Prediction of Cancer Driver Mutations in Protein Kinases

Ali Torkamani[1,2,3] and Nicholas J. Schork[2,3]

[1]Graduate Program in Biomedical Sciences and [2]Center for Human Genetics and Genomics, University of California, San Diego, San Diego, California and [3]Scripps Genomic Medicine and The Scripps Research Institute, La Jolla, California

Abstract

A large number of somatic mutations accumulate during the process of tumorigenesis. A subset of these mutations contribute to tumor progression (known as "driver" mutations) whereas the majority of these mutations are effectively neutral (known as "passenger" mutations). The ability to differentiate between drivers and passengers will be critical to the success of upcoming large-scale cancer DNA resequencing projects. Here we show a method capable of discriminating between drivers and passengers in the most frequently cancer-associated protein family, protein kinases. We apply this method to multiple cancer data sets, validating its accuracy by showing that it is capable of identifying known drivers, has excellent agreement with previous statistical estimates of the frequency of drivers, and provides strong evidence that predicted drivers are under positive selection by various sequence and structural analyses. Furthermore, we identify particular positions in protein kinases that seem to play a role in oncogenesis. Finally, we provide a ranked list of candidate driver mutations. [Cancer Res 2008;68(6):1675–82]

Introduction

Cancers are derived from genetic changes that result in a growth advantage for cancerous cells. These genetic changes, or mutations, either occur as a result of errors during replication or may be induced by exposure to mutagens. More than 1% of all human genes are known to contribute to cancer as a result of acquired mutations (1). The family of genes most frequently contributing to cancer is the protein kinase gene family (1), which are both implicated in, and confirmed as drug targets for, a number of tumorigenic functions, including, immune evasion, proliferation, antiapoptotic activity, metastasis, and angiogenesis (2, 3). As mutations accumulate in a precancerous cell, some mutations confer a selective advantage by contributing to tumorigenic functions (known as "drivers"), whereas others are effectively neutral (known as "passengers"). Passenger mutations may occur incidentally because of mutational processes, and are often observed in the mature cancer cells, but are not ultimately responsible for any pathogenic characteristics exhibited by the tumor.

Recent systematic resequencing of the kinome in cancer cell lines has revealed that most somatic mutations are likely to be passengers that do not contribute to the development of cancers (4). A challenge posed by these systematic resequencing efforts is to differentiate between passenger and driver mutations. Differentiating passengers from drivers not only is critical for understanding the molecular mechanisms responsible for tumor initiation and progression but, ultimately, also provides prognostic and diagnostic markers as well as targets for therapeutic intervention. An effective method for identifying cancer drivers is also critical for customizing or individualizing the treatment of a cancer patient based on his or her specific tumorigenic profile. Currently, statistical models comparing nonsynonymous to synonymous mutation rates are used to both identify and estimate the number of possible cancer drivers of a total set of identified genetic variations (5). These methods are excellent for estimating the overall number and frequency distribution of potential drivers of a larger set of variations but do not have sufficient power or resolution to pinpoint particular drivers.

Recent evidence suggests that cancer drivers have characteristics similar to Mendelian disease mutations (6). Based on this information, a computational tool for predicting cancer-associated missense mutations, CanPredict, was developed (7). CanPredict is a generalized prediction method but is limited to predictions made on missense mutations falling within specific functional domains of proteins. We have recently developed a support vector machine (SVM)–based method to differentiate common, likely nonfunctional genetic variations from Mendelian disease-causing polymorphisms, specifically within the protein kinase gene family (8), and here we have applied this method to somatic cancer mutations.

We have evaluated the utility of this method in a number of ways. First, we show that our method outperforms CanPredict on classification of known drivers within the protein kinase gene family. Second, we show that our method shows excellent agreement with previous statistical estimates of the number of likely drivers observed in the resequencing study by Greenman et al. (i.e., 159 specific drivers versus 158 predicted drivers by our method). Third, we present sequence, structural, and frequency analyses of mutations catalogued within the Cosmic database (9), which strongly suggest that predicted driver mutations by our method are under positive selection during oncogenesis and are, in fact, true cancer drivers. Fourth, we identify specific positions, including a position corresponding to BRAF V600, whereby mutations at these positions are observed across eight different kinases, suggesting a generalized role for this position in mediating oncogenesis. A ranked list of candidate driver mutations, as well as suspected cancer predisposing germ-line mutations, is provided in Supplementary data.

Materials and Methods

Known somatic driver mutations were obtained by searching OMIM (10). Somatic and germ-line mutations from cancer cell lines were obtained from the kinome resequencing study by Greenman et al. (4). The catalogue of observed somatic mutations was obtained from the Cosmic database (9). Our protein kinase sequences and residue numbering correspond to the

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

Requests for reprints: Nicholas J. Schork, Scripps Genomic Medicine, The Scripps Research Institute, MEM-275A, 10550 North Torrey Pines Road, La Jolla, CA 92037. Phone: 858-784-2308; Fax: 858-784-2910; E-mail: nschork@scripps.edu.

