Bioinformatics, 31(22), 2015, 3561–3568
doi: 10.1093/bioinformatics/btv430
Advance Access Publication Date: 25 July 2015
Original Paper

Genome analysis

Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data

Runjun D. Kumar[1,2], Adam C. Searleman[1], S. Joshua Swamidass[2,3], Obi L. Griffith[1,4] and Ron Bose[1,]*

[1]Division of Oncology, Department of Medicine, Washington University School of Medicine, [2]Computational and Systems Biology Program, Washington University in St Louis, [3]Department of Pathology and Immunology, Washington University School of Medicine and [4]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO 63110, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on March 24, 2015; revised on June 25, 2015; accepted on July 18, 2015

Abstract

Motivation: Several tools exist to identify cancer driver genes based on somatic mutation data. However, these tools do not account for subclasses of cancer genes: oncogenes, which undergo gain-of-function events, and tumor suppressor genes (TSGs) which undergo loss-of-function. A method which accounts for these subclasses could improve performance while also suggesting a mechanism of action for new putative cancer genes.

Results: We develop a panel of five complementary statistical tests and assess their performance against a curated set of 99 HiConf cancer genes using a pan-cancer dataset of 1.7 million mutations. We identify patient bias as a novel signal for cancer gene discovery, and use it to significantly improve detection of oncogenes over existing methods (AUROC = 0.894). Additionally, our test of truncation event rate separates oncogenes and TSGs from one another (AUROC = 0.922). Finally, a random forest integrating the five tests further improves performance and identifies new cancer genes, including CACNG3, HDAC2, HIST1H1E, NXF1, GPS2 and HLA-DRB1.

Availability and implementation*:* All mutation data, instructions, functions for computing the statistics and integrating them, as well as the HiConf gene panel, are available at www.github.com/Bose-Lab/Improved-Detection-of-Cancer-Genes.

Contact*:* rbose@dom.wustl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Since the first cancer genome was sequenced in 2008, large-scale studies surveying multiple tumor types have been released (Kandoth *et al.*, 2013; Lawrence *et al.*, 2014; Ley *et al.*, 2008). By sequencing paired tumor-normal exomes and genomes, these studies generate catalogs of tumor-specific mutations, such as single nucleotide variants and small insertions/deletions (indels). Mutation data can be used to detect evidence of positive selection and new potential cancer genes, defined here as genes that exert a pro-tumor influence as either oncogenes or tumor suppressors genes (TSGs). However, tumor mutation rates vary by several orders of magnitude, increasing from acute myeloid leukemia to melanoma (Lawrence *et al.*, 2013). Particularly in tumors with high mutation rates, it is likely that most mutations occur incidentally to tumor development. Therefore, mutations and genes must be experimentally characterized to identify biologically relevant mutations; however, due to the number of background mutations, these experiments pose a high risk, low reward opportunity. In general, only genes with strikingly nonrandom mutation patterns or new mutations in known cancer genes prompt further investigation (Bose *et al.*, 2013).

Computational solutions can alleviate this problem by directing investigators to genes that are more likely to be cancer-causing. Existing methods detect several signals of cancer gene positive selection, including: mutation rate (Dees *et al.*, 2012), functional impact scores (Gonzalez-Perez and Lopez-Bigas, 2012), intra-gene mutation clustering and recurrence (Tamborero *et al.*, 2013), post-translational modifications (Reimand and Bader, 2013), or DNA lesion likelihood (Hua *et al.*, 2013). Earlier work also demonstrated the importance of patient-specific mutation rates in detecting cancer genes (Youn and Simon, 2011). Furthermore, integrating these complementary approaches can lead to improved performance (Davoli *et al.*, 2013; Tamborero *et al.*, 2013).

One shortcoming of these methods is that they identify cancer genes, but cannot separate likely oncogenes and tumor suppressors. Oncogenes are of particular interest to biologists, as they can provide a direct target for small molecule inhibitors. However, recent studies show that tumor suppressors and oncogenes are separable using rates of truncating mutations, mutation clustering and copy-number data (Schroeder *et al.*, 2014) (see Supplementary Fig. S1 for representative examples). It is possible that an ensemble method that treats oncogenes and TSGs as separate classes would improve performance over existing methods, in addition to suggesting whether new putative cancer genes operate through gain-of-function (oncogene) or loss-of-function (TSG).

In this study, we use a pan-cancer dataset of 1.7 million mutations and manually curated set of 99 high confidence (HiConf) cancer genes to develop a panel of five statistical tests. The tests detect different signals of positive selection and are designed to detect putative oncogenes and TSGs. In particular, we identify patient and cancer type bias as new cancer gene signals, and leverage them to markedly improve detection of oncogenes. We then integrate these tests into a random forest model which can simultaneously identify new putative cancer genes, and classify them as likely oncogenes and tumor suppressors. We validate by assessing the performance of previous tools and our new methods against several independent panels of known and putative cancer genes. Finally, we explore the performance of these methods in specific cancer types and suggest new putative cancer genes.

2 Materials and methods

2.1 Data gathering and quality control

Mutation Annotation Files were drawn from data repositories for the TCGA, ICGC and COSMIC (Supplementary Table S1). Only columns for Cancer Type, Study, Patient Identifier, Chromosome, Start Position, End Position, Reference and Variant Allele were retained. A small number of hg18-based studies (accounting for ~2% of the dataset) were converted to hg19 using the UCSC Genome Browser liftOver utility with default settings (Fujita *et al.*, 2011). Patient samples were frequently included in more than one dataset, potentially producing duplicate or contradictory mutations. For a given patient and genomic position, only mutations from the most recent dataset were retained. Data were annotated with the ANNOVAR software suite using RefSeq libraries (Wang *et al.*, 2010). Mutations were also labeled with functional impact scores to allow the use of Oncodrive-fm (Gonzalez-Perez and Lopez-Bigas, 2012). In cases where a gene was related to multiple transcripts and isoforms, the transcript which preserved the most mutations was used first, with mutations from alternate isoforms being annotated as such. Data were gathered July 27 to August 1, 2013.

