Quantifying gene selection in cancer through protein functional alteration bias

Nadav Brandes[1,*], **Nathan Linial**[1] **and Michal Linial**[2,*]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel and [2]Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Israel

Received April 10, 2019; Revised June 03, 2019; Editorial Decision June 08, 2019; Accepted June 16, 2019

ABSTRACT

Compiling the catalogue of genes actively involved in cancer is an ongoing endeavor, with profound implications to the understanding and treatment of the disease. An abundance of computational methods have been developed to screening the genome for candidate driver genes based on genomic data of somatic mutations in tumors. Existing methods make many implicit and explicit assumptions about the distribution of random mutations. We present FABRIC, a new framework for quantifying the selection of genes in cancer by assessing the effects of *de-novo* somatic mutations on protein-coding genes. Using a machine-learning model, we quantified the functional effects of ∼3M somatic mutations extracted from over 10 000 human cancerous samples, and compared them against the effects of all possible single-nucleotide mutations in the coding human genome. We detected 593 protein-coding genes showing statistically significant bias towards harmful mutations. These genes, discovered without any prior knowledge, show an overwhelming overlap with known cancer genes, but also include many overlooked genes. FABRIC is designed to avoid false discoveries by comparing each gene to its own background model using rigorous statistics, making minimal assumptions about the distribution of random somatic mutations. The framework is an open-source project with a simple command-line interface.

INTRODUCTION

Cancer is a genetic disease, dominated by somatic genetic mutations altering key cellular processes such as DNA repair and cell cycle (1). Most arising somatic mutations are considered passenger mutations, whereas only a small fraction of them have a direct role in oncogenesis, and are thus referred to as cancer driver mutations (2–4).

In recent years, cancer genomic research has benefited from increasing quantities (and quality) of molecular data. The Cancer Genome Atlas (TCGA) is a valuable resource of genomic data from cancer patients covering >10 000 samples in over 30 cancer types (5). An ongoing effort in cancer research is compiling a comprehensive catalogue of cancer genes which have a role in tumorigenesis. Knowledge of these genes is crucial for diagnosis and treatment of the disease (6,7).

Numerous computational frameworks have been designed for the purpose of identifying suspect cancer genes (8–12). Most of these frameworks, regarded as ‘frequentist’, are based on the premise that cancer genes are recurrent across samples and can be recognized by excessive numbers of somatic mutations. In contrast, passenger mutations are expected to appear at random. Assessing whether a gene shows an excessive number of mutations must be considered in view of an accurate null background model. Since cancer is characterized by order-of-magnitudes variability in mutation rates among cancer types, samples and genomic loci (9,13), the frequentist approach requires complex modeling of gene mutation rates as a function of the composition of samples and cancer types that produced the mutations. It must also incorporate variations in mutation rates based on genomic regions or chromatin structures under study (14–16). Modeling all these variables introduces numerous assumptions about the observed somatic mutations, which, if violated, may result in false discoveries (9,17,18). The sensitivity of the frequentist approach to modeling choices leads to lingering uncertainty and controversy (8).

An alternative to the frequentist approach, which can be regarded as ‘functionalist’, considers the content of mutations rather than their numbers. It is based on the premise that somatic mutations in cancer genes, regardless of their number, are subjected to positive selection and, as a result, are more damaging than expected at random. Under the functionalist approach, each gene has its own inherent background model which only depends on static properties of the gene and the number of mutations. It then determines whether the observed mutations appear more damaging than the same number of random mutations. Other vari-

[*]To whom correspondence should be addressed. Tel: +972 25494608; Fax: +972 25494500; Email: nadav.brandes@mail.huji.ac.il
Correspondence may also be addressed to Michal Linial. Email: michall@cc.huji.ac.il

ables, such as the samples or cancer types that the mutations have originated from, or the specific genomic region of the gene under study, do not need to be part of the model. As a result, the functionalist approach can make fewer assumptions about the background distribution of random mutations.