doi:10.1158/0008-5472.CAN-07-5283

Table 1. Known cancer drivers and passenger

Kinase	Mutation	Driver?	Prediction	CanPredict
BRAF	R461I	Yes	Yes	Yes
BRAF	I462S	Yes	Yes	Yes
BRAF	G463E	Yes	Yes	Yes
BRAF	G465V	Yes	Yes	Yes
BRAF	L596R	Yes	Yes	Yes
BRAF	L596V	Yes	Yes	Yes
BRAF	V600E	Yes	Yes	Yes
BRAF	K600E	Yes	Yes	Yes
EGFR	G719C	Yes	Yes	Yes
EGFR	G719S	Yes	Yes	Yes
EGFR	**T790M**	Yes	Yes	No
EGFR	L858R	Yes	Yes	Yes
FGFR2	S267P	Yes	Yes	Yes
FGFR3	*R248C*	*Yes*	*Yes*	*ND*
FGFR3	*S249C*	*Yes*	*Yes*	*ND*
FGFR3	E322K	Yes	Yes	Yes
FGFR3	K650E	Yes	Yes	Yes
ErbB2	L755P	Yes	Yes	Yes
ErbB2	**G776S**	Yes	Yes	No
ErbB2	**N857S**	Yes	Yes	No
ErbB2	E914K	Yes	Yes	Yes
KIT	V559D	Yes	Yes	Yes
KIT	V560G	No	No	No
KIT	D816V	Yes	Yes	Yes
LKB1/STK11	Y49D	Yes	Yes	Yes
LKB1/STK11	G135R	Yes	Yes	Yes
PDGFRa	V561D	Yes	Yes	Yes
PDGFRa	D842V	Yes	Yes	Yes
RET	**M918T**	**Yes**	**Yes**	**No**

NOTE: Mutations incorrectly predicted by CanPredict are boldfaced. Mutations with no CanPredict predictions are italicized.
Abbreviation: ND, not determined.

position in KinBase[4] sequences (11). Single-nucleotide polymorphisms (SNP) were mapped to protein kinases by blasting KinBase sequences versus Cosmic database sequences (12). SNPs from the Cosmic database were assigned to KinBase sequences with the best *E* values and mapped to specific positions as described by Torkamani and Schork (13). SNPs mapping to Obscurin and Titin were filtered out because these proteins are currently unamenable to our prediction method. This filtering resulted in 563 SNPs from Greenman et al. and 1,036 SNPs from the Cosmic database.

Subdomain distribution and motif-based alignments of 175 kinase catalytic domains containing somatic mutations found within the Cosmic database were generated as described by Torkamani and Kannan.[5] Previously, motif-based alignments were generated by implementation of the Gibbs motif sampling method of Neuwald et al. (14, 15). Given a set of protein kinase sequences used to generate conserved motifs, as in Kannan et al. (16), the Gibbs motif sampling method identifies characteristic motifs for each individual subdomain of the kinase catalytic core, which are then used to generate high-confidence motif-based Markov chain Monte Carlo multiple alignments based on these motifs (17). These subdomains define the core structural components of the protein kinase catalytic core. Intervening regions between these subdomains were not aligned.

The quality of these alignments was assessed by the APBD (18) method using available crystal structures of human protein kinases that contained at least one cancer-associated somatic mutation (CASM). The sequences and crystal structures used in APBD were 1A9U (p38a), 1AQ1 (CDK2), 1B6C (TGFbR1), 1BI7 (CDK6), 1CM8 (p38g), 1QPJ (LCK), 1FGK (FGFR1), 1FVR (TIE2), 1GAG (INSR), 1GJO (FGFR2), 1GZN (AKT2), 1IA8 (CHK1), 1K2P (BTK), 1M14 (EGFR), 1MQB (EphA2), 1MUO (AurA), 1QCF (HCK), 1R1W (MET), 1RJB (FLT3), and 1U59 (ZAP70). The average alignment accuracy was 92%. After visual inspection of the multiple alignment score distribution, manual tuning of the alignments was deemed unnecessary. Score accuracy was evenly distributed across the entire alignment, suggesting no loss of alignment resolution at any particular region.

Calculations about the enrichment of somatic mutations within particular subdomains were executed as follows. The average length of each subdomain was calculated as the weighted average of the region length in each kinase considered, where weights correspond to the total number of SNPs occurring within each kinase. Although subdomains are generally of the same length, these weights are used to avoid biases in the length of intervening regions between subdomains (those labeled "a" in Table 2) due to the large inserts occurring in a few protein kinases. The probability of a SNP occurring within a particular region purely by chance was computed as its weighted average length over the sum of every region's weighted average length. The probability (*P* value) of the observed total number of SNPs occurring within each region was then calculated using the general binomial distribution. A simulation study to determine the significance of the position-specific distribution of CASMs was carried out by randomly placing the same number of SNPs observed in the Cosmic database per kinase 10,000 times. The results were used to determine the

[4] http://kinase.com/kinbase/
[5] Manuscript in preparation.

95% confidence interval of the expected number of sites where one to eight kinases would be expected to be mutated by chance.

Predictions were done as described by Torkamani and Schork (8). Briefly, a SVM was trained on common SNPs (presumed neutral) and congenital disease–causing SNPs characterized by a variety of sequence, structural, and phylogenetic variables. The SNP characteristics used to predict disease causing status were (*a*) kinase group; (*b*) wild-type amino acid; (*c*) SNP amino acid; (*d*) domain; (*e*) subPSEC score; (*f*) the change in hydrophobicity, polarity, and charge coded as 1, 0, or −1, where 1 is a gain in the respective factor, 0 is no change, and −1 is a loss in the respective factor; (*g*) the secondary structure coded as coil, helix, or sheet; (*h*) the solvent accessibility coded as accessible, inaccessible, or intermediate; (*i*) protein flexibility; and (*j*) the differences in the following characteristics: the five amino acid metrics, Kyte-Doolittle Hydropathy, water/octanol partition energy, and volume [described in detail by Torkamani and Schork (8)]. For mutations falling within the kinase catalytic domain, an additional eleventh predictor, whether the mutations fall within the NH_2-terminal or COOH-terminal lobe, was used. Predictions are done using somatic mutations occurring within and outside of the kinase catalytic core separately. As in Torkamani and Schork (8), the threshold taken for calling a SNP a driver is 0.49 for catalytic domain mutations and 0.53 for all other mutations.