2.2 HiConf cancer gene panel construction

As our goal is to use mutation data to find potential cancer genes that can be confidently carried into biological experiments, we sought HiConf cancer genes that are biologically established as a training data set. This HiConf panel was focused towards known cancer genes that have previously been detected through genetic criteria, and which could plausibly be detected with exome sequencing data. The following steps were taken to ensure the HiConf cancer gene panel met these criteria. The Cancer Gene Census (CGC) provided candidates (Futreal *et al.*, 2004). Genes which have only been observed in translocations (as per the CGC annotations) were immediately eliminated, as our dataset lacks translocation events and it is often unclear whether translocation partners are active cancer genes individually. This left 204 candidate genes.

A literature search was then performed. A gene qualified for the HiConf panel if a scientific publication could be found which fulfilled one of the following: (i) Demonstrated a cancer-like phenotype in cell lines when the gene was activated or inhibited. (ii) Demonstrated a change in disease progression in mouse models of cancer when the gene was activated or inhibited. (iii) Demonstrated the gene as a causative agent of a Mendelian human tumor syndrome. Importantly, all means of gene alteration (RNAi, ectopic expression, drug or antibody targeting, null models, etc.) were accepted for animal and cell studies, and any phenotype outlined in the Hallmarks of Cancer was accepted as cancer-like (Hanahan and Weinberg, 2011). The sources for the literature search included OMIM and PubMed. The literature search left 99 HiConf cancer genes. Based on the preponderance of literature recovered, these genes were further categorized as oncogenes (ONCs, gain of function causes pro-cancer phenotype) or TSGs (loss of function causes pro-cancer phenotype).

2.3 Tool import

Three existing tools were applied to the dataset: MutSigCV (Lawrence *et al.*, 2013), OncodriveCLUST (Tamborero *et al.*, 2013) and Oncodrive-fm (Gonzalez-Perez and Lopez-Bigas, 2012).

MutSigCV identifies likely cancer genes by detecting genes with elevated mutation rates. MutSigCV v1.3 was run on the dataset using scripts downloaded from http://www.broadinstitute.org/cancer/cga/mutsig following the provided instructions and default settings. The MutSigCV *P*-value was used to assess tool performance. OncodriveCLUST is a cancer gene detection method which uses intra-protein mutation clustering to identify possible cancer genes. Software was downloaded from http://bg.upf.edu/group/projects/oncodrive-clust.php, and run on the dataset using the included instructions and default settings. Oncodrive-fm detects genes with unusually impactful mutations, as judged by a suite of functional impact scores (SIFT, PolyPhen2 and MutationAssessor) (Gonzalez-Perez and Lopez-Bigas, 2012). The Oncodrive-fm method was re-implemented in R, using the dataset itself as an internal null distribution.

2.4 Test development

We assembled five statistical tests that target several signals of positive selection in cancer genes. 'Patient Distribution' and 'Cancer Type Distribution' operate similarly and detect genes that are mutated in nonrandom sets of patients or cancer types. 'Unaffected

Residues' is our method to identify genes with unusual levels of mutation recurrence. 'VEST Mean' uses VEST scores (Carter *et al.*, 2013) to identify functional impact bias among genes. Finally, 'Truncation Rate' is our approach to detecting genes that have unusual numbers of truncation events; either an enrichment (as is expected of TSGs) or depletion (as is expected of oncogenes).

'Patient Distribution' and 'Cancer Type Distribution' are calculated similarly. Each mutation occurs within a patient (or cancer type). A randomly mutated gene should be mutated in a random set of patients (or cancer types). This null hypothesis can be tested using the Pearson Chi-Square Goodness-of-Fit test, with the entire dataset providing the null expectations. For each gene g, a chi-square statistic was calculated:

$$X_g^2 = \sum_{p=1}^{p} \frac{(O_p - E_p)^2}{E_p} \qquad E_p = \frac{N_p N_g}{N}$$

where O is the observed count of mutations for a given patient (or cancer type), E is the expected number of mutations for the same patient (or cancer type), and P is the number of unique patients (or cancer types). The expected count for a given patient (or cancer type) and gene is the product of the total number of mutations in the patient (N_{p}) and the total number of mutations in the gene (N_{g}) divided by the number of mutations in the dataset (N). The P-value is calculated by simulation since the low expectations would violate normality assumptions required to use the theoretical chi-square distribution. Given the number of mutations in a gene, the test statistic is calculated for 10 000 random draws from the full list of patient (or cancer type) labels with replacement, and the upper tail probability of a higher test statistic under the null distribution is reported. All mutations, including synonymous mutations, are used when calculating 'Patient Distribution' and 'Cancer Type Distribution'.