Example of a simple functionalist model is the non-synonymous to synonymous (dN/dS) ratio (19,20) which is a common metric for the evolutionary selection of a gene. A richer functionalist model was recently explored by OncodriveFML (21). OncodriveFML estimates the pathogenicity of mutations using CADD (22), which provides numeric scores for the clinical effects of mutations. OncodriveFML then compares the CADD effect scores of the somatic mutations observed within a gene to those of random mutations using permutation tests. Despite being a functionalist framework, OncodriveFML still uses a rather complex background model that includes sample identities and cancer types. As a result of its complex background model, it is unable to calculate probabilities analytically, and requires computationally demanding permutation tests.

With the goal of developing an analytical functionalist model, we introduce a new framework called FABRIC (Functional Alteration Bias Recovery In Coding-regions). FABRIC is a purely functionalist framework, with a simple background model that is completely agnostic to samples, cancer-types and genomic regions. This simplicity allows analytical calculation of precise *P*-values per gene. As a result, FABRIC can provide a detailed ranking of all genes by significance.

FABRIC is comprised of three components: (i) a machine-learning prediction model used to assign quantitative effect scores to mutations in coding regions, based on their rich proteomic context, (ii) a simple background model per gene, which doesn't require any covariates in the input data, and (iii) precise calculation of the probability for the extent of the damage caused by the observed mutations compared to the background model.

We illustrate the performance of FABRIC by comparing its results to commonly used catalogues of cancer driver genes. We further compare between OncodriveFML and FABRIC, used on the same input data (TCGA, >10 000 samples). We demonstrate the applicability of FABRIC in both pan-cancer and cancer-type specific analyses.

MATERIALS AND METHODS

Dataset of somatic mutations

The dataset of somatic mutations in cancer, used in the analyses throughout this work, was extracted from TCGA (5). We used the somatic mutations processed by the MuTect2 workflow for variant aggregation and masking (23), downloaded through NIH's GDC Data Portal (24). We selected only the 33 open access files, corresponding to the 33 open-access cancer type projects.

In total, these 33 projects contained 3 175 929 somatic mutations across 10 182 samples. 2 956 550 of these mutations were SNVs (i.e. substitutions of single nucleotides), and 2 235 884 of these SNVs were in coding regions (i.e. substituting a nucleotide within the open reading frame of a protein-coding gene). Each of these coding-region SNVs was assigned effect score(s) for the gene(s) it affected (occasionally it happens that the same mutation affects multiple overlapping genes).

