Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms

Joshua S. Kaminker, Yan Zhang, Allison Waugh, et al.

Cancer Res 2007;67:465-473.

Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms

Joshua S. Kaminker,[1] Yan Zhang,[1] Allison Waugh,[1] Peter M. Haverty,[1] Brock Peters,[2] Dragan Sebisanovic,[2] Jeremy Stinson,[2] William F. Forrest,[3] J. Fernando Bazan,[4] Somasekar Seshagiri,[2] and Zemin Zhang[1]

Departments of [1]Bioinformatics, [2]Molecular Biology, [3]Biostatistics, and [4]Protein Engineering, Genentech, Inc., South San Francisco, California

Abstract

Missense variants are commonly identified in genomic sequence but only a small fraction directly contribute to oncogenesis. The ability to distinguish those missense changes that contribute to cancer progression from those that do not is a difficult problem usually only accomplished through functional *in vivo* analyses. Using two computational algorithms, Sorting Intolerant from Tolerant (SIFT) and the Pfam-based LogR.E-value method, we have identified features that distinguish cancer-associated missense mutations from other classes of missense change. Our data reveal that cancer mutants behave similarly to Mendelian disease mutations, but are clearly distinct from either complex disease mutations or common single-nucleotide polymorphisms. We show that both activating and inactivating oncogenic mutations are predicted to be deleterious, although activating changes are likely to increase protein activity. Using the Gene Ontology and data from the SIFT and LogR.E-value metrics, a classifier was built that predicts cancer-associated missense mutations with a very low false-positive rate. The classifier does remarkably well in a number of different experiments designed to distinguish polymorphisms from true cancer-associated mutations. We also show that recurrently observed mutations are much more likely to be predicted to be cancer-associated than rare mutations, suggesting that our classifier will be useful in distinguishing causal from passenger mutations. In addition, from an expressed sequence tag–based screen, we identified a previously unknown germ line change (P1104A) in tumor tissues that is predicted to disrupt the function of the TYK2 protein. The data presented here show that this novel bioinformatics approach to classifying cancer-associated variants is robust and can be used for large-scale analyses. [Cancer Res 2007;67(2):465–73]

Introduction

A central focus of cancer genetics is the study of mutations that are causally implicated in tumorigenesis. The identification of such causal mutations not only provides insight into cancer biology but also presents anticancer therapeutic targets and diagnostic markers. For example, recent work has provided details of the biology underlying cancer, including information about the types of gene families involved in various stages of cancer (1) as well as the complex nature of the mutational spectra associated with different cancers (2). In clinical settings, these mutations have proved to be extremely valuable in distinguishing patient populations that are responsive to a particular therapy (3–7). In addition to somatic mutations, which are more prevalent in cancers, germ line mutations can confer a predisposition to cancer risks (8, 9). Further study of both somatic and germ line mutations associated with cancer is likely to lead to a deeper understanding of the biology of cancer and possibly will reveal additional targets for therapeutic design.

Targeted sequencing has been done to characterize novel cancer-associated mutations by identifying variants found in tumor tissue (1, 2, 10–16). However, the identification of true cancer-associated variants from such approaches is a challenging problem as these studies often yield large numbers of changes that are not necessarily causally associated with cancer (2, 16). This is partly a result of non–cancer-causing somatic variants, termed passenger mutations (2, 16), which accumulate in cancer tissue due to the high mutation rate and multiple cell divisions seen during tumor growth (17). Different types of tumors may display different point mutation rates. Cells with microsatellite instability, for example, are known to exhibit much higher mutation rates than those with chromosomal instability (reviewed in ref. 18), although tumors with DNA copy number alteration may also have a large number of point mutations (19). In large-scale screening for mutations, passenger mutations were estimated to account for two thirds of all variants identified (16). The identification of cancer-associated mutations can also be hindered by tumor heterogeneity as oncogenic changes may not be present equally throughout a tumor (20). As a consequence, such variants may be present only in a small fraction of the tissue sample used in a targeted sequencing experiment and are not easily detected (20). Thus, infrequently occurring cancer-associated changes can be difficult to distinguish in a directed sequencing approach. Aside from functional analysis, there is no method available to differentiate those variants that are responsible for tumor progression from other changes.

In addition to targeted sequencing approaches, expressed sequence tag (EST) sequences have also been used to identify missense changes overrepresented in cancer libraries (21, 22). However, similar to the targeted sequencing work, these analyses yielded large numbers of variants of which only a small number would likely contribute to cancer progression. This is not only due to factors such as passenger mutations but is also a result of poor data quality from high-throughput sequencing of EST libraries. Perhaps for this reason, none of the novel variants identified from such studies have been shown to exist in independent tumor samples. Although some of the putative cancer variants identified

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

Requests for reprints: Zemin Zhang, Bioinformatics, Genentech, Inc., South San Francisco, CA 94404. Phone: 650-225-4293; E-mail: zemin@gene.com.

doi:10.1158/0008-5472.CAN-06-1736

from EST data are expected to be real, such screens are hindered by a lack of additional filtering mechanisms that can enhance true signals from the high level of background noise.

Aside from missense cancer mutations and common polymorphisms, other classes of missense change have been identified. Missense mutations known to cause Mendelian disease have been annotated in Swiss-Prot and studied for their deleterious effects on protein function (23). In a recent work, missense mutations associated with complex disease were collected and shown to have characteristics distinct from Mendelian disease variants (24). The ability to distinguish missense changes by their disease phenotype suggests that comparison of cancer-associated variants to other classes of variant could be a valuable tool in understanding cancer progression. Furthermore, the availability of data sets of missense mutations associated with complex disease (24), Mendelian disease (23), and cancer (25) make such an analysis possible. Several computational methods, including Sorting Intolerant from Tolerant (SIFT; ref. 26), Polymorphism Phenotyping (PolyPhen; ref. 27), the Pfam-based LogR.E-value (28), large-scale annotation of coding nonsynonymous SNPs (LS-SNP; ref. 29), statistical geometry methods (30), support vector machine methods (31), decision trees (32), and random forest (RF) classifiers (33), have been developed to identify deleterious variants. However, it remains to be determined if these algorithms are able to distinguish cancer-associated variants from other types of change.

Here, we have used a number of computational methods to define characteristics of known cancer mutations and subsequently developed a novel approach for predicting cancer-associated mutations from among a large set of missense variants. This method provides a means of analyzing large-scale data sets and will likely prove to be increasingly relevant to genome-scale efforts currently under way to identify mutations involved in cancer progression.