'Unaffected Residues' detects high levels of recurrence by considering the number of un-mutated residues in a gene. First, given the number of mutations and the protein length, the probability of a residue being un-mutated is calculated based on the Poisson distribution. Because the mutation count is zero, the estimated probability of an unaffected residue simplifies to:

$$\hat{P}_{\text{zero}} = e^{-n/l}$$

Where n is the number of mutations in the protein, l is protein length, and $\hat{P}_{\text{zero}}$ is the estimated probability of a given residue being un-mutated. Once $\hat{P}_{\text{zero}}$ is calculated, the binomial distribution is used to calculate the probability of a gene having at least the observed number of unaffected residues:

$$P(X \geq x) = \sum_{i=x}^{l-1} \binom{l}{i} \hat{P}_{\text{zero}}{}^{i} (1 - \hat{P}_{\text{zero}})^{l-i}$$

Where x is the observed number of unaffected residues and l is the protein length. 'Unaffected Residues', represents the probability of a gene having as many or more unaffected residues as observed if mutation location is entirely random. Only nonsynonymous protein-coding mutations are used to calculate this test, as recurrent synonymous mutations can suggest alignment errors and may produce false positives.

'VEST Mean' is calculated in a very similar manner as the individual sub-scores used within Oncodrive-fm (Gonzalez-Perez and Lopez-Bigas, 2012), but uses the Variant Effect Scoring Tool as the base functional impact score (Carter *et al.*, 2013). It is the upper tail probability of a gene having a mean VEST score greater than that observed, given the number of mutations, based on 10 000 random draws with replacement from all observed VEST scores. VEST scores are limited to missense mutations, so imputation was required for other mutations. We used the same rationale as was used in Oncodrive-fm. Synonymous and non-coding mutations were assigned a value of 0, the lowest functionality score under VEST, while in-frame and frameshift indels, premature stop, nonstop and splice site mutations were assigned the highest value of 1. Synonymous and nonsynonymous mutations are used in this calculation.

To use 'Truncation Rate', a gene's mutations are categorized as truncating (i.e. Splice Site, Frameshift Insertion/Deletion, Premature Stop/Nonsense) or non-truncating. Then the upper tail binomial probability of at least the observed number of truncation events (using the truncation rate across the dataset for the null distribution) is calculated as:

$$P(T > t) = \sum_{i=t}^{n} \binom{n}{i} \hat{P}_{Trunc}{}^{i} (1 - \hat{P}_{Trunc})^{n-i}$$

Where $\hat{P}_{\text{trunc}} = 182\,030$ truncation events / $1\,703\,709$ mutations $= 0.107$, t is the observed number of truncating events for a given gene and n is the total number of mutations in the gene. Synonymous and non-synonymous mutations are used in this calculation.

2.5 Imputation

Our tests rely on very basic annotations (e.g. Sample ID, Cancer Type, Mutation Type, etc.) and consequently we had very low rates of missingness. Two exceptions warrant note, and both are related to 'Unaffected Residues'. This test requires a valid protein length to be calculated; however, after integrating datasets, ~4% of genes had protein lengths smaller than the most downstream mutations. In these cases, the test uses the most downstream mutation position as a conservative proxy of protein length. The other exception is in model training. Most of our tests are calculable for virtually all genes. The exception is 'Unaffected Residues', which cannot be calculated for the ~10% of genes with no coding nonsynonymous mutations. The data matrix was filled in by mean imputation prior to model training. Missing values were excluded from the calculation or assessment of individual tests.

2.6 Model generation

We compared Random Forests, SVMs and Naïve Bayes classifiers in separating the three gene classes (Unknown Function, HiConf Oncogenes, HiConf TSGs) using the individual tests of our panel. Random Forests and SVMs both performed well. Random Forests were chosen because they have been used in previous tools such as OncodriveROLE (Schroeder *et al.*, 2014) and worked well with default settings ($mtry = 2$, trees $= 500$).

To generate the scores and predictions used in the study, we trained a random forest (RF5) on the five individual tests ('Patient Distribution', 'Cancer Type Distribution', 'Unaffected Residues', 'Truncation Rate', 'VEST Mean') and labels generated from the HiConf panel (22 801 unknown genes, 48 TSGs, 51 Oncogenes). TSGs and ONCs were up-sampled during training to better calibrate the model (trees were trained on 300 unknowns, 30 TSGs and 30 ONCs). Five-fold cross validation was used to generate predictions, repeated 50 times. The repeated cross validation runs were averaged to generate the stable predictions presented in the study.

2.7 Validation gene panels

In addition to the manually curated HiConf gene panel, we also sought out additional panels of established cancer genes.

These panels are necessary to validate the performance of our random forest model, since even with cross validation its performance on the HiConf panel could be over-optimistic.

We gathered the high confidence driver (HCD), Cancer5000 and TSGene lists as presented in Schroeder *et al.* (2014). These were originally generated by Tamborero *et al.* (2013), Lawrence *et al.* (2014) and Zhao *et al.* (2013), respectively (Lawrence *et al.*, 2014; Tamborero *et al.*, 2013; Zhao *et al.*, 2013). In addition to the filters applied by Schroeder and colleagues, we ensured independence by depleting these lists for any members of the HiConf panel, leaving 149, 96 and 55 genes in the respective panels. Note that while they are independent of the HiConf list, they do overlap with one another.

Although the HCD and Cancer5000 lists may contain both oncogenes and TSGs, the TSGene list is composed of TSGs exclusively. To generate an oncogene-only list, we defined the Kinase list as any kinase bearing a known activating cancer mutation in Kin-Driver (Simonetti *et al.* 2014). This panel consists of 29 genes after being depleted of HiConf genes.

2.8 Cancer subset analysis

Cancers with at least 500 patients or 200 000 mutations were considered in the cancer type analysis. Tests and RF5 models were applied using identical procedures to the pan-cancer analysis.