Framework overview

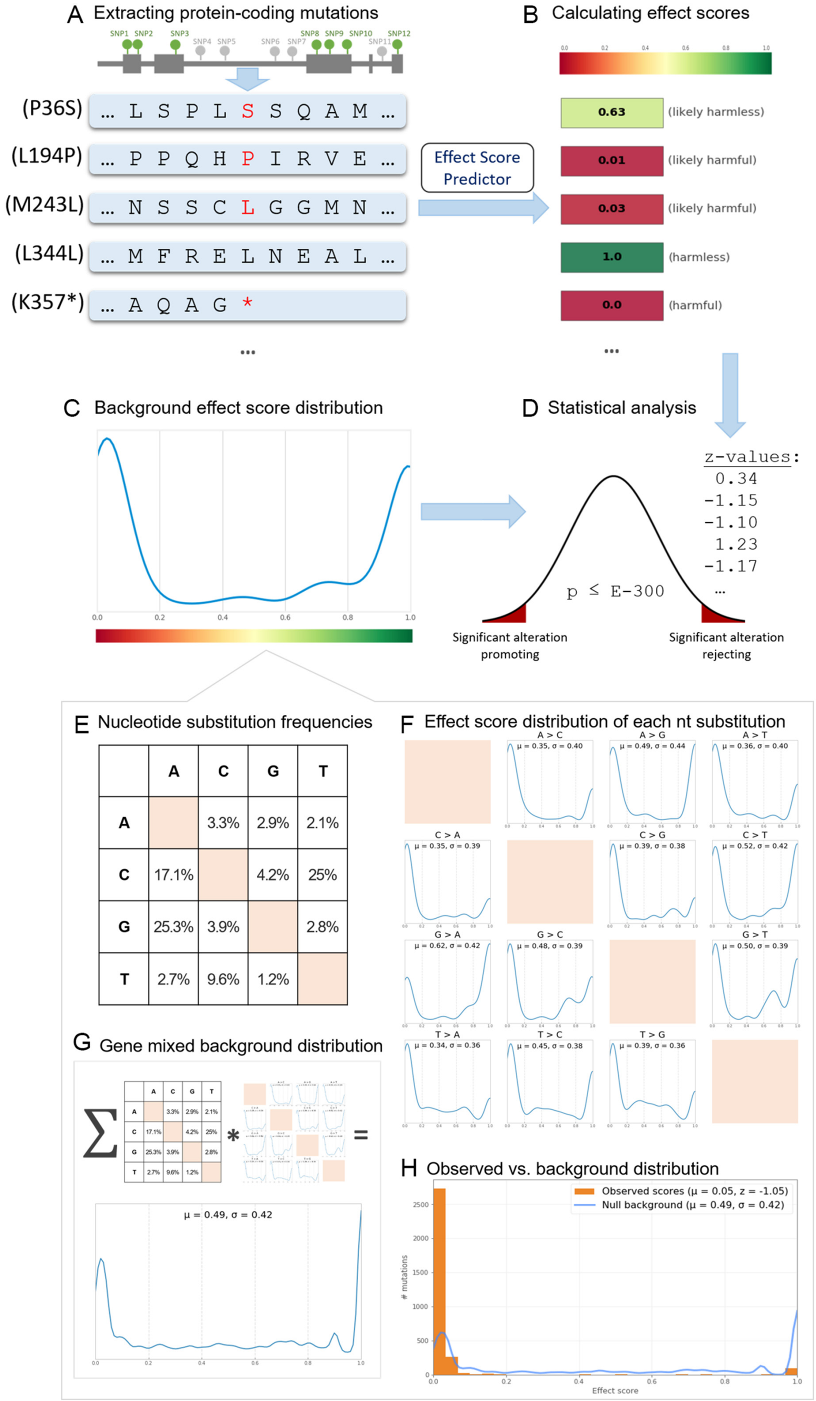
FABRIC analyzes each protein-coding gene independently, extracting all the single nucleotide variations (SNVs) observed within the coding regions of that gene (Figure 1A). It then uses a machine-learning model to assign functional effect scores to each SNV (Figure 1B), which measure the predicted effects of those variants explicitly on the protein function (see details below). Intuitively, this score can be thought of as the probability of the protein to retain its original biochemical function given the mutation. Simplistically, all synonymous mutations are assigned a score of 1 (gene retains full function), nonsense mutations are assigned a 0 score (gene retains no function), and missense mutations are processed through the machine-learning model to obtain a score between 0 to 1. The machine-learning model was trained in advance on an independent dataset.

Independently to the calculation of scores for the observed mutations, a background distribution for the expected scores is also constructed, assuming that unselected passenger mutations occur at random by a uniform distribution across the gene (Figure 1C). This background model is precise, and calculated individually for each gene. Significant deviations between the null background distribution to the observed effect scores are then detected (Figure 1D). *z*-values measure the strengths of deviations between observed to expected scores, and, using routine statistical tools, exact *P*-values are derived.

If a gene's average z-value is significantly negative, it means its observed scores are significantly lower than expected. This indicates that they are more damaging to the gene function than expected by the same number of mutations randomly distributed along the gene's coding sequence. In such case, the gene is deemed to be 'alteration promoting', reflecting its tendency to harbor damaging mutations, which are presumably beneficial to the development and evolution of the tumor. An observed score that is significantly higher than expected indicates the opposite, namely genes less damaging and more constrained than expected. We refer to these genes as 'alteration rejecting'.