The Ingenuity Pathway Analysis[6] tool was used to determine which pathways each protein kinase gene participates in. Standard least squares regression, with pathways as the independent variable and the SVM predicted probability that a polymorphism is deleterious as the dependent variable, was then applied to all germ-line mutations with the number of times a germ-line mutation is observed as its weight. All statistical analyses were done using JMP IN 5.1.[7]

Results

Prediction of Known Drivers and Comparison with Previous Methods

All known CASMs occurring within the kinase gene family were extracted from the Cosmic database. A nonredundant set of CASMs was generated from this data set and subjected to predictions by our SVM method. Within this data set of 1,036 CASMs, 512 (49.42%) were predicted to be driver mutations. The OMIM database contains a small number of these mutations that are known to be drivers and whose functional significance in sporadic, nonfamilial cases of cancer is supported by substantial evidence (Table 1). These 28 known driver mutations and 1 known passenger mutation are predicted with 100% accuracy by our SVM method. Given that 49.42% of the mutations within the CASMs data set are predicted to be driver mutations, this degree of accuracy for these 29 mutations can be expected to occur, at random, once in a billion. Given that most of these known driver mutations occur within the kinase catalytic core, and that mutations within the catalytic core are more likely to be predicted as driver mutations (74.50% of mutations within the catalytic core are predicted to be drivers), the probability with which this predictive accuracy can be expected at random, adjusted for the rate at which catalytic core mutants are predicted to be drivers, is $P = 6.71 \times 10^{-5}$, and thus is highly statistically significant. The performance of our method on this small subset of known cancer drivers suggests that predictions of drivers by our method are highly accurate. The performance of our method on the protein kinase gene family is also superior to that of CanPredict (7), a whole genome cancer driver prediction method (Table 1). CanPredict only performs predictions on the 27 SNPs falling within functional domains. Of these SNPs, four are incorrectly predicted as passengers.

Agreement with Resequencing-Based Predictions

Our SVM prediction technique was applied to 583 missense mutations identified by Greenman et al. (4) in cancer cell lines to identify which of these mutations are likely to be cancer drivers. One hundred fifty-nine missense mutations (28.24% of missense mutations) in 99 kinases were predicted to be cancer drivers (Supplementary Table S1). These figures show excellent agreement with the analysis of selection pressure using synonymous versus nonsynonymous mutational frequencies by Greenman et al., which suggested that 158 (95% confidence interval, 63–246) driver mutations in 119 kinase (95% confidence interval, 52–149) exist within this data set. The analysis by Greenman et al. revealed that selection pressure is only slightly higher within the catalytic domain (1.40) as compared with mutations outside this domain (1.23). Consistent with this finding, we predict that 66.67% of drivers fall within the catalytic domain, whereas the rest of the predicted drivers fall outside, especially within receptor structures (11.95%) and unstructured interdomain linker regions (13.84%). Within the kinase catalytic domain, Greenman et al. showed that mutations within the P-loops and activation segments showed a higher selection pressure (1.75) than the remainder of the catalytic domain. In agreement with their analysis, our method also predicts a higher proportion of drivers (64.29%) within these regions as opposed to the rest of the catalytic domain (44.63%; $P = 0.0258$).

Additionally, our SVM prediction technique was applied to germ-line mutations observed by Greenmen et al. to predict which mutations may underlie cancer predisposition. Interestingly, SNPs predicted to underlie inherited cancer predisposition were observed less often than those predicted to be neutral ($P = 0.0006$), suggesting that, potentially, a variety of rare

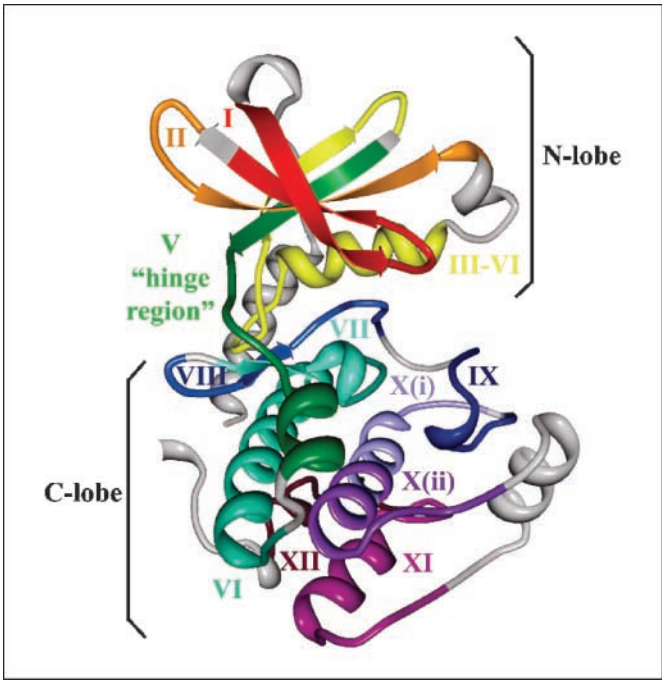

Figure 1. Subdomains mapped to PKA. The subdomains of PKA (PDB ID 1ATP) are colored and labeled by color-matched roman numerals. Obscuring COOH-terminal residues beyond subdomain XII have been removed.

[6] Ingenuity Systems, http://www.ingenuity.com.
[7] JMP IN 5.1 (SAS Inistitute, Inc.).