2.9 Statistics and software

All comparisons of AUROCs were performed as two-sided DeLong Tests (DeLong *et al.*, 1988) with adjustment for ties. All analyses were performed in R v2.15. Modeling was performed using methods available through the randomForest (Breiman, 2001) and e1071 (Support Vector Machines, Naïve Bayes) packages. AUROCs, ROC plots and DeLong Tests were performed using the pROC package.

2.10 Data availability

The dataset is available along with a script, instructions and sample data to be used to train RF5 models on any dataset. Please see www.github.com/Bose-Lab/Improved-Detection-of-Cancer-Genes.

3 Results

3.1 Data assembly and quality control

The analytic flow follows the schema in Supplementary Figure S2. Data sources are listed in Supplementary Table S1. The final dataset includes 1 703 709 mutations across 10 239 patients (Supplementary Fig. S3). In total 22 902 genes appear at least once in the dataset, with a median of 49 mutations per gene. This is one of the largest assembled pan-cancer data sets and is publicly accessible (See Materials and Methods for more details).

3.2 Developing a panel of known cancer genes

Based on a criteria-driven literature review, 99 genes were collected into a high-confidence cancer gene panel (HiConf, see 'Methods' and Supplementary Table S2 for details). The HiConf gene panel was further divided into 48 TSGs (with 15 698 mutations) and 51 ONCs (with 11 243 mutations). Rather than defining a separate set of presumptively neutral genes, the remaining 22 801 genes were labeled as 'unknown'. Most unknown genes are neutral with regards to cancer progression, and the set as a whole is treated as neutral for the purposes of training and assessment.

3.3 Test design and assessment

Individual methods of cancer gene prediction must separate the distinct mutation patterns of ONCs, TSGs and neutral genes. In particular, TSGs tend to be enriched in truncation events, while oncogenes are depleted; in addition, oncogenes tend to have clustered mutations (Supplementary Fig. S1). We performed statistical tests for each of five signals of positive selection, and refer to them as follows: 'Truncation Rate' (rate of truncating events), 'Unaffected Residues' (intra-gene mutation clustering/recurrence), 'VEST Mean' (functional impact bias), 'Patient Distribution' (bias in patient labels) and 'Cancer Type Distribution' (cancer type bias). OncodriveCLUST, Oncodrive-fm, and MutSigCV were also applied to the dataset (Gonzalez-Perez and Lopez-Bigas, 2012; Lawrence *et al.*, 2013; Tamborero *et al.*, 2013).

We use the Area Under Receiver Operator Characteristic (AUROC) to gauge performance as it is threshold independent and testable (DeLong *et al.*, 1988). In particular, we consider the following classification tasks: separation of the HiConf ONCs and TSGs from other genes of unknown function (UK) as separate and pooled classes, and separation of ONCs and TSGs from one another.

'Patient Distribution' is notable because it relies on a novel cancer gene signal which we call patient bias. The contribution of tumors to the pan-cancer dataset is highly unequal because tumor mutation rates vary by up to four orders of magnitude (Supplementary Fig. S3). However, mutations within HiConf TSGs and oncogenes are much more evenly distributed between patients (Fig. 1A). 'Patient Distribution' makes use of a chi-square statistic to detect genes which are frequently mutated in relatively hypomutated tumors. Unlike many of the other statistics and tools we assessed, 'Patient Distribution' detects HiConf oncogenes and HiConf TSGs equally well (Fig. 1B and Table 1). In fact, it is the single best test for detecting oncogenes and the HiConf panel as a whole, with AUROCs of 0.894 and 0.900, respectively. Known oncogenes

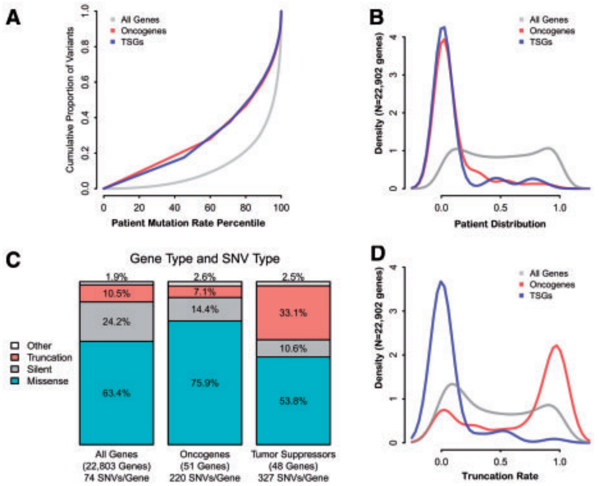

Fig. 1. Patient distribution and truncation rate. **(A)** Patients have unequal mutation rates, but this effect is less pronounced when considering only HiConf ONCs and TSGs. **(B)** Patient Distribution is the *P*-value from a chi-square goodness-of-fit test for the distribution of patients a gene is mutated in, versus the distribution of patients generally. Patient distribution can separate ONCs and TSGs from most other genes, but not from one another. **(C)** Distribution of mutation types for each of three gene types. TSGs are relatively enriched for truncating events (nonsense, frameshift and splice site) while ONCs are depleted. **(D)** Truncation Rate is the binomial upper tail probability of a gene having an equal or higher percentage of truncating mutations. Truncation Rate can separate HiConf TSGs and ONCs from one another

including TRAF7 and ALK are missed by previously published tools at the $P < 0.05$ cutoff, but are easily detected by 'Patient Distribution' (Supplementary Table S3).

'Cancer Type Distribution' is very similar to 'Patient Distribution', but relies on cancer type bias to identify cancer genes. For instance, it easily highlights VHL, a HiConf tumor suppressor which is frequently truncated in renal clear cell carinomas (Supplementary Table S3). It also identifies the known tumor suppressor PTCH1, which is not identified by existing tools.