We illustrate FABRIC's background model by a detailed specific example of *TP53* (Figure 1E–H). Importantly, the 12 background distributions of the *TP53* gene, corresponding to the 12 possible single-nucleotide substitutions (Figure 1F), are completely independent of the input data, and represent only the inherent properties of the gene. The only part of the background model actually dependent on the input is the 12 frequencies (Figure 1E). Hence, the background model accounts for the exact number of mutations and their single nucleotide substitution frequencies as observed in the data for the studied gene. This per-gene background model doesn't rely on any additional covariates. We avoided the more complex signature of 96 trinuleotide frequencies (25,26) as it would result in a too detailed background model compatible only with long proteins.

In order to keep the model simple and minimize the required assumptions, we restricted our framework to the



	A	C	G	T
A		3.3%	2.9%	2.1%
C	17.1%		4.2%	25%
G	25.3%	3.9%		2.8%
T	2.7%	9.6%	1.2%	

Figure 1. FABRIC framework. (A–D) Framework overview, (E–H) background model (*TP53* as an example). (**A**) All somatic mutations within a particular gene are collected from a variety of samples and cancer types. SNVs within protein-coding regions are analyzed to study their effects on the protein sequence (synonymous, missense or nonsense). (**B**) Using a machine-learning model, we assign each mutation a score for its effect on the protein biochemical function, with lower scores indicating mutations that are more likely harmful. (**C**) In parallel, a precise null background score distribution is constructed (details in E–H). (**D**) By comparing the observed scores to their expected distribution, we calculate *z*-values for the mutations, and overall *z*-value and *P*-value for the gene. (**E**) 3167 SNVs were observed in coding regions of *TP53* from which a 4 × 4 matrix of single-nucleotide substitution frequencies was derived. Note that this matrix is non-symmetric (e.g. 25.3% of the substitutions are G to A, while only 2.9% are A to G). (**F**) For each of the 12 possible nucleotide substitutions, an independent background effect score distribution was calculated, by considering all possible substitutions within the coding region of *TP53* and processing them with the same effect score prediction model used in (B). (**G**) By mixing the 12 distributions calculated in (F) with the weights of the substitution frequencies calculated in (E), we obtained the gene's final effect score distribution, used as its null background model for the analysis. (**H**) According to the null background distribution, we would expect mutations within the *TP53* gene to have a mean score of $\mu = 0.49$. However, the observed mean score of the 3167 analyzed mutations is $\mu = 0.05$, which is 1.05 standard deviations below the mean (*P*-value < E–300). The observed mean (0.05) was calculated from the 3167 SNVs observed in *TP53* which are categorized as follows: 92 synonymous mutations (effect scores of 1), 512 nonsense mutations (effect scores of 0) and 2563 missense mutations with an average score of 0.02.

analysis of SNVs, accounting for 93% of the somatic mutations in the analyzed dataset. Modeling non-SNV variations (e.g. indels, copy-number variations and chromosomal rearrangements) would require complex modeling, thereby jeopardizing the robustness and validity of the results. Likewise, we restricted FABRIC to protein-coding genes, and considered the functional effects of genetic variations only within the context of their proteins, allowing direct proteomic-based interpretation of the results. By ignoring complex variations and effects (e.g. frameshifts and splicing events), it is likely that in many instances we underestimate the damage to gene function.

Statistical framework & background model

Our framework uses a pre-trained prediction model for the effect scores of missense variants, denoted ϕ (the training of ϕ is discussed in the next section). For each variant v (in the context of a protein-coding gene) we assign a deterministic effect score $ES(v) \in [0, 1]$ by the following rule:

$$ES\,(v) = \begin{cases} 0 & v \textit{ is nonsense} \\ \phi\,(v) & v \textit{ is missense} \\ 1 & v \textit{ is synonymous} \end{cases}$$

In order to construct a background distribution for the effect scores expected at random (Figure 1), we first consider each single-nucleotide substitution individually. Let $nt_1, nt_2 \in \{A, C, G, T\},\ nt_1 \neq nt_2$ be two different nucleotides. The background model for the substitution $nt_1 \rightarrow nt_2$ in gene i is determined by calculating $ES(v)$ for all possible substitutions $nt_1 \rightarrow nt_2$ within the open-reading frame sequence of the gene. Specifically, let $l_1, \ldots, l_k$ be all the occurrences of nt_1 within the open-reading frame sequence of the gene. For each $j \in \{1, .., k\}$, let us denote by $\hat{v}_j$ the variant that results in upon substituting the occurrence l_j of nucleotide nt_1 by nucleotide nt_2 within the context of gene i. The background distribution for the substitution $nt_1 \rightarrow nt_2$ in gene i, denoted by D_{i,nt_1,nt_2}, is a uniform distribution over $ES(\hat{v}_1), \ldots, ES(\hat{v}_k)$ (each chosen with probability $\frac{1}{k}$).