Materials and Methods

Data sets. Variants were assembled as follows: (*a*) common variants were downloaded from National Center for Biotechnology Information (NCBI)[5] and overall minor allele frequencies were determined from the file SNPAlleleFreq.bcp from the NCBI ftp site;[6] (*b*) cancer-associated variants were collected from the COSMIC ftp site[7] and, based on the analysis of Forbes et al. (25), include only variants in those genes most likely to be involved in oncogenesis (25); (*c*) Mendelian disease–associated variants were obtained from identifying those records in the file uniprot_sprot.dat from the Swiss-Prot ftp site[8] that contains nucleotide change data, were human records, were of the type "disease" but not of the subtype "cancer," and did not overlap with records in single nucleotide polymorphism database (dbSNP); (*d*) complex disease–associated variants were collected from previous work (24). All data sets will be available online.[9]

SIFT and Pfam-based LogR.E-value scores. The SIFT program was downloaded[10] and installed and run locally. Scores were obtained from SIFT output, and only those variants with a median sequence information of <3.25 were included in the analyses. Pfam-based LogR.E-value scores were derived from scores provided by the HMMER 2.3.2 software. The ls mode was used to search against the Pfam protein family database. LogR.E-value scores were calculated as $\log_{10}(E_{\text{variant}}/E_{\text{canonical}})$ (28). Note that for the discussion and display of data (Figs. 1 and 2; Table 1), we have used a negative version of this value (see ref. 28).

Gene Ontology log-odds analysis and RF classification. The log-odds scores were calculated to represent the relative frequency with which a Gene Ontology (GO) term was used to annotate cancer or noncancer gene sets. All genes represented in the COSMIC database associated with an oncogenic phenotype were used as the cancer gene data set. Further details of the GO analysis are presented as Supplementary Data.

The RF classifier was built using the package randomForest 4.5-16[11] for the R statistical environment.[12] The classifier was trained on 200 cancer mutations and 800 noncancer mutations by using the SIFT score, LogR.E-value score, and GO log-odds score for each variant. The mutation to SNP ratio was empirically determined based on the numbers of somatic mutations and background polymorphisms identified in the tyrosine kinome (12), tyrosine phosphatome (34), and serine-threonine kinase data (10). Technical details of this method as well as details of the pathway analysis are presented as Supplementary Data. All training data are freely available online.[9]

EST-based identification of variants. A set of 22,332 RefSeq mRNAs was downloaded from NCBI, and each mRNA was aligned to ESTs from public and Incyte collections. ESTs were aligned to the genome using GMAP (35) to determine overlaps with RefSeq mRNA genomic coordinates. Those ESTs that overlapped with a particular gene region were aligned to the corresponding RefSeq mRNA using NCBI BLAST run with default parameters. All mismatches were required to be flanked by 30 bp of sequence identical to the RefSeq mRNAs. Rare events that occurred in only one EST or were seen in only one library were eliminated, and variants were required to be observed in at least 3% of the total number of cancer ESTs. Any variant that overlapped with a record in dbSNP was eliminated to avoid identifying known or common SNPs. For similar reasons, variants identified in normal libraries were also eliminated. Lastly, a z-score was calculated to test the significance of the difference in the number of ESTs contributed from cancer or normal libraries. Variants were eliminated for $P > 0.1$. This list of putative computationally defined cancer variants was then further enriched by selecting those changes that were identified as cancer-associated by the RF classifier. Variants identified from the screen are presented online.[9]

Validation of missense mutations. For traditional sequencing, PCR products were amplified from genomic DNA extracted from cell lines. Sequence reactions were done using conventional Sanger sequencing methods for both sense and antisense directions on an Applied Biosystems 3730xl DNA Analyzer (Applied Biosystems, Foster City, CA). Traces were analyzed using Sequencher (Gene Codes, Ann Arbor, MI). For mass spectrometry analysis, variation validation assays were designed using MassARRAY 3.0.2.0 software (Sequenom, San Diego, CA). See Supplementary Data for further details. The cancer sequence data were obtained from the following tumor tissue types (and numbers) of samples: lung (64), breast (37), stomach (10), colon (9), and liver (8). The oligonucleotide primer used for TYK2 P1104A variant was 5′-TGACTCCAGCCAGAGC-3′.

Results

Comparison of variant data sets. We collected four groups of missense variants as an initial step in understanding differences among variant groups. A set of 5,747 common polymorphisms used as a baseline data set was collected from dbSNP by extracting those polymorphisms with an overall minor allele frequency (MAF) >20%. Mendelian disease mutations were identified by isolating variants from the Swiss-Prot database annotated as disease but not cancer. These 11,456 noncancer disease missense mutations represent

[5] http://www.ncbi.nlm.nih.gov/projects/SNP/.
[6] ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data.
[7] ftp://ftp.sanger.ac.uk/pub/CGP/cosmic.
[8] ftp.uniprot.org/pub/databases/uniprot/knowledgebase/uniprot_sprot.dat.gz.
[9] http://share.gene.com/mutation_classification.
[10] http://blocks.fhcrc.org/sift/SIFT.html.
[11] http://stat-www.berkeley.edu/users/breiman/RandomForests.
[12] http://www.r-project.org.