Table 2. Subdomain distribution of cancer SNPs

Subdomain	% Catalytic core	% SNPs	Ratio	Distribution *P*	% Driver	% Passenger	Regression *P*
I	6.32	11.09	1.75	<0.0001*	86.67%	13.33%	0.0038*
Ia	1.50	1.66	1.11	0.4505	88.89%	11.11%	0.0443*
II	5.38	5.18	0.96	0.4307	67.86%	32.14%	0.1319
Iia	2.00	2.59	1.30	0.1304	71.43%	28.57%	0.1289
III-IV	10.71	10.35	0.97	0.2202	73.21%	26.79%	0.0550
Iva	0.81	0.74	0.91	0.9657	75.00%	25.00%	0.2388
V	6.72	6.84	1.02	0.2053	81.08%	18.92%	0.0196*
Va	*5.82*	*2.40*	0.41	*0.0069**	*61.56%*	*35.29%*	*0.2897*
VI	7.46	6.28	0.84	0.9167	64.71%	35.29%	0.1699
VIa	0.07	0.18	2.57	0.5185	100.00%	0.00%	0.8334
VII	**5.69**	6.65	1.17	0.0426*	86.11%	13.89%	0.0076*
VIIa	0.73	0.92	1.26	0.4496	80.00%	20.00%	0.1554
VIII	**5.36**	16.82	3.14	**<0.0001***	87.91%	12.09%	0.0018*
VIIIa	**4.19**	9.98	2.38	**<0.0001***	83.33%	16.67%	0.0094*
IX	4.98	4.25	0.85	0.8983	82.61%	17.39%	0.0236*
Ixa	1.00	1.29	1.29	0.3139	71.43%	28.57%	0.7150
X(i)	3.91	2.03	0.52	0.1398	72.73%	27.27%	0.1342
X(ii)	5.55	3.33	0.60	0.1992	50.00%	50.00%	0.5567
X(ii)a	*7.52*	*2.77*	0.37	*0.0004**	*53.33%*	*46.67%*	*0.4716*
XI-XII	*11.79*	*3.33*	0.28	*<0.0001**	*27.78%*	*72.22%*	*0.6213*
XIIa	2.50	1.29	0.52	0.2701	14.29%	85.71%	0.3259

NOTE: Subdomains enriched in CASMs are boldfaced and subdomains devoid of CASMs are italicized. % Catalytic core, the fraction of the catalytic core composed of the individual subdomain; % SNPs, the percentage of CASMs occurring within the individual catalytic core; % Driver and % Passenger, the fraction of SNPs within the individual subdomain that are drivers or passengers, respectively. Subdomains are labeled by roman numerals; those followed by "a" correspond to intervening regions.
*Statistically significant. $P < 0.05$.

polymorphisms underlie inherited cancer predisposition (Supplementary Table S2). Furthermore, when pathway analysis is done (see Materials and Methods), the majority of identified pathways encompassing the genes that the predisposing variations are within seem to lend to a predisposition to developing cancer by reducing the effectiveness of the immune response or by allowing immune evasion. These pathways include toll-like receptor signaling ($P < 0.0001$), integrin signaling ($P = 0.0001$), transforming growth factor-β signaling ($P = 0.0143$), T-cell receptor signaling ($P = 0.0143$), and IFN signaling ($P = 0.0446$) pathways. This analysis suggests that immune deficiencies are a major mechanism underlying cancer predisposition.

Analyses of the Cosmic Database

Predicted drivers are observed frequently in different cancer samples. To further validate the accuracy of our SVM approach, we extracted a nonredundant set of CASMs occurring within the kinase gene family from the Cosmic database (9), noting the number of times each specific mutation is recorded within the database (9), and performed predictions on the CASMs using our SVM method. Within this data set of 1,036 CASMs, 512 (49.42%) were predicted to be driver mutations (Supplementary Table S3). We postulate that driver mutations are positively selected; if so, they should be observed within the Cosmic database more often than random passenger mutations. We compared the number of times predicted driver mutations (mean of 19.5 ± 9.4 observations of 512 SNPs) have been observed in cancer against predicted passenger mutations (mean of 1.4 ± 0.07 observations of 524 SNPs), using the nonparametric Wilcoxon rank sum test. Nonparametric analysis allows us to control for major outliers, such as the BRAF V600E mutation, which has been observed in cancer >3,000 times. The result of this analysis was that the predicted driver mutations (mean rank score, 559.8) are indeed observed more frequently than predicted passenger mutations (mean rank score, 478.14; standardized score, 5.41; $P < 0.0001$).

Subdomains enriched with CASMs are enriched with predicted drivers. Further validation was sought by generating multiple motif-based alignments of the kinase catalytic core and mapping cancer mutants to catalytic core subdomains and specific positions, as described by Torkamani and Kannan (Fig. 1; Supplementary Table S4).[8] A simulation study suggested that cancer mutations are not observed in a statistically significant position-specific manner, likely due to random noise generated by passenger mutations (see Materials and Methods). However, analysis of the subdomain distribution of cancer mutations using the method described by Torkamani and Kannan (see Materials and Methods) suggested that cancer mutations, regardless of the noise of passenger mutations, do show a bias in distribution throughout the catalytic core (Table 2, *left*). For example, subdomain I, containing the glycine loop, which is directly involved in ATP binding, and subdomains VII, VIII, and VIIIa, comprising the catalytic and activation loops, are significantly enriched for cancer-associated mutations, whereas subdomains Va, X(ii)a, and

[8] A. Torkamani, N. Kannan, S.S. Taylor, N.J. Schork, submitted for publication.

XI–XII, which are not directly involved in either ATP binding or catalysis, are significantly devoid of cancer-associated mutations. Suprisingly, the "hinge region" (subdoman V), involved in ATP binding, is not significantly enriched for cancer-associated mutations. However, mutations within this region are predominantly predicted as drivers (described below), suggesting a robustness of the hinge region, possibly mediated through the relative importance of backbone amide interactions versus specific amino acid residue interactions for the majority of residues within this region.