In order to construct the background distribution D_i for the entire gene i, we first calculate the frequencies of the nucleotide substitutions of the observed variants within the gene, denoted f_{nt_1,nt_2} for the observed frequency of the $nt_1 \rightarrow nt_2$ substitution. These frequencies satisfy: $\sum_{nt_1,nt_2 \in \{A,C,G,T\},\ nt_1 \neq nt_2} f_{nt_1,nt_2} = 1$. We then take D_i to be a mixture of the twelve D_{i,nt_1,nt_2} distributions with f_{nt_1,nt_2} as coefficients (i.e. to sample from D_i one first samples a pair of nucleotides nt_1, nt_2 with probabilities f_{nt_1,nt_2} and then samples from D_{i,nt_1,nt_2}).

Let $v_1, \ldots, v_n$ be the observed variants in gene i. We calculate the mean observed score of the gene $\mu_i = \frac{ES(v_1)+\ldots+ES(v_n)}{n}$ and compare it to the background model of the gene, D_i. We do this by calculating the gene's mean z-value $z_i = \frac{\mu_i - \hat{\mu}_i}{\hat{\sigma}_i}$, where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the mean and standard-deviation of D_i. This is equivalent to calculating the z-value for each variant individually (given by $\frac{ES(v)-\hat{\mu}_i}{\hat{\sigma}_i}$) and then averaging them. This value summarizes the overall strength of alteration bias in the variants observed for gene i, but it gives no indication of statistical significance. When $z_i < 0$, gene i is potentially alteration promoting, as the observed effect scores are lower than those expected at random, indicating more harmful variants. Similarly, $z_i > 0$ indicates a potential alteration rejecting gene.

When $z_i < 0$, we can derive the one-tailed P-value by calculating:

$$p_i = P_{\hat{s}_1,\ldots,\hat{s}_n \sim \tilde{D}_i}\ (\hat{s}_1 + \ldots + \hat{s}_n \leq ES(v_1) + \ldots + ES(v_n))$$

In other words, the P-value is the probability of obtaining scores at least as low as the observed ones, assuming they are independent and identically distributed (i.i.d.) according to the background distribution D_i. Similarly, when $z_i > 0$ we calculate the probability of obtaining scores at least as high as the observed ones. All the reported P-values throughout this work are two-tailed, obtained by multiplying the one-tailed P-values by a factor of 2.

In order to compute the P-values, we need to calculate the distribution of the sum of n i.i.d random variables, each with distribution D_i. The distribution of the sum is given by convolving D_i with itself n times. To facilitate this computation, we round all the values (both the observed values, and in the background model) to two decimal digits, obtaining 101 distinct bins in the range [0, 1]: 0, 0.01, 0.02, …, 0.99, 1. The distribution of the i.i.d sum is then given by $100n + 1$ bins in the range $[0, n]$. This computation results in a precise probabilistic calculation given that the missense effect score predictor ϕ outputs scores in a resolution of 2 decimal places.

As evident from this mathematical formulation, FABRIC makes only a single assumption: mutations under no selective pressure distribute uniformly across a gene's sequence (corrected for the observed intrinsic biases from single-nucleotide substitution tendencies). If this one assumption is accepted then the statistical results, which are precise probabilistic calculations, are indisputable. In particular, FABRIC makes no assumptions about the validity of the pre-trained prediction model ϕ, or the effect-score calculation schema ES in general. Formally, even if the scoring function gives arbitrary scores, the calculated P-values are still accurate, and significantly low P-values provide strong evidence against the null hypothesis, namely that the observed variants do not seem to distribute independently and uniformly across the gene. A bad scoring function would undoubtedly diminish the statistical power of the framework, but should not result in false discoveries. For the same reason, false discoveries should not result in from including hyper-mutated genomic regions, samples or cancer types. As the background model controls for the prediction model, the number of observed variants and their nucleotide frequencies, the assumptions of our framework are minimal.