'Unaffected Residues' is our test of mutation clustering and recurrence. Rather than testing for clustering directly, as was the approach taken by OncodriveCLUST (Tamborero *et al.*, 2013), we instead examine the number of unmutated residues. 'Unaffected Residues' is a one-tailed binomial test for the number of unmutated residues, assuming the number of mutations per residue is poisson distributed. It is the second best method for detecting HiConf ONC (AUROC = 0.855) and third best for the whole HiConf panel (AUROC = 0.861). It is superior to Oncodrive-CLUST in these tasks (Table 1). The top four genes according to 'Unaffected Residues' are KRAS, PIK3CA, BRAF and TP53 (Supplementary Table S3), all of which have well known mutation clusters.

'VEST Mean' tests for functional impact bias. It reports the probability of a randomly mutated gene having a higher average functional impact, very similar in concept to the method used in Oncodrive-fm (Gonzalez-Perez and Lopez-Bigas, 2012), but with better oncogene detection (AUROC = 0.796 versus 0.710, Table 1). It is also the best method for detecting TSGs (AUROC = 0.938).

'Truncation Rate' is the final test and has unique properties. It is well appreciated that TSGs are enriched in protein-truncating mutations (Fig. 1C). This pattern lead Vogelstein and colleagues to suggest that genes with >20% truncating events be considered putative TSGs (Vogelstein *et al.*, 2013). 'Truncation Rate' formalizes this concept using a one-sided binomial test, which reports the probability of a randomly mutated gene bearing an equal or greater number of truncation events (i.e. splice site, premature stop and frameshift indels), given a fixed number of mutations. However, as a one-sided test, 'Truncation Rate' is also sensitive to the relative depletion of truncation events in oncogenes (Fig. 1D). It is by far the best method for separating oncogenes and TSGs (AUROC = 0.922, Table 1, ROC curves in Supplementary Fig. S4).

3.4 Integration into a single model

As Table 1 illustrates, the tests we have identified are complementary, each having different performances in our classification tasks. We hypothesized that a model integrating the full panel would be able to separate all three gene classes (HiConf Oncogenes, HiConf TSGs, all other genes of unknown function) from one another. To test this hypothesis, we trained a Random Forest model on the five individual test values as predictor variables. Gene labels were generated from the HiConf panel, resulting in 51 ONC, 48 TSG, and 22 801 unknown (UK) genes. Most of the UK genes are passenger genes, so this large class serves as a neutral class for training. Training was performed in five-fold cross validation, with results averaged over 50 repetitions.

The five-test model, which we refer to as RF5, produces a score for probability of membership in each class. These scores summate to 1, allowing genes to be visualized in a ternary plot (Fig. 2). UK genes which are placed near the ONC and TSG regions are putative cancer genes, while HiConf ONCs and TSGs which are assigned to the Unknown region are false negatives.

Table 1. AUROCs of individual tests and RF5 model

	ONC + TSG versus UK	TSG versus UK + ONC	ONC versus UK + TSG	ONC versus TSG	In RF5	Description
RF5	**0.935**	**0.980**	**0.891**	**0.924**^{b}		**Cross-validation predictions of a Random Forest trained on the five indicated features and the HiConf panel.**
Patient Distribution	0.900	0.905	0.894*	0.556	Yes	Detects deviation from expected patient distribution with a chi-square statistic. *P* values by resampling.
Truncation Rate	0.788^a	0.904	0.694	0.922*	Yes	Detects enrichment or depletion of Frameshift Indels, Nonsense and Splice Site events using binomial distribution.
Unaffected Residues	0.861	0.865	0.855*	0.479	Yes	Detects clustering using poisson and binomial distributions to calculate probability of unaltered residues.
VEST Mean	0.866	0.938	0.796	0.710	Yes	Detects high-functional impact (based on VEST3). *P*-values from resampling.
Cancer Type Distribution	0.853	0.905	0.803	0.612	Yes	Detects deviation from expected cancer distribution with a chi-square statistic. *P* values by resampling.
MutSigCV	0.760	0.896	0.632	0.723	No	*P*-value retrieved from MutSigCV. Detects high rates of mutation based on gene-specific background mutation rate.
Oncodrive CLUST	0.776	0.741	0.808	0.597	No	*P*-value retrieved from OncodriveCLUST summary report. Detects high rates of clustering.
Oncodrive-fm	0.818	0.932	0.710	0.725	No	*P*-value retrieved from Oncodrive-fm. Detects high rates of functional events using several functional impact scores.

AUROC, area under receiver operator characteristic, for the separation of the indicated gene classes; ONC, HiConf Oncogenes; TSG, HiConf Tumor Suppressor Genes; UK, genes of unknown relevance to cancer growth.

*These performances are *not* significantly different from RF5 performance at $P < 0.05$.

[a] Truncation Rate is converted to a two-tail test when identifying the combined HiConf panel.

[b] The ratio of the RF5 TSG and ONC scores is used to separate these classes.

As Figure 2 shows, RF5 is able to delineate most HiConf ONCs and TSGs from the bulk of UK genes. It also suggests a large number of UK genes which appear similar to ONCs and TSGs. Simultaneously, RF5 separates the ONCs and TSGs from one another. When assessed for performance at each task separately, RF5 is significantly better or not significantly different from the best individual tests. It performs markedly better than 'VEST Mean' in detecting TSGs (AUROC = 0.980 versus 0.938), the same as 'Patient Distribution' in detection ONCs (AUROC = 0.891 versus 0.894), and the same as 'Truncation Rate' in separating ONCs and TSGs (AUROC = 0.924 versus 0.922, Table 1). HiConf genes which are identified by few of the individual tests can often be identified confidently by RF5, demonstrating the importance of integrating multiple approaches (Supplentary Fig. S5A).