If driver mutations are positively selected, driver mutations should be more likely to occur within the subdomains where cancer-associated mutations are enriched in general, and passenger mutations should occur more frequently in subdomains where cancer-associated mutations occur less frequently in general. To test this hypothesis, a nominal logistic regression analysis, with subdomains taken as the independent variables and predicted driver/passenger status (i.e., predictions about whether a variation is likely to be driver or passenger based on our SVM method) taken as the dependent variable, was done (Table 2, *right*). If our proposed prediction method has randomly selected residues from within the catalytic core as possible cancer drivers, at a rate of 74.50% drivers and 25.50% passengers, then the proportion of mutations predicted as drivers versus passengers should not stray far from this ratio on a subdomain-by-subdomain basis. However, if the variations chosen by our method to be drivers are biased toward residing in particular kinase subdomains, then a higher proportion of mutations within particular subdomains should be predicted as driver mutations. As can be seen in Table 2, this is indeed the case. Subdomains enriched in cancer-associated mutations in general show a higher proportion of predicted driver mutations than the rest of the catalytic domain, whereas subdomains devoid of cancer-associated mutations in general are populated more frequently by passenger mutations. This is depicted visually in Fig. 2, where the driver and CASM densities are depicted in color. Note that both the CASM and driver densities are enriched in subdomains surrounding the nucleotide binding pocket.

Predicted drivers occur at positions enriched in CASMs. The previous analysis suggested that, although the statistical signals from the position-specific distribution of cancer-associated mutations are dampened on a position-by-position basis, it is likely that cancer driver mutations will occur more often at positions harboring a larger number of cancer-associated mutations across all kinases, whereas passenger mutations will occur at positions mutated rarely or in isolation within one (or a random few) kinase only. Therefore, as further validation that our SVM-based prediction technique is identifying true driver mutations, a nonredundant set of the cancer-associated mutations was mapped to specific catalytic core positions based on multiple alignments of the catalytic domain. This nonredundant set ensures that each position is only considered once per individual protein kinase gene. For each cancer-associated mutation, the number of kinases harboring a mutation at its equivalent corresponding position within the multiple alignment was calculated. The frequencies at which predicted driver (mean, 3.2 $\pm$ 0.1 SNPs per position/135 total SNPs) and passenger (mean, 2.4 $\pm$ 0.1 SNPs per position/406 total SNPs) mutations fall at positions mutated in multiple kinases were then compared by the Wilcoxon rank sum test. This analysis confirmed that predicted driver mutations (score mean, 287.0) occur at positions mutated frequently among all kinase genes whereas predicted passenger mutations (score mean, 223.0) occurred at positions rarely mutated in other kinase genes (standardized score, 4.2; $P < 0.0001$). This is depicted visually in Fig. 3, where the numbers of drivers and CASMs per position are depicted in color. Note the close correspondence between the two figures and the preponderance of green CASM sites (2–3 SNPs per position), which become blue driver sites (0–1 SNPs per position).

Driver hotspots. Greenman et al. discuss the abundance of CASMs observed in the glycine loop and the DFG motif, positions which we also observe as mutational hotspots. However, on performing a simulation study to determine what positions are statistically enriched in somatic mutations, only one specific site reached significance. This site, even among the noise of passenger mutations, is mutated in eight different kinases, a frequency that is not expected to occur purely by chance in our simulation study:

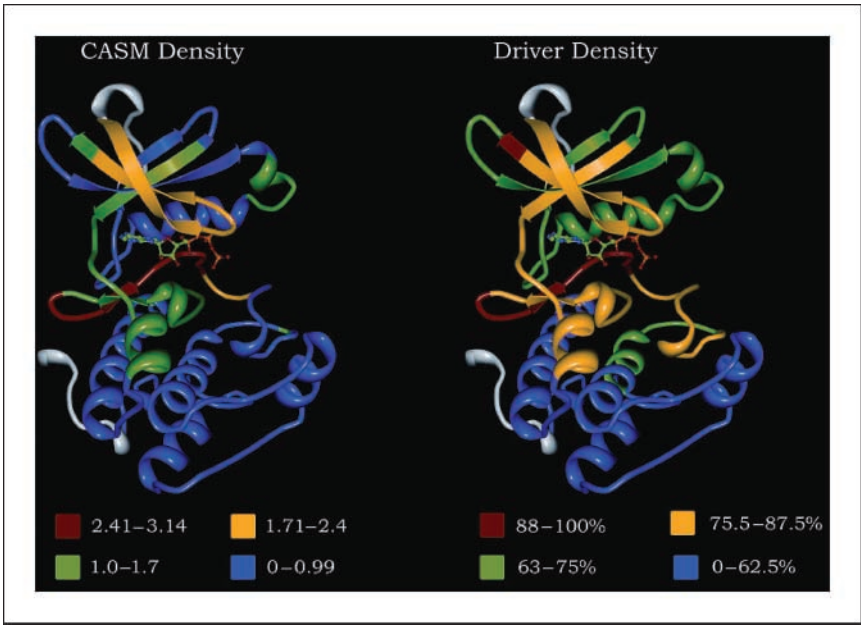

Figure 2. CASM and driver densities mapped to PKA. The subdomains of PKA (PDB ID 1ATP) are colored depending on their CASM or driver density. CASM density is the ratio of expected CASM to observed CASM from Table 2 (*left*). Driver density is the percentage of CASMs per subdomain predicted to be drivers by the SVM method. Note that CASMs and drivers are enriched around the nucleotide binding pocket. Gray regions extend before subdomain I and beyond subdomain XII.

one would expect 8 mutations at 0.4 ± 0.08 residues (95% confidence interval). This position corresponds to the known driver mutations BRAF V600, KIT D816, and PDGFRa D842 (L190 in PKA). On further examination of the literature, this mutation, which also occurs in EGFR L861 (Fig. 1), ABL L387, ErbB2 L869, FLT3 D835, and MET D1246, has been shown to cause kinase activation and, in some cases, resistance to inhibitors in KIT (19), BRAF (20), EGFR (21), ABL (22), FLT3 (23), and MET (24). Thus, mutations at this position seem to be commonly occurring activating mutations in tyrosine kinases; seem to be insensitive to inhibitors; and bear important implications for targeted inhibitor therapies.