Effect score prediction model

A key component of FABRIC is a pre-trained machine-learning model for predicting the effects of missense genetic variants on protein function. Given the details of a missense variant, it predicts a numerical effect score between 0 (harmful) to 1 (harmless). There are numerous existing tools assessing the pathogenicity of genetic variations (e.g. CADD (22), SIFT (27), Polyphen2 (28), MutationTaster2 (29); for a collection of prediction tools, see (11)). However, FABRIC's goal is to find positive selection at the gene level (i.e. alteration promoting genes). It requires a predictor capable of assessing functional biochemical effects rather than clinical pathogenicity scores (discussed in (30,31)). Since the outputs of most existing predictors provide pathogenicity scores coupled with clinical consequences (32), we developed a new tool—FIRM (Functional Impact Rating at the Molecular-level), a dedicated predictor focused solely on assessing functional proteomic effects. FIRM is the machine-learning component incorporated into FABRIC.

In order to ensure that FIRM does not capture any clinical or evolutionary information, we restricted its used features to purely biochemical properties. For examples, while most functional effect prediction tools use multiple sequence alignment and evolutionary conservation of the gene/protein sequence as a primary feature, we avoided it altogether. FIRM extracts an immense set of features (1109 in total), aimed at capturing the rich proteomic context of each missense variant. The main features included are: (i) the location of the variant within the protein sequence, (ii) the identities of the reference and alternative amino-acids, (iii) the score of the amino-acid substitution under various BLOSUM matrices, (iv) an abundance of annotations extracted from UniProt, (v) amino-acid scales (i.e. various numeric values assigned to amino-acids, as described elsewhere (33,34)), (vi) Pfam domains and Pfam clans. For more details about the extracted features, see the Supplementary Methods.

Importantly, FIRM was pre-trained on a dataset independent to TCGA used in our primary analysis of FABRIC. Specifically, it was trained on a dataset of human genetic variations extracted from ClinVar (35). ClinVar provides a comprehensive catalogue of human genetic variations together with their clinical significance (e.g. pathogenic, benign), as determined by various submitting groups (e.g. OMIM (36)). It is important to note that while ClinVar variants are labeled by pathogenicity, FIRM is capable to extract only biochemical signal, due to its restricted set of features. We extracted a final dataset of 37 008 variants from ClinVar, 22 496 labeled harmful and 14 512 labeled harmless (see Supplementary Methods).

We used 3-fold cross-validation to estimate FIRM's performance. We chose a Random Forest classifier (implemented by the scikit-learn Python library (37) with the following hyper-parameters: n_estimators = 100 and min_samples_split = 50. We report the following performance on ClinVar's validation sets (average scores of the three cross-validation folds): AUC = 90%, precision = 86%, recall = 85.5%, specificity = 78.4%, F1 = 85.8% and accuracy = 82.7%. The overall good performance of FIRM reassures that it learned to extract meaningful signal and assess gene damage, despite the imperfections in the labeling on the ClinVar dataset (38).

It is important to stress that our goal in developing FIRM was not to improve the performance of state-of-the-art pathogenicity prediction (32). Rather, our purpose was to develop a model for predicting functional effects that does not use any evolutionary selection information, thereby allowing a separation between the goals of FIRM (measuring functional alteration) and FABRIC (quantifying evolutionary selection). As stated above, FABRIC isn't sensitive, in terms of false discoveries, to inaccuracies in FIRM, due to its probabilistic model which accounts for the predictions of FIRM as part of the background model. Both the observed somatic mutations and the background mutations are calculated identically by FIRM.