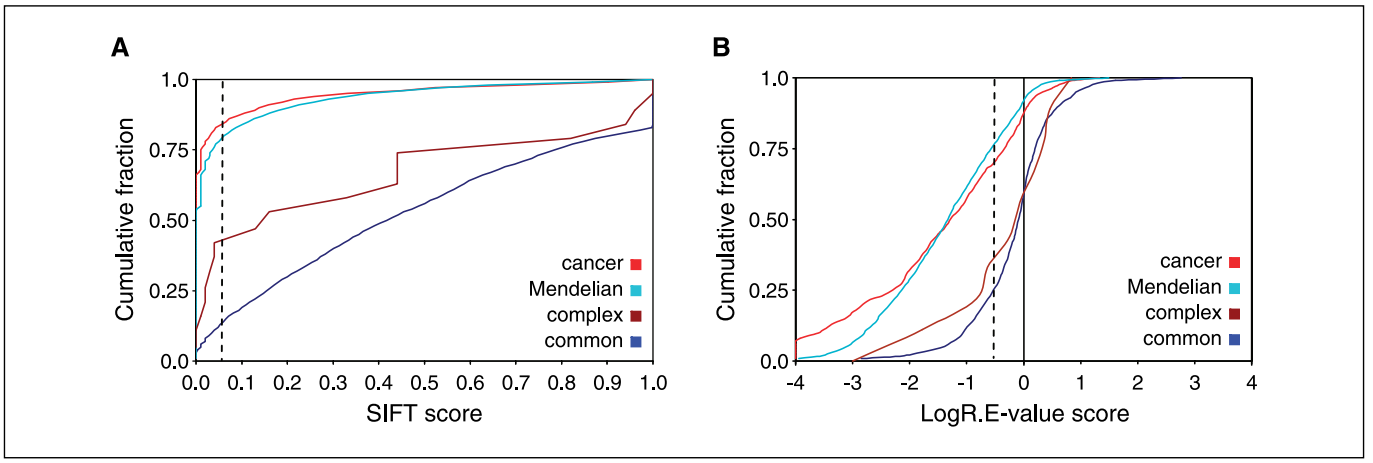

Figure 1. Cumulative distributions of SIFT scores (*A*) and LogR.E-value scores (*B*) for different variant classes. Intolerant changes are predicted by SIFT scores less than 0.05 or LogR.E-value scores less than −0.5 (*dashed lines*). The cancer and Mendelian variants show similar distributions of scores by either SIFT (*A*) or LogR.E-value (*B*) metrics. The uneven distribution of the complex data is likely due to the small number of variants included in this data set. Data have been connected with a smoothed line in Excel. The two-tailed Wilcoxon rank-sum test was used to test the null hypothesis that data from the individual distributions were drawn from the same populations. Significant *P* values for pairwise comparisons for SIFT scores are as follows: cancer-common, 2.20e−16; complex-cancer, 3.41e−9; Mendelian-cancer, 1.68e−7; Mendelian-common, 2.20e−16; Mendelian-complex, 4.57e−7. Significant *P* values for pairwise comparisons for Pfam-based LogR.E-value scores are as follows: cancer-common, 2.20e−16; complex-cancer, 2.55e−3; Mendelian-common, 2.20e−16; Mendelian-complex, 7.72e−4.

changes to single genes that follow standard Mendelian rules of inheritance and are thought to contribute to disease progression (23). From a previous study (24), we collected a set of 27 complex disease–associated variants that are known to contribute to disease progression but do not follow standard Mendelian rules of inheritance. Last, a set of 1,091 cancer-associated somatic mutations were gathered from the COSMIC database (25) in which cancer-associated mutations manually curated from literature sources are compiled. Because some mutations in COSMIC may be passenger or bystander mutations, we attempted to enrich for true cancer-associated mutations by using variants identified in genes likely to be involved in cancer progression (25).

Two distinct approaches were used to measure the effect of each variant on protein function. First, the SIFT program (26) was used to predict if a variant is likely to affect protein function. SIFT uses sequence homology between closely related protein species to measure this effect, and low SIFT scores (<0.05) are predictive of intolerant changes. The SIFT scores from all variants were calculated and plotted by class as cumulative distributions (Fig. 1*A*). The distributions of SIFT scores suggest that these variants can be divided into at least two classes: Mendelian/cancer–associated and complex/common. The differences between these groups are most apparent for the intolerant SIFT scores between 0.0 and 0.05 where 83% and 77% of the cancer and Mendelian variants are predicted to be intolerant compared with 42% and 12% of the complex and common missense changes. Not surprisingly, the distinction between cancer and normal variants becomes smaller for cumulative distributions of SIFT scores derived from SNPs with an overall MAF of <20% (not shown). Although cancer is often thought of as a complex disease involving changes to multiple loci, in this assay, cancer mutations seem more similar to Mendelian disease variants than complex disease variants.

The second approach used to predict the effect of a variant on protein function is the Pfam-based LogR.E-value, which calculates

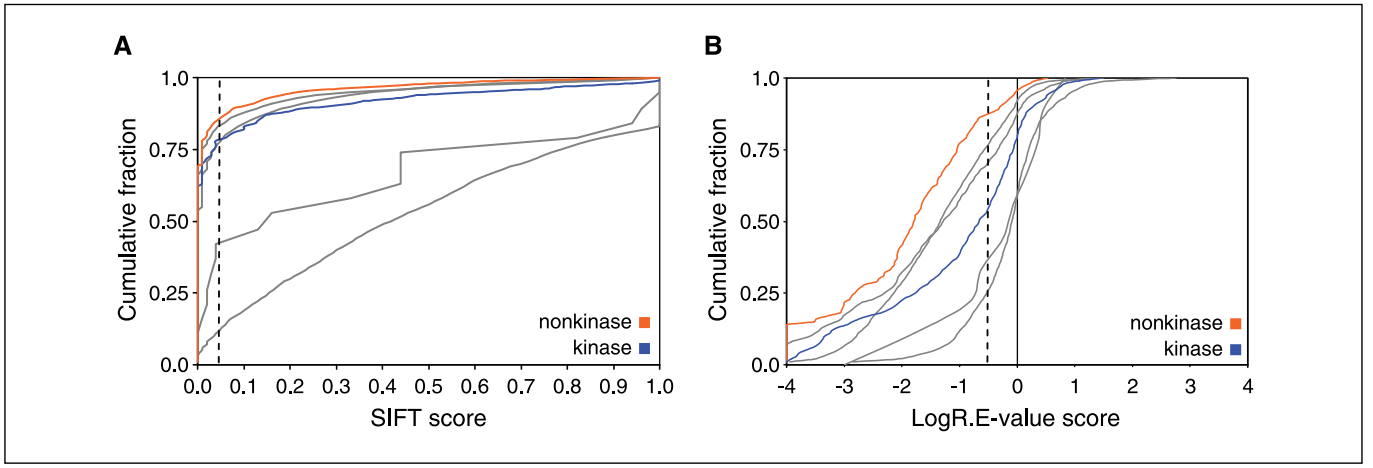

Figure 2. Cumulative distributions of SIFT (*A*) and LogR.E-value (*B*) scores for variants in kinase and nonkinase-containing genes. Intolerant changes are predicted by SIFT scores <0.05 or LogR.E-value scores less than −0.5 (*dashed lines*). *Gray lines,* data presented in Fig. 1. Data have been connected with a smoothed line in Excel.