Although other sites are not statistically enriched in CASMs, the functional significance of other high ranking positions (i.e., those positions mutated in 6 or more protein kinases) is immediately apparent. Two sites are mutated in six separate kinases. The first is the glycine of the DFG motif. The second corresponds to M120 of PKA. This site too seems to mediate resistance to inhibitors targeting ABL T315 (25), EGFR T790 (26), KIT T670 (27), and PDGFRa (28). We observe additional mutations at this site in NEK11 T108, suggesting that it may be involved in colorectal cancer, and FGFR4 V550. Although FGFR4 carries a valine, rather than threonine, at this position, it should be noted that mutations in RET, which also carries a valine at this position, are implicated in inhibitor resistance (29).

Discussion

Tumorigenesis is an evolutionary process, acting on the accumulation of somatic mutations during tumor progression. The underlying source of this accumulation of mutations, whether it be successive rounds of selection and clonal expansion (30) or the acquisition of a mutator phenotype (31), is controversial. However, the underlying theme is that of an accrual of a large number of mutations, of which only a subset contribute to cancer progression. Identification of these driver mutations among a preponderance of passenger mutations is of utmost importance for the successful exploitation of information obtained by large-scale tumor resequencing studies (32). These predictions will be particularly important in protein kinases, which are major participants in tumor progression and especially important targets for pharmaceutical intervention (2, 3). Thus, the large number of observed somatic mutations in protein kinases (4) and their importance in tumorigenesis substantiate the value of a specialized method capable of highly accurate predictions within the protein kinase gene family.

The accuracy of our prediction method is supported by a battery of tests including (*a*) perfect accuracy based on a small set of known driver mutations; (*b*) excellent agreement with previous statistical estimates of the number of likely drivers on an overall basis, within particular functional domains, and within key functional elements of the catalytic core; and (*c*) frequency analyses at various levels, including individual mutations, the subdomain distribution of mutations, and the occurrence of mutations at positions within motif-based multiple alignments, indicating that predicted driver mutations are under positive selection. This preponderance of evidence strongly suggests that our method is capable of quickly identifying driver mutations in large kinase mutation data sets.

The subdomain distribution of CASMs suggests that enrichment of subdomains with CASMs is indicative of the presence of drivers. Specifically, subdomains I, VII, VIII, and VIIIa are greatly enriched in CASMs and predicted drivers (Table 2; Fig. 2). Subdomain I contains the G-loop, one of the most flexible elements of the catalytic core, which plays a key role in nucleotide binding and phosphoryl transfer. All glycines of this loop are mutated heavily. Mutations in this loop are known to affect kinase activity; for example, substitutions of the third glycine by serine or alanine are known to increase activity in BRAF (33). Subdomain VII participates in phosphoryl transfer, substrate binding, and regulation. Interestingly, the histidine and regulatory arginine of

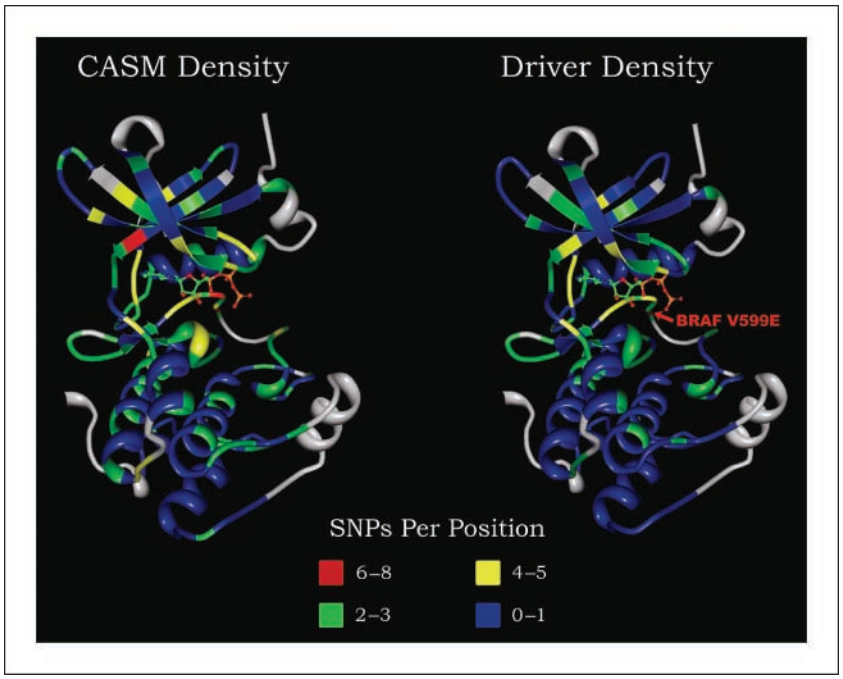

Figure 3. Position-specific distribution of CASM and driver SNPs. The position-specific distribution of CASM and driver SNPs mapped to PKA (PDB ID 1ATP). The positions are colored by the number of SNPs per site (either CASMs or drivers) and correspond to the information provided in Supplementary Table S4. *Arrow,* high-density position V600E (V599E). Note the preponderance of green CASM sites, which become blue driver sites, especially in the COOH-terminal lobe. *Gray regions,* unaligned positions falling within intervening regions (labeled "a" in Table 2).