When a machine-learning classifier is used, usually only the predicted label (e.g. harmful or harmless variant) is of interest, while the exact score given by the prediction model has no significance. Furthermore, the exact scores (usually in the range 0–1) produced by algorithms like Random Forests have no simple interpretable meaning. FABRIC required refined effect scores spanning the entire 0–1 range, preferably with meaningful probabilistic interpretation. To this end, we rescaled the outputs produced by FIRM such that an effect score of $s \in [0, 1]$ would indicate that roughly s percentage of the validation-set variants with a similar score were benign (e.g. ∼85% of ClinVar's variants with an effect score of 0.85 were benign). This way, it can be useful to think of a variant with an effect score of 0.85 as having 85% chance of being harmless, although this is by no means guaranteed as we move from ClinVar to another dataset (e.g. to TCGA), especially considering that ClinVar is highly imbalanced and biased towards having mostly pathogenic variants.

RESULTS AND DISCUSSION

A pan-cancer catalogue of alteration promoting genes

We applied FABRIC on 2 235 884 SNVs in the coding regions of 17 828 genes containing at least one mutation (see Materials and Methods, and 'Constructing gene sequences & annotations' in the Supplementary Methods). Of these genes, the somatic mutations in 593 genes were significantly more harmful than expected at random (FDR q-value < 0.05; full ranked list of all analyzed genes is provided in Supplementary Table S1-TCGA_combined). A short excerpt with the top 15 results is given in Table 1.

Notably, significant alteration promoting genes can dramatically vary in their total number and density of mutations (i.e. number of mutations per nucleotide). For example, *TP53* has 2.69 SNV mutations per coding-region nucleotide, while *KMT2D* has a 38-fold lower mutation density (0.07). Even though *TP53* is the most significant alteration promoting gene (with respect to the calculated q-value), the effect score z-values of *APC* (–1.31) and *ARID1A* (–1.47) are lower than that of *TP53* (–1.05), indicating a potentially stronger effect size.

Notably, FABRIC is completely symmetric, detecting genes either more or less damaged than expected. Despite the methodological symmetry, we found only six significant alteration rejecting genes, compared to the 593 alteration promoting genes, confirming the dominance of positive selection over negative selection in cancer (Supplementary Table S1-TCGA_combined).

Evaluation of the pan-cancer results

We compared our results against prominent resources of cancer genes, used as a benchmark: the COSMIC-Census catalogue (39), and the recently compiled PanSofware catalogue of 299 cancer driver genes (11). A substantial and sig-

Table 1. Top 15 alteration promoting genes

Gene symbol	Gene name	Chr	# Observed mutations	Mutations per nt	Observed/expected mean score	Score *z*-value	FDR *q*-value	Census annotations[a]
TP53	tumor protein p53	17	3167	2.69	0.05 / 0.49	−1.05	0	OG, TSG, F
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	3	1508	0.47	0.15 / 0.39	−0.68	4E-240	OG
APC	APC, WNT signaling pathway regulator	5	997	0.12	0.49 / 0.84	−1.31	7.3E-215	TSG
KRAS	KRAS proto-oncogene, GTPase	12	783	1.38	0.04 / 0.24	−0.61	1.5E-177	OG
ARID1A	AT-rich interaction domain 1A	1	640	0.09	0.45 / 0.82	−1.47	3.6E-169	TSG, F
BRAF	B-Raf proto-oncogene, serine/threonine kinase	7	819	0.36	0.10 / 0.41	−0.77	3.5E-145	OG, F
PTEN	phosphatase and tensin homolog	10	656	0.54	0.06 / 0.33	−0.70	9.5E-116	TSG
IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic	2	536	0.43	0.05 / 0.33	−0.68	4.8E-94	OG
CDKN2A	cyclin dependent kinase inhibitor 2A	9	294	0.63	0.19 / 0.55	−0.99	9.9E-82	TSG
FBXW7	F-box and WD repeat domain containing 7	4	442	0.21	0.20 / 0.52	−0.86	8.8E-81	TSG
KMT2D	lysine methyltransferase 2D	12	1191	0.07	0.72 / 0.88	−0.65	5.9E-72	OG, TSG
NF1	neurofibromin 1	17	697	0.08	0.52 / 0.72	−0.67	6.9E-56	TSG, F
NRAS	neuroblastoma RAS viral oncogene homolog	1	286	0.50	0.06 / 0.34	−0.72	3.3E-52	OG
RB1	RB transcriptional corepressor 1	13	335	0.12	0.28 / 0.54	−0.82	1.2E-50	TSG
CTNNB1	catenin beta 1	3	442	0.19	0.30 / 0.53	−0.68	6.7E-49	OG, F