Table 1. SIFT and LogR.E-value scores for activating and inactivating mutations in select genes

Mutation type	Gene	Mutation	SIFT score	Pfam-based LogR.E-value
Activating	*BRAF*	V600E	0	−0.3
	KRAS	G12V	0.01	−1.97
	KIT	D816V	0	0.5
	PDGFRA	V561D	0	NA
	EGFR	L858R	0	−2.03
	JAK2	V617F	0	−0.27
Inactivating	*p53*	R175H	0	−1.51
	PTEN	C124S	0	$-\infty$
	P16	D84G	0	NA

Abbreviation: NA, not applicable.

the difference between a wild-type and variant protein by measuring their fit to a Pfam model (28). The LogR.E-value score is derived from the *E* value provided by HMMER2 software (36) and was used previously to predict whether specific changes to a protein were likely to affect protein activity or stability (28). The underlying scoring systems behind LogR.E-value and SIFT are different (28), and it has been shown that these algorithms produce distinct metrics that can be used to independently analyze variant data (28).

Based on previous work with the LogR.E-value, scores that are less than −0.5 are predicted to alter protein function (ref. 28; see Materials and Methods). The values generated from analysis of all four variant sets using this method were plotted as cumulative distributions (Fig. 1*B*). As seen with the analysis of SIFT data, the distribution of the cancer and Mendelian disease–associated variants are fairly similar to one another, but both are significantly different from the distributions of either the complex or common variant data sets (Fig. 1*B*). Moreover, similar to the SIFT analysis, the cancer and Mendelian data sets have a larger percentage of changes predicted to have a significant effect on protein function (70% and 77%, respectively) compared with the complex and common variant data sets (33% and 26%, respectively). These data once again point to differences between cancer-associated variants and complex disease variants. Further, and perhaps even more apparent than the SIFT analysis, these data reveal similarities between Mendelian and cancer-associated variants that underscore that these changes may affect protein function in a similar manner.

Characterization of variants in different classes of genes. As the SIFT and Pfam-based LogR.E-value metrics proved useful in studying differences between populations of variants, we used these tools to identify characteristics of variants defined by biochemical analysis as activating or inactivating. Well-characterized inactivating mutations in the genes *p53* (37), *PTEN* (38), and *p16* (39) were analyzed. Each of these mutations was predicted to impair protein function in at least one of the assays (Table 1). This result is consistent with the debilitating characteristics of these mutations. Next, activating mutations in BRAF (40), K-Ras (41), Kit (42), PDGFR α (43), JAK2 (44), and epidermal growth factor receptor (EGFR; refs. 3, 5, 6) were analyzed in the same manner. Although these activating mutations increase the output of different signaling pathways and might not be considered intolerant, it was striking that all mutations were predicted to impair protein function in at least one of the assays (Table 1). Together, these results suggest that both activating and inactivating oncogenic mutations seem to be untolerated by wild-type proteins although the activating mutations are able to increase the signaling ability of these molecules.

As many activating changes have been identified in kinase genes, we were interested in examining a larger pool of such mutations. Using data from the COSMIC database, mutations were collected either from genes with kinase domains or genes without kinase domains. The distribution of SIFT scores for these two variant data sets suggests that these changes are largely predicted to be intolerant (78% and 85%, for the variants in kinase and nonkinase genes, respectively; Fig. 2*A*). The Pfam-based LogR.E-value scores follow a similar trend to the SIFT data in that 54% of the kinase variants and 87% of the nonkinase variants are predicted to affect protein function (Fig. 2*B*). A recent analysis of cancer-associated variants in genes with kinase domains (45) revealed that a majority of such changes are likely to be activating as they often affect residues that regulate the catalytic activity of the kinase (45). Thus, the results presented in Fig. 2 not only may reflect the similarities between kinase and nonkinase genes but also may suggest that, in general, activating and inactivating changes are not tolerated.

GO analysis of cancer genes and the development of a cancer-variant predictor. We next attempted to identify an additional metric that could be used to distinguish the cancer-associated variants by classifying the genes in which they reside. Using a standard set of GO annotations, a log-odds score was calculated to measure the difference in frequency that a particular GO term was used to annotate genes that were either known to be involved in cancer (25), or, not known to be involved in cancer and represented by a RefSeq mRNA. For example, genes annotated with the term "ion transport" are underrepresented in the cancer data set, and this term has a negative log-odds score of −2.30. On the other hand, genes annotated with the term "cell cycle" are overrepresented in the cancer data set, and this term has a positive log-odds score of 2.06 (Fig. 3*A*). For a gene of interest annotated with a set of GO terms, one can sum the log-odds score for each GO term. This cumulative score reveals inherent differences between genes containing common variants and genes containing cancer variants as shown in Fig. 3*B*. The different distributions are not a result of the varying levels to which these gene sets were annotated, as scores that were normalized by the number of annotated terms produced very similar data (not shown). Although these data could reflect an inherent bias in the types of genes chosen for analysis, it is likely that this approach reveals distinguishing features of the cancer and noncancer gene sets.

Based on the above analysis, it seemed possible that a computational method could be developed that would combine the information from the SIFT, Pfam-based LogR.E-value, and GO log-odds metrics to predict whether a variant was likely to be cancer associated. The RF classifier provides such a method by dividing a large pool of data into smaller subsets based on characteristics of each datum. This method has been used successfully in many biological applications, including distinguishing harmful SNPs from harmless SNPs (33) and classifying microarray data (46). In our analysis, variants included in the training set used to construct the classifier were assigned a measurement from each of the three different algorithms described above (SIFT, LogR.E-value, and the GO log-odds score). During the construction of the classifier using these sets of training data, an

internal measurement of the rate at which training data are misclassified was determined to be 3.19% [this value is termed the out-of-bag (OOB) error rate; for further details, see Materials and Methods].

Further validation of the classifier was achieved by performing a cross-validation experiment in which a group of known variants was entirely excluded from the training process during the construction of the classifier. These excluded variants were then analyzed using the RF predictor to determine how often they were misclassified. From this analysis of 730 variants, only 10 of 581 (1.7%) normal variants were misclassified as cancer and only 13 of 149 (8.7%) cancer variants were misclassified as normal. Together with the OOB error rate, these data illustrate that this is a strong classifier with a very low rate of false-positive predictions. It should be noted that the noncancer variant training data set consists only of a subset of variants from dbSNP with a MAF of >20%. We expect to observe a somewhat higher rate of misclassification if more rare SNPs are included.