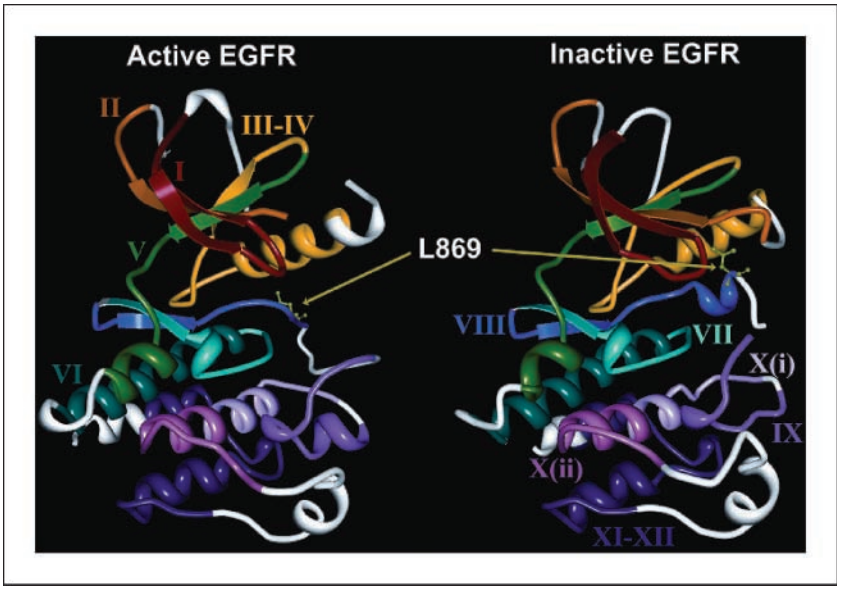

Figure 4. Subdomains and driver hotspot in EGFR. The subdomains of EGFR are colored and labeled by color-matched roman numerals. The structure on the left represents EGFR in the active conformation (PDB ID 2GS6), whereas the structure on the right represents EGFR in the inactive conformation (PDB ID 2GS7). Note that L861 interacts with the N-lobe in the inactive conformation but it does not in the active conformation, suggesting that mutations of L861 disrupt the inactive conformation leading the increased kinase activity.

the HRD motif as well as the tyrosine kinase–specific arginine (E170 in PKA), which is involved in substrate binding (34), are mutated, whereas the HRD aspartate, responsible for the orientation of the P-site hydroxyl acceptor group in the substrate (35), is not. This implies that residues involved in regulation, rather than those more directly involved in catalysis, are targeted. Similarly, in subdomain VIII the DFG-glycine and residues downstream of this glycine in both subdomains VIII and VIIIa, which contribute to the flexibility and rearrangements of this loop (36) and the adoption of the active conformation through phosphorylation of subdomain VIIIa residues, are highly mutated. However, the catalytic aspartate is mutated in proapoptotic proteins LKB1 and DAPK3 (as well as BRAF and HCK), suggesting that this subdomain is involved heavily in both activation and deactivation of protein kinases.

Interestingly, subdomain Ia, at the distal end of the P-loop, is not enriched in CASMs but is enriched with predicted drivers. These drivers occur in tyrosine kinases, RET, MET, EGFR, and EphA6, suggesting that this region may be involved in the dynamics of P-loop motion, specifically in tyrosine kinases. This region is an interesting target for further investigations.

As a result of using motif-based multiple alignments, as opposed to multiple pairwise alignments, a specific position, corresponding to BRAF V600, was observed and predicted to be a driver in BRAF, EGFR, ABL, ErbB2, FLT3, KIT, MET, and PDGFRa. This position is involved in modulating transitions between the active and inactive conformations [e.g., by interaction with the P-loop in BRAF (20) and interaction with the C-helix in EGFR (37)]. Our analysis suggests a generalized role for this position in mediating oncogenesis by disrupting these transitions, especially in tyrosine kinases (Fig. 4).

Another interesting position is the M120 (PKA) "gatekeeper" position of subdomain V, which forms part of the hydrophobic binding pocket for ATP. M120 is important for the shape of the nucleotide binding pocket and is frequently mutated in drug-resistant tumors (38). In fact, although subdomain V is not statistically enriched with CASMs, we do predict an enrichment of drivers in this subdomain, showing the importance of residues involved in nucleotide binding. Another highly mutated residue in this subdomain, G126 (PKA; mutated in five different kinases, all predicted to be drivers), is responsible for interlobe movements (39), providing another example of the importance of protein kinase residues involved in transitions between the active and inactive conformations in cancer progression.

In addition to the positions mentioned above, three positions contain four or more predicted drivers. One of them, L49, provides an additional example of the importance of residues involved in determining the size and shape of the nucleotide binding pocket (40). The other two, K105 and S109, lie in the αC-β4 region; do not seem to be conserved; are not positioned to disrupt the K72-E91 salt bridge, which forms on activation; and their side chains extend away from the nucleotide binding pocket. The functional significance of these residues is unclear and thus would be interesting targets for further investigation.

Overall, our analyses indicate that our method is capable of accurately determining driver mutations in protein kinases. These driver mutations seem to be involved heavily in nucleotide binding, possibly driven by resistance to inhibitors mimicking ATP, and regulatory functions, especially movements from the inactive to active conformation. Although protein kinases are key players in cancer development and progression, accurate predictions of drivers in other protein families, such as transcription factors or phosphatases, will also be useful in determining a more "holistic" picture of tumorigenesis and cancer treatment. Despite this limitation, application of our method to upcoming resequencing studies should be extremely useful in identifying cancer driver mutations among a sea of passenger mutations.

Acknowledgments

Received 9/13/2007; revised 1/9/2008; accepted 1/14/2008.

Grant support: N.J. Schork and his laboratory are supported in part by the following research grants: National Heart Lung and Blood Institute Family Blood Pressure Program grant U01 HL064777-06, National Institute on Aging Longevity Consortium grant U19 AG023122-01, National Institute of Mental Health Consortium on the Genetics of Schizophrenia grant 5 R01 HLMH065571-02, NIH grants R01 HL074730-02 and HL070137-01, and Scripps Genomic Medicine. A. Torkamani is supported in part by the UCSD Genetics Training Grant for the Biomedical Sciences.

