42	0.00237	74	FP21 Zinc Finger Domain	71	-0.000638
35	FP17 Transmembrane	67	0.00234	75	pmiddlewt	36	-0.000728
36	Pfirstwt	34	0.00231	76	Region Composition WYF	54	-0.000822
37	Region Composition G	49	0.00231	77	Region Composition ILVM	55	-0.000926
38	Pmiddlemut	37	0.00226	78	Pred stability @ res: Low	12	-0.00139
39	pdiff_first	31	0.00213	79	Pred stability @ res: Med	13	-0.00147
40	Region Composition_P	47	0.00205	80	Pred stability @ res: High	14	-0.00226

Supplementary Table 4. *Comparison of CHASM with other methods for missense mutant function prediction*. Performance of each method is shown at its minimum error point. Relative Coverage is the fraction of mutations in the CHASM training set (5749 mutations) that could be classified by each method. AUCs could not be calculated for SIFT/PolyPhen Consensus (Methods).

	CHASM				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	NA	0.82	0.58	0.91	0.79
TP53 test set	NA	0.98	0.97	0.996	0.99
EGFR test set	NA	0.92	0.88	0.98	0.96
	PolyPhen PSIC score				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	56%	0.33	0.003	0.66	0.29
TP53 test set	59%	0.64	0.65	0.85	0.64
EGFR test set	64%	0.57	0.1	0.68	0.42
	SIFT Score				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	42%	0.39	0.49	0.62	0.36
TP53 test set	81%	0.58	0.83	0.81	0.56
EGFR test set	67%	0.09	0.29	0.53	0.09
	SIFT/PolyPhen Consensus				
	Relative Coverage	Precision	Recall	AUC -ROC	AUC -PR
Training set	22%	0.44	0.59	NA	NA
TP53 test set	51%	0.57	0.96	NA	NA
EGFR test set	42%	0.13	0.38	NA	NA

Supplementary Table 5. Method Comparison Statistics. Methods are assessed by their coverage and performance on three datasets, namely the CHASM training drivers and passengers, held out TP53 mutations and passengers and held out EGFR mutations and passengers. The same set of held out passengers, distinct from those used in the training set, are used in both the TP53 and EGFR test sets.

	PolyPhen PSIC score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	590/1248	2623/4500	0.33	0.003	0.66	0.29
TP53 test set	133/196	332/590	0.64	0.65	0.85	0.64
EGFR test set	132/133	332/590	0.57	0.1	0.68	0.42
	SIFT Score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	650/1248	1754/4500	0.39	0.49	0.62	0.36
TP53 test set	195/196	444/590	0.58	0.83	0.81	0.56
EGFR test set	41/133	444/590	0.09	0.29	0.53	0.09
	SIFT/PolyPhen Consensus					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	428/1248	858/4500	0.44	0.59	NA	NA
TP53 test set	133/196	266/590	0.57	0.96	NA	NA
EGFR test set	40/133	266/590	0.13	0.38	NA	NA
	CanPredict Score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	685/1248	1885/4500	0.62	0.63	NA	NA
TP53 test set	190/196	259/590	0.8	0.995	NA	NA
EGFR test set	114/133	259/590	0.58	0.55	NA	NA
	KinaseSVM Score					
	Coverage		Precision	Recall	AUC-ROC	AUC-PR
	drivers	passengers				
Training set	218/1248	123/4500	0.7	0.92	0.71	0.81
TP53 test set	NA	NA	NA	NA	NA	NA
EGFR test set	112/133	29/590	0.92	0.8	0.71	0.95