Validation and practical application of the classifier. A useful application of our classifier is in distinguishing relevant cancer-associated mutations from the expected polymorphic variants often identified during sequencing projects to discover novel somatic mutations. We attempted to model this process by collecting data from several protein family–based studies, including variants in the tyrosine kinome (12), tyrosine phosphatome (34), and serine/threonine kinase gene family (10). For each group of genes, two sets of variants were identified: the set of somatic cancer–associated mutations reported in each publication and a set of all missense SNPs isolated from dbSNP, which represent those changes that would be expected as background polymorphisms identified in sequence data. The variants were then classified by our predictor. It should be noted that none of these variants were used to train the RF classifier. As shown in Fig. 4*A*, although 64% of the somatic mutations in the tyrosine kinome were predicted to be cancer-associated, a significantly smaller percentage of the changes from dbSNP (23%) were predicted to be cancer-associated (P = 3.9e−4; two-tailed Fisher's exact test). Analysis of the tyrosine phosphatase data set reveals that the somatic mutations were also more often predicted to be cancer associated than the variants isolated from dbSNP (P = 0.034; Fig. 4*B*). A similar trend is observed for the serine/threonine data set in which 36% of the somatic variants and only 19% of the variants from dbSNP were predicted to be cancer associated (Fig. 4*C*). Thus, our predictor is capable of distinguishing true

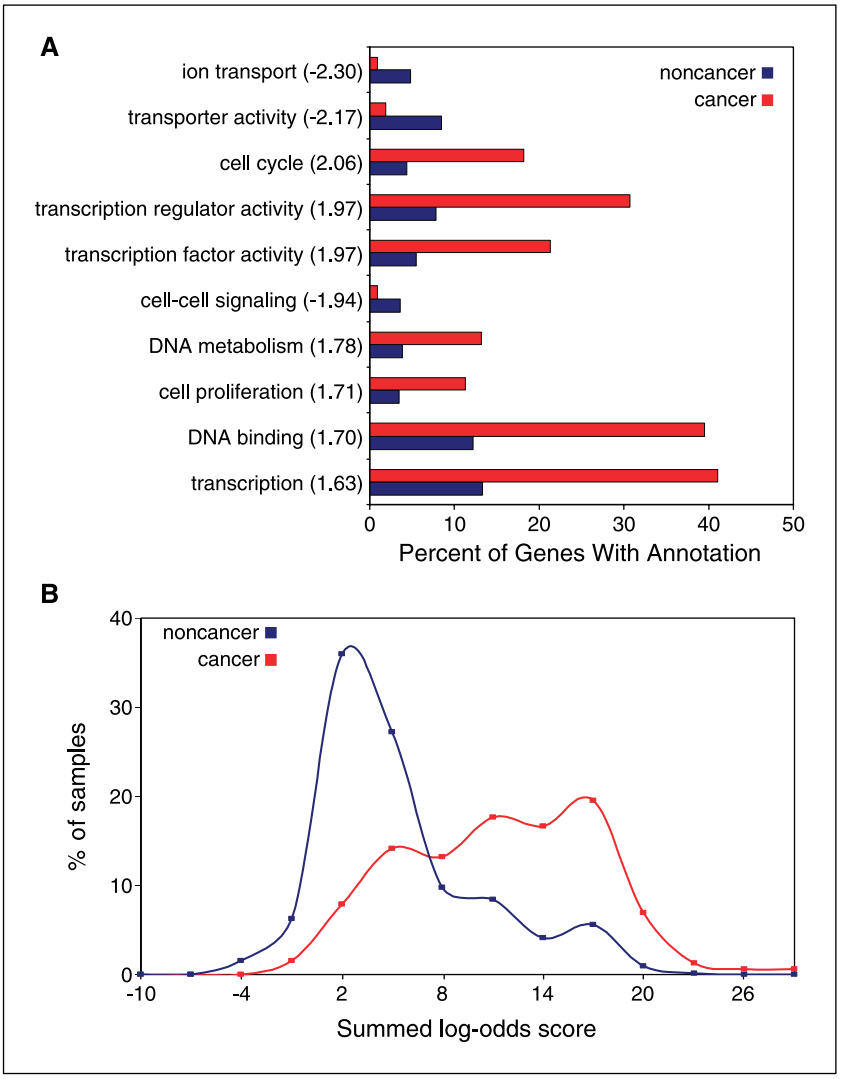

Figure 3. Log-odds scores are distinct across cancer and noncancer data sets. *A,* the frequency that a particular term is used to annotate genes in noncancer (*blue*) or cancer (*red*) data sets. *Numbers in parentheses,* log-odds scores. Although only the 10 terms with the largest log-odds scores are shown, the entire term list from the GO slim generic data set was used in assigning summed scores to particular genes (see Materials and Methods). *B,* summed log-odds scores were calculated for noncancer genes (*blue*) or genes containing cancer variants (*red*) and plotted as a frequency histogram. Binned values (*boxes*) have been connected with a smoothed line in Excel.