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 thank Kannan Natarajan for his work on the motif-based sequence alignments.

References

1. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
2. Baselga J. Targeting tyrosine kinases in cancer: the second wave. Science 2006;312:1175–8.
3. Garber K. The second wave in kinase cancer drugs. Nat Biotechnol 2006;24:127–30.
4. Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153–8.
5. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF. Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics 2006;173: 2187–98.
6. Kaminker JS, Zhang Y, Waugh A, et al. Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res 2007;67:465–73.
7. Kaminker JS, Zhang Y, Watanabe C, Zhang Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 2007; 35:W595–8.
8. Torkamani A, Schork NJ. Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 2007;23:2918–25.
9. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004;91:355–8.
10. McKusick VA. Mendelian inhertance in man. Catalogs of human genes and genetic disorders. 12th edition. Baltimore: John Hopkins University Press; 1998.
11. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–34.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215: 403–10.
13. Torkamani A, Schork NJ. Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family. Genomics 2007;90:49–58.
14. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci 1995;4:1618–32.
15. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 1993;262:208–14.
16. Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G. Structural and functional diversity of the microbial kinome. PLoS Biol 2007;5:e17.
17. Neuwald AF, Liu JS. Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. BMC Bioinformatics 2004;5:157.
18. O'Sullivan O, Zehnder M, Higgins D, Bucher P, Grosdidier A, Notredame C. APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. Bioinformatics 2003;19:i215–21.
19. Furitsu T, Tsujimura T, Tono T, et al. Identification of mutations in the coding sequence of the proto-oncogene c-kit in a human mast cell leukemia cell line causing ligand-independent activation of c-kit product. J Clin Invest 1993;92:1736–44.
20. Wan PT, Garnett MJ, Roe SM, et al. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell 2004;116:855–67.
21. Fu YN, Yeh CL, Cheng HH, et al. EGFR mutants found in non-small cell lung cancer show different levels of sensitivity to suppression of Src: implications in targeting therapy. Oncogene. Epub 2007 Jul 23.
22. Corbin AS, La Rosée P, Stoffregen EP, Druker BJ, Deininger MW. Several Bcr-Abl kinase domain mutants associated with imatinib mesylate resistance remain sensitive to imatinib. Blood 2003;101:4611–4.
23. Yamamoto Y, Kiyoi H, Nakano Y, et al. Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. Blood 2001;97:2434–9.
24. Maritano D, Accornero P, Bonifaci N, Ponzetto C. Two mutations affecting conserved residues in the Met receptor operate via different mechanisms. Oncogene 2000;19:1354–61.
25. Gorre ME, Mohammed M, Ellwood K, et al. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. Science 2001;293: 876–80.
26. Kobayashi S, Boggon TJ, Dayaram T, et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med 2005;352:786–92.
27. Wardelmann E, Merkelbach-Bruse S, Pauls K, et al. Polyclonal evolution of multiple secondary KIT mutations in gastrointestinal stromal tumors under treatment with imatinib mesylate. Clin Cancer Res 2006;12: 1743–9.
28. Cools J, DeAngelo DJ, Gotlib J, et al. A tyrosine kinase created by fusion of the PDGFRA and FIP1L1 genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. N Engl J Med 2003;348:201–14.
29. Carlomagno F, Anaganti S, Guida T, et al. BAY 43–9006 inhibition of oncogenic RET mutants. J Natl Cancer Inst 2006;98:326–34.
30. Tomlinson IP, Novelli MR, Bodmer WF. The mutation rate and cancer. Proc Natl Acad Sci U S A 1996;93: 14800–3.
31. Loeb LA. Mutator phenotype may be required for multistage carcinogenesis. Cancer Res 1991;51:3075–9.
32. Cho WC. A future of cancer prevention and cures: highlights of the Centennial Meeting of the American Association for Cancer Research. Ann Oncol. Epub 2007 Sep 13.
33. Ikenoue T, Hikiba Y, Kanai F, et al. Different effects of point mutations within the B-Raf glycine-rich loop in colorectal tumors on mitogen-activated protein/extracellular signal-regulated kinase kinase/extracellular signal-regulated kinase and nuclear factor κB pathway and cellular transformation. Cancer Res 2004;64:3428–35.
34. Hubbard SR. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. EMBO J 1997;16:5572–81.
35. Adams JA. Activation loop phosphorylation and catalysis in protein kinases: is there functional evidence for the autoinhibitor model? Biochemistry 2003;42:601–7.
36. Kornev AP, Haste NM, Taylor SS, Eyck LF. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. Proc Natl Acad Sci U S A 2006;103:17783–8.
37. Choi SH, Mendrola JM, Lemmon MA. EGF-independent activation of cell-surface EGF receptors harboring mutations found in gefitinib-sensitive lung cancer. Oncogene 2007;26:1567–76.
38. Zhou T, Parillon L, Li F, et al. Crystal structure of the T315I mutant of AbI kinase. Chem Biol Drug Des 2007; 70:171–81.
39. Nolen B, Ngo J, Chakrabarti S, Vu D, Adams JA, Ghosh G. Nucleotide-induced conformational changes in the *Saccharomyces cerevisiae* SR protein kinase, Sky1p, revealed by X-ray crystallography. Biochemistry 2003; 42:9575–85.
40. Bonn S, Herrero S, Breitenlechner CB, et al. Structural analysis of protein kinase A mutants with Rho-kinase inhibitor specificity. J Biol Chem 2006;281:24818–30.

Prediction of Cancer Driver Mutations in Protein Kinases

Ali Torkamani and Nicholas J. Schork

Cancer Res 2008;68:1675-1682.