References

1. Karchin R, Kelly L, Sali A. Improving functional annotation of non-synonomous SNPs with information theory. Pac Symp Biocomput 2005:397-408.
2. Wolpert, Wolf. Estimating functions of probability distributions from a finite set of samples. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 1995;52(6):6841--54.
3. Cover T, Thomas J. Elements of information theory. 1st ed: Wiley and Sons; 1991.
4. Siepel A, Bejerano G, Pedersen JS*, et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15(8):1034-50.
5. Blanchette M, Kent WJ, Riemer C*, et al.* Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 2004;14(4):708-15.
6. Hinrichs AS, Karolchik D, Baertsch R*, et al.* The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 2006;34(Database issue):D590--D8.
7. R Core Development Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2008.
8. Sayers EW, Barrett T, Benson DA*, et al.* Database resources of the National Center for Biotechnology Information. Nucl Acids Res 2009;37(suppl_1):D5-15.
9. Hubbard TJP, Aken BL, Ayling S*, et al.* Ensembl 2009. Nucl Acids Res 2009;37(suppl_1):D690-7.
10. Wu CH, Apweiler R, Bairoch A*, et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. 2006.
11. Fawcett T. ROC graphs: Notes and practical considerations for researchers. Machine Learning 2004;31.
12. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39(4):561-77.
13. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. 2006: ACM New York, NY, USA; 2006. p. 233-40.
14. 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) 1995:289-300.
15. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 2001;96(456):1151-60.
16. Parzen E. On estimation of a probability density function and mode. The Annals of Mathematical Statistics 1962:1065-76.
17. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100(16):9440-5.
18. Yan H, Parsons DW, Jin G*, et al.* IDH1 and IDH2 Mutations in Gliomas. The New England Journal of Medicine 2009;360(8):765.
19. Breiman L. Random forest. Machine Learning 2001;45:5-32.
20. Karchin R, Agarwal M, Sali A, Couch F, Beattie MS. Classifying Variants of Undetermined Significance in BRCA2 with Protein Likelihood Ratios. Cancer informatics 2008;6:203.
21. Torkamani A, Schork NJ. Prediction of cancer driver mutations in protein kinases. Cancer Res 2008;68(6):1675--82.
22. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Oxford Univ Press; 1999. p. 387-94.
23. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11(5):863-74.
24. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Oxford Univ Press; 2004. p. 1006-14.
25. Kaminker JS, Zhang Y, Waugh A*, et al.* Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms. Cancer Res 2007;67(2):465-73.
26. Torkamani A, Schork NJ. Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family. Genomics 2007;90(1):49--58.
27. Zamyatnin AA. Protein volume in solution. Prog Biophys Mol Biol 1972;24:107-23.
28. Engelman DM, Steitz TA, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 1986;15:321-53.

29. Karplus K, Karchin R, Barrett C*, et al.* What is the value added by human intervention in protein structure prediction? Proteins 2001;Suppl 5:86-91.
30. Karchin R, Diekhans M, Kelly L*, et al.* LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 2005;21(12):2814-20.
31. Kullback S. Information theory and statistics. New York: Wiley; 1959.
32. Schneider A, Dessimoz C, Gonnet GH. OMA Browser--exploring orthologous relations across 352 complete genomes. Bioinformatics 2007;23(16):2180-2.
33. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302(1):205-17.
34. Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185(4154):862-4.
35. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589-91.
36. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res 2005;33(Web Server issue):W382-8.
37. Yampolsky LY, Stoltzfus A. Untangling the effects of codon mutation and amino acid exchangeability. Pac Symp Biocomput 2005:433-44.
38. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12(5):1060-72.
39. Schwartz RM, Dayhoff MO. Improved scoring matrix for identifying evolutionary relatedness among proteins. Biophys J 1978;21(3):A198-A.
40. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89(22):10915-9.
41. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18(3):534-52.
42. Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. Journal of Molecular Modeling 2001;7(12):445-53.
43. Forbes S, Clements J, Dawson E*, et al.* Cosmic 2005. Br J Cancer 2006;94(2):318-22.
44. Schmidt CW. HapMap: building a database with blocks. EHP Toxicogenomics 2003;111(1T):A16.
45. Kent WJ, Sugnet CW, Furey TS*, et al.* The human genome browser at UCSC. Genome Res 2002;12(6):996--1006.
46. Parsons DW, Jones S, Zhang X*, et al.* An integrated genomic analysis of glioblastoma multiforme. Science 2008;321(5897):1807-12.