3.5 Validation panels

The HiConf panel serves as our primary method of assessment for pre-existing tools and our new methods. However, RF5 is trained to detect the HiConf panel, and it is possible the RF5 performance estimates are optimistic even with cross-validation. Therefore, we retrieved four validation panels and depleted them of the HiConf panel members to ensure independence (see 'Methods'). We then assessed the ability of our methods to prioritize the validation panels over other genes of unknown function (Table 2).

The HCD panel was defined by Tamborero *et al.* (2013) using a variety of existing tools including Oncodrive-fm (Tamborero *et al.*, 2013). After excluding HiConf cancer genes, it consists of 149 members. We find that RF5 has the best performance (AUROC = 0.884) on this set, but that 'VEST Mean' and Oncodrive-fm are not significantly different. This expected, as the list was defined in part using Oncodrive-fm. We also examined the Cancer5000 gene panel, which has 96 members after depletion of HiConf genes (Lawrence *et al.*, 2014). It overlaps by roughly 50% with the HCD panel, and RF5 still has the strongest performance (AUROC = 0.943). This panel was defined using MutSigCV, which performs well as expected (AUROC = 0.882).

Unlike the HCD and Cancer5000 panels, the TSGene and Kinase panels are largely composed of TSGs and oncogenes, respectively (see 'Methods'). The TSGene panel consists of 55 manually curated tumor suppressors (Zhao *et al.*, 2013). 'VEST Mean' has the highest performance (AUROC = 0.876), but RF5 is not significantly worse. The Kinase panel consists of 29 manually curated kinases that are known to harbor activating mutations in cancer (Simonetti *et al.*, 2014). RF5 again has the strongest performance (AUROC = 0.801).

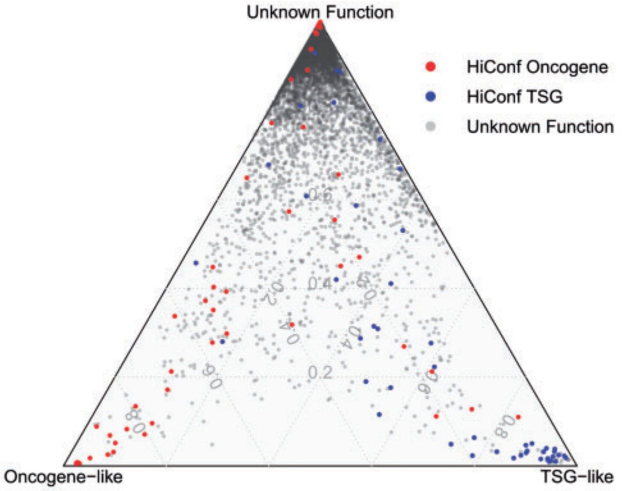

Fig. 2. RF5 model predictions. Cross-validated predictions of the random forest model. N = 22 902 genes. Using the three class-specific scores generated by RF5, genes can be stratified as oncogene- or TSG-like. Genes which are judged as oncogene- or TSG-like, but are not on the HiConf panel, are putatively related to cancer

3.6 Predicted cancer genes

For brevity and clarity we will focus on the top 100 predictions made by RF5 in the pan-cancer setting. They include many potentially new cancer genes, of which we will highlight a few (Supplementary Fig. S5B and S6). Several genes are related to chromatin structural and epigenetic regulation. GPS2 and HDAC2 are members of the NCOR-HDAC3 complex (Zhang *et al.*, 2002) and are predicted TSGs. HIST1H1E, a linker protein in nucleosomes, is predicted to be an oncogene. Other novel predicted cancer genes are drawn from a range of biological classes: CACNG3 (predicted oncogene) is a voltage-dependent calcium channel subunit; NXF1 (predicted TSG) is a nuclear RNA export factor; and HLA-DRB1 (predicted TSG) is a subunit of MHC Class II. Additionally, several experimentally known cancer genes are linked to human tumors through somatic mutation data for the first time. Among these are the oncogenes SGK1 (Towhid *et al.*, 2013) and TMEM30A (Kato *et al.*, 2013) as well as the TSGs RBM5 (Sutherland *et al.*, 2010), CHD4 (Cai *et al.*, 2014) and CHD2 (Nagarajan *et al.*, 2009). Although these are not new cancer genes, their identification by patterns of somatic mutations supports their relevance in human disease. None of these genes are listed in the Cancer Gene Census.

Many top predicted cancer genes are potentially druggable. A query of the Drug Gene Interaction Database reveals that 26 of the top 100 predicted cancer genes have known interactions with drugs, and an additional 43 belong to a potentially druggable gene category (Griffith *et al.*, 2013). With the majority of top predictions being potentially druggable, the prioritized gene list presents opportunities for both new discoveries in cancer biology and more immediate pharmacologic interventions.