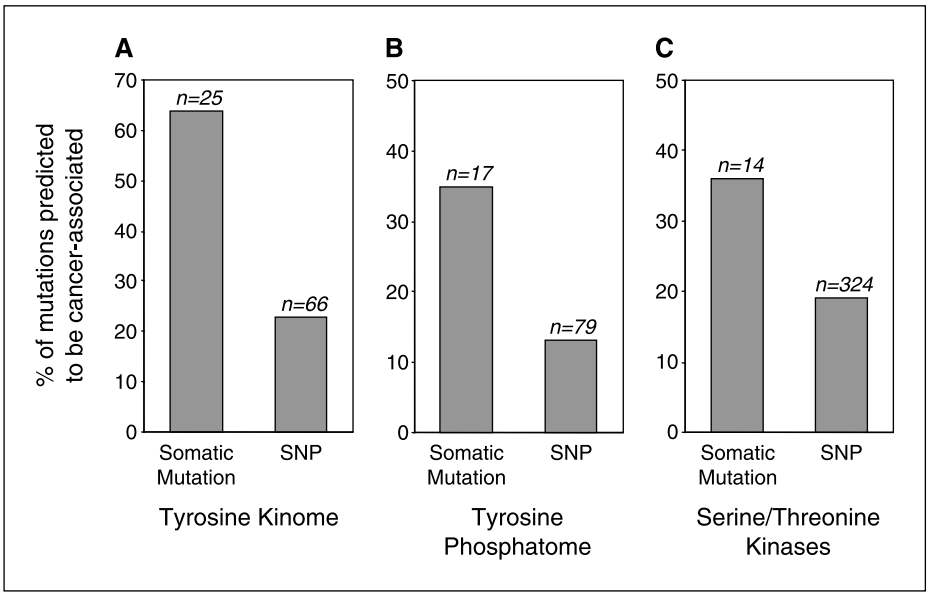

Figure 4. Data sets of somatic mutations are more likely to be predicted to be cancer-associated than variants in dbSNP. Variants in genes of the tyrosine kinome (*A*), tyrosine phosphatome (*B*), or serine/threonine kinases (*C*) were classified using the RF predictor described in the text. Classes of somatic variants and variants isolated from dbSNP are indicated below each panel. Only variants for which there was a GO log-odds score, a LogR.E-value score, and a SIFT score were used in this analysis. The total number of variants is labeled above each column. The percentage of variants predicted to be cancer-associated for somatic and SNP variants is as follows: *A,* 64% and 23%; *B,* 35% and 13%; *C,* 36% and 19%.

cancer mutations from among large sets of sequence data even in the absence of extensive knowledge of SNPs. It is worth noting that the percentage of cancer-associated variants from these data sets was lower than that observed during the analysis of the training data. This likely reflects the fact that the experimental data described above presumably contain fewer true cancer causal variants than the training data mutations, which were derived from a select group of well-studied mutants.

Whereas true causal mutations are under positive selection and are therefore observed recurrently, bystander or passenger mutations occur in a stochastic fashion and are expected at lower frequencies. We were interested in determining if those variants that occur infrequently are less likely to be predicted to be cancer associated by our classifier. To address this issue, the entire set of cancer-associated variants in COSMIC was divided into groups according to the number of times that a change was observed. The percentage of variants predicted to be cancer associated was determined for each class using the RF classifier. As shown in Fig. 5*A*, whereas 58% of variants observed at least 10 times are predicted to be cancer associated, only 43% of variants occurring only a single time are predicted to be cancer associated ($P = 0.018$, two-tailed Fisher's exact test). This analysis not only shows the usefulness of our classifier but also suggests that a large fraction of infrequent variants in COSMIC could be passenger mutations.

As an additional test of the usefulness of our classifier, we attempted to predict cancer-associated mutations from a recent large-scale sequencing effort of human colorectal and breast tumors (47). In this recent analysis, genes were classified into two groups based on mutation frequencies: those likely to be involved in tumor progression (CAN genes) and those less likely to be involved in tumor progression (non-CAN genes; ref. 47). Variants identified in either CAN genes or non-CAN genes were classified by our predictor. As shown in Fig. 5*B*, variants identified in CAN genes are much more likely to be classified as cancer associated (26.3%) than variants in non-CAN genes (13.3%; $P = 8.8e{-}6$; two-tailed Fisher's exact test). The larger number of cancer-associated variants predicted from the CAN genes is consistent with the suggestion that these mutations were under selection during tumor progression (47). Importantly, independent of the method used for deriving the CAN genes, our classifier provides a novel method that prioritizes mutations for *in vivo* functional analysis.

EST-based identification of novel variants. We designed an EST-based screen to collect novel variants that could then be classified using our predictor. In this approach, ESTs from libraries derived from either normal or cancer tissue were aligned to RefSeq mRNAs to identify mismatches. From 2,600 candidate variants that were specifically present in ESTs from cancer libraries, 494 variants (19%) were identified as cancer associated based on the RF predictor. This list includes a number of known cancer-related mutations, such as the C135F change in p53.

Such an EST-based approach is likely to be complicated by the sequence noise commonly associated with EST data. To address this issue, eight novel variants were selected for validation by determining if the changes were in fact present in genomic sequence or were identified as a consequence of noisy EST sequences. Two of the variants, a V1304M change in MAST2 and a Y11H change RIOK2, were identified in genomic DNA from their respective EST tissue sources (Supplementary Fig. S1), which show that some of these events indeed occur in genomic sequences and are not a result of EST sequencing artifacts. Although it was not surprising to identify these changes in their respective EST tissue sources, it was important to determine whether some of the predicted mutations were present in unrelated tumor DNA samples. This was addressed by performing mass spectrometry genotyping for 65 predicted cancer variants over a collection of 128 tumor tissue samples that were independent of any EST libraries used in the screen. From this analysis, we identified one novel variant, the P1104A variants in the kinase domain of the *TYK2* gene, which was present in four independent tumor tissues (Supplementary Fig. S2*A* and S2*B*). This variant was classified as a germ line change as it was also identified in matched normal samples.

It is likely that the P1104A variant is a novel cancer-associated germ line mutation that affects TYK2 function. This change has not been found in other normal samples and is absent from the comprehensive dbSNP. Analysis of sequence data covering the entire *TYK2* gene of >47 normal individuals by the Seattle SNP project (48) and 102 normal individuals catalogued at the SNP500Cancer project failed to reveal any samples with this change. The P1104A variant

mutates a conserved proline that underlies the substrate-binding groove in the COOH-terminal, helical lobe of the TYK2 kinase domain. In addition, the proline is positioned under a key tryptophan residue in a ring-stacking interaction that stabilizes the inactive conformation of the activation loop. The mutation of proline to an alanine could help precipitate an activated catalytic state of TYK2, which may lead to an oncogenic phenotype.

Discussion

Missense sequence variants can exhibit a wide range of effects on proteins, producing diverse phenotypic outcomes. From the analysis in this article, we have shown that it is possible to use a combination of computational metrics to predict those changes that are more likely to contribute to oncogenesis. This approach becomes increasingly relevant in light of the recent initiation of large-scale cancer genome initiatives (11), which are likely to produce large numbers of variants of which only a small proportion will be implicated in tumorigenesis.

The contribution of multiple somatic changes makes it difficult to classify cancer as either a Mendelian or a complex disease. Oncogenesis often requires the accumulation of mutations in multiple genes, and one might be tempted to consider cancer mutations as more similar to those involved in complex diseases. Thus, it is somewhat surprising to observe that cancer mutations behave more similarly to Mendelian disease mutations than to complex disease mutations. Although one could argue that these results simply reflect the fact that variants resembling Mendelian mutations are those most easily studied in the laboratory and subsequently collected in databases such as COSMIC, it is more likely that these results point to cancer as a disease that progresses through a series of stepwise deleterious mutations (49). Such mutations are likely to lie in the class of genes that, when altered, mediate tumor progression by providing cancer cells growth advantages, metastatic capabilities, and the ability to escape from homeostasis control (49). Our analysis does not imply that variants similar to those found in complex disease are less important for tumor progression. However, those changes may be less frequent and more diverse than the changes we identified.