RF5 also makes high-quality predictions. For instance, very few of the top 100 predicted cancer genes are biologically implausible. Among these genes, there is one olfactory receptor (OR4C5) and one collagen (COL2A1) (Lawrence *et al.*, 2013). However, technical artifacts remain a concern. For instance, the highly ranked genes IL32 and PLAC4 have multiple recurrent frameshift and synonymous events. An examination of alignment files from several of the

Table 2. AUROCs of individual tests and RF5 in validation panels

	HCD	Cancer5000	TSGene	Kinases
RF5	**0.884**	**0.943**	**0.843**	**0.801**
Patient Distribution	0.713	0.768	0.644	0.745*
Truncation Rate	0.751[a]	0.836[a]	0.711	0.515
Unaffected Residues	0.825	0.849	0.792*	0.761*
VEST Mean	0.876*	0.921*	0.876*	0.731*
Cancer Type Distribution	0.811	0.861	0.760	0.612
MutSigCV	0.801	0.882	0.771*	0.525
OncodriveCLUST	0.686	0.742	0.699	0.657
Oncodrive-fm	0.876*	0.923*	0.835*	0.627

*These performances are *not* significantly different from RF5 at $P < 0.05$.

[a] Truncation Rate is calculated as a two-tail test for the HCD and Cancer5000 panels, as they combine TSGs and oncogenes.

affected patients suggests these genes are prone to alignment errors (data not shown). These examples illustrate the need for human expertise in scrutinizing prioritized gene lists, and the quality of data and associated mutation calls in particular.

3.7 Application to specific cancer types

Of our tests, only 'Cancer Type Distribution' relies on multiple cancer types; the others may perform differently in individual cancers. To address this possibility, the tests from Table 1 as well as RF5 models were generated for each cancer with at least 500 patients or 200 000 mutations (breast, colorectal, lung, melanoma and endometrial cancers, Supplementary Tables S4–S8). This analysis demonstrates that the relative performance of these tests is quite consistent across cancer types and that our new methods outperform previous tools in a variety of settings (Supplementary Fig. S7).

Analyzing the pan-cancer set may allow us to detect additional cancer genes due to increased power, but it may also mask cancer-specific cancer genes. To explore this possibility, we examined the detection of the HiConf panel in the pan-cancer and cancer-specific datasets. We found that 11 HiConf cancer genes were detected in the pan-cancer dataset, but not in the individual cancers, while 10 were detected in at least one of the specific cancers, but not in the pan-cancer set (Supplementary Fig. S8A). This suggests that we are likely to make some predictions only in the pan-cancer set, and others only in specific cancers. In fact, we found that 30 of our top 100 pan-cancer predicted cancer genes could only be detected in the pan-cancer set (Supplementary Fig. S8B). These included many promising potential cancer genes such as HDAC2, NXF1 and TMEM30A, illustrating the value of pooling cancers.

We then sought cancer genes that are cancer-specific and compared their detection across cancer types. We gathered the top 100 predictions for each of breast, colorectal, lung, melanoma and endometrial cancers. Roughly half of the top predictions were cancer-specific (Supplementary Fig. S9). A few examples include: MED23 and MYB as putative TSGs in breast cancer; TGIF1 and B3GNT6 as putative TSGs in colorectal cancer; CDK14, IRF2BPL and NTRK2 as putative oncogenes in lung adenocarcinoma; CCDC28B and ATAD2 as potential TSGs in melanoma; and EIF3C as a potential oncogene in endometrial cancer. We conclude that large numbers of cancer genes may be cancer-specific. Taken together, these results suggest the importance of searching for cancer-genes in the pan-cancer and specific-cancer settings.

Discussion

One use of cancer genome sequencing results is the identification of novel cancer genes. This problem has two stages: first, cancer genes must be separated from genes bearing only passenger mutations; second, cancer genes must be sorted into likely tumor suppressors and oncogenes. Both stages are crucial because mechanism-specific predictions are needed to guide downstream analyses and experiments. In this study we gathered a pan-cancer dataset of 1.7 million variants and a manually curated set of 99 known cancer genes (HiConf panel). Using these data, we designed and assessed a panel of statistical tests which identify cancer genes using several signals of positive selection, as well as separate cancer genes by mechanism of action. We also compared the performance of these tests to previous tools in accomplishing these tasks.

In general, we found that HiConf TSGs were easier to detect than HiConf oncogenes. Several methods had AUROCs of 0.9 or higher, including the published tool Oncodrive-fm and our tests of patient and cancer type bias ('Patient Distribution', 'Cancer Type Distribution'). However, the best single method for detecting TSGs was our test of functional impact bias, 'VEST Mean', with an AUROC of 0.938.

In contrast, HiConf oncogenes were less easily identified. This is concerning because oncogenes provide more direct targets for drug development. The best performing existing tool for detecting the HiConf oncogenes was OncodriveCLUST with an AUROC of 0.808. With the exception of 'Truncation Rate', all of the tests in our panel had AUROCs of 0.80 or greater when detecting HiConf oncogenes. Particular improvement was observed with 'Unaffected Residues', which tests for mutation clustering/recurrence and had an AUROC of 0.855, and 'Patient Distribution', which was the best performer with an AUROC of 0.894.

Two of our tests warrant emphasis. 'Patient Distribution' uses a novel signal of positive selection. It identifies genes with mutations that occur in nonrandom sets of patients, particularly genes with mutations that occur in relatively *hypo*-mutated tumors, as would be anticipated of genes bearing driver mutations. For identifying the HiConf panel as a whole (TSG + ONC), 'Patient Distribution' is the strongest performer with AUROC of 0.900. Another important member of our statistical panel is 'Truncation Rate'. This test is a formalized version of the 20/20 rule for TSGs put forward by Vogelstein *et al.* (2013), and uses the binomial distribution to model the expected number of truncation events per gene. 'Truncation Rate' can be used to separate TSGs and oncogenes with an AUROC of 0.922. It is the only method that usefully accomplished this task.

Because the individual tests of our panel offered complementary strengths, we also integrated them into a single model. We found that a random forest built on our five tests (RF5) was effective at separating HiConf oncogenes and TSGs from passenger genes, and from one another. Moreover, this integration did not require any loss in performance: RF5 is as good as or better than the individual methods at every classification tasks we assessed. We also confirmed these results in several independent validation gene panels.

RF5 identifies many potential pan-cancer cancer genes. These include the predicted oncogenes CACNG3 and HIST1H1E, and the predicted TSGs HDAC2, GPS2, NXF1 and HLA-DRB1. It also identifies several known cancer genes through genome sequencing for the first time, including SGK1, TMEM30A, CHD2, CHD4 and RBM5. Many RF5 predictions are potentially druggable. Furthermore, additional cancer genes can be identified when focusing on single cancer types. In fact, we found that half of RF5 predictions within tumor types were cancer-specific. These results illustrate the importance of searching for cancer genes in both the pan-cancer and specific-cancer settings and suggest many new potential avenues of research.

However, there remains room for improvement. As Figure 2 illustrates, some oncogenes and TSGs could not be detected by RF5, and some were not detectable by any individual test or pre-existing tool (Supplementary Fig. S5A). There are two major explanations. Foremost, cancer genes will be undetectable if they are primarily altered through means other than somatic mutations in the exome. Additionally, our cancer gene panels may include genes that are involved in later stages of disease progression, such as metastasis and drug resistance. These are true cancer genes, but may be undetectable in the available data as tumor samples largely come from newly diagnosed patients (Kandoth *et al.*, 2013). Fortunately, our methods are highly expandable, and multiple strategies could improve performance, such as: (i) Introduction of additional, heterogeneous data types. (ii) Improved tests. (iii) Improved model design and training. (iv) Expansion of the HiConf cancer gene panel.

In conclusion, our study demonstrates that the detection of putative cancer genes requires a mix of complementary methods. We have developed a panel of five statistical tests that outperform previous methods. In particular, 'Patient Distribution' detects oncogenes especially well. We have also integrated these tests into a single classifier, and demonstrated that it performs as well or better than previous tools in both training and validation cancer gene panels. These expansions to computational methods, along with targeted functional experiments, will lead to a more complete understanding of the cancer genome.

Acknowledgements

We thank The Cancer Genome Atlas, International Cancer Genome Consortium, and Catalogue of Somatic Mutations in Cancer for data access, as well as Elaine Mardis for helpful discussions and comments. We also thank Don Conrad, John Edwards and Kristen Naegle for critical readings of the article.

Funding

Our work was supported by the Alvin J. Siteman Cancer Center, the Ohana Breast Cancer Research Fund, and the Foundation for Barnes-Jewish Hospital (to R.B.), Canadian Institutes of Health Research (DFS-134967 to R.D.K.) and National Institutes of Health (K22CA188163 to O.L.G.).

Conflict of Interest: none declared.

References

Bose,R. *et al*. (2013) Activating HER2 mutations in HER2 gene amplification negative breast Cancer. *Cancer Discov.*, **3**, 224–237.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Cai,Y. *et al*. (2014) The NuRD complex cooperates with DNMTs to maintain silencing of key colorectal tumor suppressor genes. *Oncogene*, **33**, 2157–2168.

Carter,H. *et al*. (2013) Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, **14**(Suppl 3), S3.

Davoli,T. *et al*. (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, **155**, 948–962.

Dees,N.D. *et al*. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.

DeLong,E.R. *et al*. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.

Fujita,P.A. *et al*. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.

Futreal,P.A. *et al*. (2004) A census of human cancer genes, *Nat. Rev. Cancer*, **4**, 177–183.

Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers, *Nucleic Acids Res.*, **40**, e169.

Griffith,M. *et al*. (2013) DGIdb: mining the druggable genome, *Nat. Methods*, **10**, 1209–1210.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hua,X. *et al*. (2013) DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am. J. Human Genet.*, **93**, 439–451.

Kandoth,C. *et al*. (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.

Kato,U. *et al*. (2013) Role for phospholipid flippase complex of ATP8A1 and CDC50A proteins in cell migration. *J. Biol. Chem.*, **288**, 4922–4934.

Lawrence,M.S. *et al*. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.

Lawrence,M.S. *et al*. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.

Ley,T.J. *et al*. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66–72.

Nagarajan,P. *et al*. (2009) Role of chromodomain helicase DNA-binding protein 2 in DNA damage response signaling and tumorigenesis. *Oncogene*, **28**, 1053–1062.

Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637.

Schroeder,M.P. *et al*. (2014) OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics*, **30**, i549–i555.

Simonetti,F.L. *et al*. (2014) Kin-Driver: a database of driver mutations in protein kinases. *Database (Oxford)*, **2014**, bau104.

Sutherland,L.C., *et al*. (2010) RBM5 as a putative tumor suppressor gene for lung cancer. *J. Thorac. Oncol.*, **5**, 294–298.

Tamborero,D. *et al*. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.

Tamborero,D. *et al*. (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, **3**, 2650.

Towhid,S.T. *et al*. (2013) Inhibition of colonic tumor growth by the selective SGK inhibitor EMD638683. *Cell. Physiol. Biochem.*, **32**, 838–848.

Vogelstein,B. *et al*. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164–e164.

Youn,A. and Simon,R. (2011) Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, **27**, 175–181.

Zhang,J. *et al*. (2002) The N-CoR-HDAC3 nuclear receptor corepressor complex inhibits the JNK pathway through the integral subunit GPS2. *Mol. Cell*, **9**, 611–623.

Zhao,M. *et al*. (2013) TSGene: a web resource for tumor suppressor genes, *Nucleic Acids Res.*, **41**, D970–D976.