An interesting observation from our analysis is that activating mutations seem to impair protein function. Although one might hypothesize that such changes would be desirable because they increase the output of the molecule, our data suggest that proteins that function at a wild-type level are selectively maintained. Our analysis of variants in kinase and nonkinase genes supports the notion that activating mutations in kinases primarily affect residues involved in the control of their enzymatic activity (12, 15, 50). An apparent gain-of-function phenotype could be manifested as a damaging change to a regulatory domain that controls kinase activity. In fact, this theory has been supported by known activating mutations in BRAF and EGFR (3, 5, 6, 40).

With the aid of our computational classifier, we were able to identify a novel germ line variant from EST sequences, the P1104A change in the kinase domain of TYK2. This variant is present in four different tumor samples independent of the EST libraries where it was originally found. As this variant was not found in other normal samples, it is likely that P1104A is a germ line mutation associated with increased cancer risk. Although we are encouraged by this finding, we also observed many limitations intrinsic to EST-based mutation screening. First, a majority of the observed variants from ESTs are likely sequence artifacts. In our hands, only two of eight expected variants were found in the genomic DNA from the same tissue sources where the variant ESTs were derived. Second, the ability to detect mutations is heavily influenced by EST coverage and library bias. Although we were able to observe some of the known cancer mutations, many of them were missed due to insufficient EST coverage at expected locations. For example, the V600E mutation of BRAF, a prevalent somatic mutation in melanoma, was missed owing to a lack of EST sequences from melanoma libraries covering the expected location. Furthermore, true mutations identified in the EST libraries may not be easily found in unrelated tumor samples if they occur at a low frequency. Despite these difficulties, we show here that it is still possible to identify novel cancer-associated variants from the EST

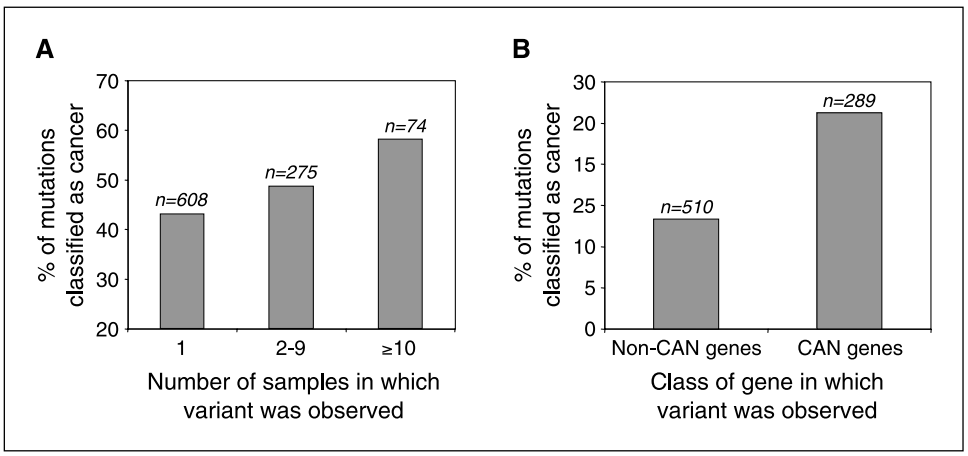

Figure 5. Cancer mutation prediction in different subsets of somatic variants. *A,* frequently occurring mutations are more often predicted to be cancer-associated than rare mutations. Somatic mutations isolated from the COSMIC database were divided into classes based on the number of times a particular change was seen in independent tissue samples (*1, 2–9*, or *≥10*). Only variants for which there was a GO log-odds score, a LogR.E-value score, and a SIFT score were used in this analysis. The number of variants in each respective class is indicated above each column. The difference between the variants with one mutation and those with ≥10 mutations is statistically significant (P = 1.8e−2; two-tailed Fisher's exact test). *B,* variants in genes described as CAN are classified as cancer-associated more frequently than variants in non-CAN genes. Only variants for which there was a GO log-odds score, a LogR.E-value score, and a SIFT score were used in this analysis. The number of variants in each class is indicated above each column. The difference between the number of variants predicted to be cancer-associated between the CAN and non-CAN genes is significant (P = 8.8e−6; two-tailed Fisher's exact test).

data. Given the reduced cost of genotyping, it is becoming feasible to validate large numbers of putative cancer variants.

The approach described here will likely facilitate the identification of new mutations associated with oncogenesis. With efforts such as The Cancer Genome Atlas project under way, a large volume of sequence variation data is expected to accumulate in the next few years. Such data will be most useful if it is possible to distinguish the rare, meaningful mutations from among the large number of missense variants. As shown in our analysis, using a variety of different data sets, the method we developed provides statistically significant enrichment of cancer-associated changes. Consistent with these data, a preliminary analysis of an unpublished large sequence data set using this classifier revealed that 47% of variants identified as somatic or germ line were predicted to be cancer associated, whereas only 19% of polymorphisms known in dbSNP were predicted to be cancer associated.[13]

Although our classifier did well on different sets of variants, the frequencies of predicted cancer and noncancer mutations vary from those suggested by the OOB error and cross-validation data. It is likely that this is in part due to the requirement that common SNPs used to train the classifier have a MAF of >20%. This high MAF was chosen to ensure that the majority of SNPs used to train the classifier were indeed common and unlikely to be deleterious. In the experiments described above, no attempt was made to exclude SNPs based on their MAF and it would not be surprising to find that some of the variants described as common are rare and possibly deleterious. In addition, while the training data were derived from a highly studied set of variants, much of the large-scale experimental data consist of novel changes that are not yet well studied for their role in cancer progression. Although some of the changes identified by Sjoblom et al. (47) are likely to be tumorigenic, it is possible that others will be classified as passenger or background changes.

It may be desirable in the future to attempt to separate cancer mutations from other types of variants using additional structural characteristics of amino acid changes, such as those presented in the recent work by Furney et al. (51). However, the approach here is an important first step toward differentiating cancer-associated variants from other types of change. The novelty and usefulness of this algorithm provides a needed method that will enable the prioritization of large data sets of variants for further study.

[13] J. Kaminker and S. Seshagiri, unpublished observations.

Acknowledgments

Received 5/12/2006; revised 11/1/2006; accepted 11/8/2006.

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 Shiuh-Ming Luoh, Colin Watanabe, Jerry Tang, Lawrence Hon, Reece Hart, and Brian Desany for helpful discussions; Thomas Wu and Guy Cavet for their detailed comments on the manuscript; and William Wood for guidance and support throughout the project.

References

1. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
2. Stephens P, Edkins S, Davies H, et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat Genet 2005;37:590–2.
3. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 2004;350:2129–39.
4. O'Hare T, Pollock R, Stoffregen EP, et al. Inhibition of wild-type and mutant Bcr-Abl by AP23464, a potent ATP-based oncogenic protein kinase inhibitor: implications for CML. Blood 2004;104:2532–9.
5. Paez JG, Janne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304:1497–500.
6. Pao W, Miller V, Zakowski M, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A 2004; 101:13306–11.
7. Shah NP, Tran C, Lee FY, Chen P, Norris D, Sawyers CL. Overriding imatinib resistance with a novel ABL kinase inhibitor. Science 2004;305:399–401.
8. Vierimaa O, Georgitsi M, Lehtonen R, et al. Pituitary adenoma predisposition caused by germline mutations in the AIP gene. Science 2006;312:1228–30.
9. Landi MT, Bauer J, Pfeiffer RM, et al. MC1R germline variants confer risk for BRAF-mutant melanoma. Science 2006;313:521–2.
10. Parsons DW, Wang TL, Samuels Y, et al. Colorectal cancer: mutations in a signalling pathway. Nature 2005; 436:792.
11. Bonetta L. Going on a cancer gene hunt. Cell 2005; 123:735–7.
12. Bardelli A, Parsons DW, Silliman N, et al. Mutational analysis of the tyrosine kinome in colorectal cancers. Science 2003;300:949.
13. Wang ZC, Lin M, Wei LJ, et al. Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. Cancer Res 2004;64:64–71.
14. Samuels Y, Velculescu VE. Oncogenic mutations of PIK3CA in human cancers. Cell Cycle 2004;3:1221–4.
15. Samuels Y, Wang Z, Bardelli A, et al. High frequency of mutations of the PIK3CA gene in human cancers. Science 2004;304:554.
16. Davies H, Hunter C, Smith R, et al. Somatic mutations of the protein kinase gene family in human lung cancer. Cancer Res 2005;65:7591–5.
17. Beckman RA, Loeb LA. Genetic instability in cancer: theory and experiment. Semin Cancer Biol 2005;15: 423–35.
18. Lengauer C. Cancer. An unstable liaison. Science 2003;300:442–3.
19. Takano T, Ohe Y, Sakamoto H, et al. Epidermal growth factor receptor gene mutations and increased copy numbers predict gefitinib sensitivity in patients with recurrent non-small-cell lung cancer. J Clin Oncol 2005;23:6829–37.
20. Pao W, Wang TY, Riely GJ, et al. KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. PLoS Med 2005;2:e17.
21. Aouacheria A, Navratil V, Wen W, et al. In silico whole-genome scanning of cancer-associated nonsynonymous SNPs and molecular characterization of a dynein light chain tumour variant. Oncogene 2005;24: 6133–42.
22. Babenko VN, Basu MK, Kondrashov FA, Rogozin IB, Koonin EV. Signs of positive selection of somatic mutations in human cancers detected by EST sequence analysis. BMC Cancer 2006;6:36.
23. Bairoch A, Apweiler R, Wu CH, et al. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33: D154–9.
24. Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A 2004;101: 15398–403.
25. Forbes S, Clements J, Dawson E, et al. Cosmic 2005. Br J Cancer 2006;94:318–22.
26. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.
27. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30:3894–900.
28. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics 2004; 20:1006–14.
29. 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; 21:2814–20.
30. Barenboim M, Jamison DC, Vaisman II. Statistical geometry approach to the study of functional effects of human nonsynonymous SNPs. Hum Mutat 2005;26: 471–6.
31. Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol 2006;356:1263–74.
32. Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 2003;19:2199–209.
33. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 2005;21:2185–90.
34. Wang Z, Shen D, Parsons DW, et al. Mutational analysis of the tyrosine phosphatome in colorectal cancers. Science 2004;304:1164–6.
35. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005;21:1859–75.
36. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1999;27:260–2.
37. Baker SJ, Fearon ER, Nigro JM, et al. Chromosome 17

deletions and p53 gene mutations in colorectal carcinomas. Science 1989;244:217–21.
38. Teng DH, Hu R, Lin H, et al. MMAC1/PTEN mutations in primary tumor specimens and tumor cell lines. Cancer Res 1997;57:5221–5.
39. Milde-Langosch K, Ocon E, Becker G, Loning T. p16/MTS1 inactivation in ovarian carcinomas: high frequency of reduced protein expression associated with hypermethylation or mutation in endometrioid and mucinous tumors. Int J Cancer 1998;79:61–5.
40. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. Nature 2002;417:949–54.
41. Nigro JM, Baker SJ, Preisinger AC, et al. Mutations in the p53 gene occur in diverse human tumour types. Nature 1989;342:705–8.
42. Nagata H, Worobec AS, Oh CK, et al. Identification of a point mutation in the catalytic domain of the protooncogene c-kit in peripheral blood mononuclear cells of patients who have mastocytosis with an associated hematologic disorder. Proc Natl Acad Sci U S A 1995;92:10560–4.
43. Heinrich MC, Corless CL, Duensing A, et al. PDGFRA activating mutations in gastrointestinal stromal tumors. Science 2003;299:708–10.
44. Kralovics R, Passamonti F, Buser AS, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. N Engl J Med 2005;352:1779–90.
45. Bardelli A, Velculescu VE. Mutational analysis of gene families in human cancer. Curr Opin Genet Dev 2005;15:5–12.
46. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.
47. Sjoblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–74.
48. SeattleSNPs. NHLBI program for genomic applications; 2006.
49. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet 2006;7:21–33.
50. Benvenuti S, Arena S, Bardelli A. Identification of cancer genes by mutational profiling of tumor genomes. FEBS Lett 2005;579:1884–90.
51. Furney SJ, Higgins DG, Ouzounis CA, Lopez-Bigas N. Structural and functional properties of genes involved in human cancer. BMC Genomics 2006;7:3.