[a] OG, oncogene; TSG, tumor suppressor gene; F, fusion.

nificant overlap was found between the 593 detected genes to these two external lists of cancer genes (Figure 2A). A particularly remarkable enrichment is observed with respect to the PanSoftware catalogue. Of the 299 genes reported by the PanSoftware catalogue, 282 mapped into the list of 17 828 analyzed proteins-coding genes. Of these 282 genes, 147 (52%) were independently recovered by FABRIC over the TCGA dataset (×15.7 enrichment, *P*-value = 2.2E–144).

Our analysis gathers information from two distinct signals: (i) the composition of mutation types (synonymous, missense or nonsense) and, (ii) the predicted effect scores of missense mutations. As only the scores of missense mutations vary, these two components are fully orthogonal. To determine the contribution of each of the two complementary components, we considered, in addition to the full analysis (referred to as 'overall analysis' in Figure 2A), two other variations: (a) the mutation-type analysis (abbreviated Mut-Type) considers only deviations in the types of mutations (i.e. synonymous, missense or nonsense), treating all missense mutations as one category; (b) the missense analysis considers only missense mutations, looking for significant differences between their observed to expected effect scores, while disregarding the other two mutation types (i.e. synonymous and nonsense mutations). While our overall analysis found 593 significant alteration promoting genes, the mutation-type and missense analyses found 387 and 492 genes, respectively (see Supplementary Methods and Supplementary Table S1-TCGA_combined). As expected, we found a significant overlap between the mutation-type and missense analyses (*P*-value = 7.96E–30), confirming that both capture the same signal of positive selection in cancer, despite their reliance on independent properties of the data. These two components (Mut-Type and Missense, Figure 2A) are also capable of recovering many of the annotated cancer genes, yet the integrated overall analysis (overall, Figure 2A) shows superior results. This proves that the utilized machine-learning model, used for missense mutations, has an important role in our framework. In particular, FABRIC is superior to methods that only look for differences in mutation types, such as non-synonymous to synonymous (dN/dS) ratios (20) (which is reflected as the Mut-Type analysis). An exhaustive overlapping analysis is available in Supplementary Table S2.

Importantly, 510 significant genes were found in the missense analysis: 492 (96.5%) had a negative effect score *z*-value, indicating alteration promotion, and only 18 (3.5%) had a positive *z*-value, indicating alteration rejection. Unlike the overall analysis, which also considered synonymous and nonsense mutations with predefined effect scores, the missense analysis relied solely on the scores learned by FIRM. The overwhelming imbalance in the directionality of effect sizes (96.5% to 3.5%) is another strong evidence that FIRM was successfully trained over the ClinVar dataset, and was able to extract meaningful signal in the TCGA dataset, which was then utilized by FABRIC.

The PanSofware catalogue is based on a consensus from several tools for detecting driver genes (11). Among these tools, OncodriveFML is a functionalist method that does not rely on mutation rates. We compared the performance of FABRIC to OncodriveFML, the most prominent functionalist method currently available, by independently executing the two frameworks on the same TCGA dataset to find significant protein-coding genes (see Supplementary Methods). We measured the percentage of Census and PanSoftware genes recovered by each of the two frameworks (Figure 2B). We find that despite the simplicity of FABRIC, it performs slightly better than OncodriveFML. In fact, FABRIC is evidently superior when it comes to the ranking of the most significant results (up to the gene ranked ∼200). We attribute this difference to the fact that FABRIC, in contrast to OncodriveFML, is capable of analyti-

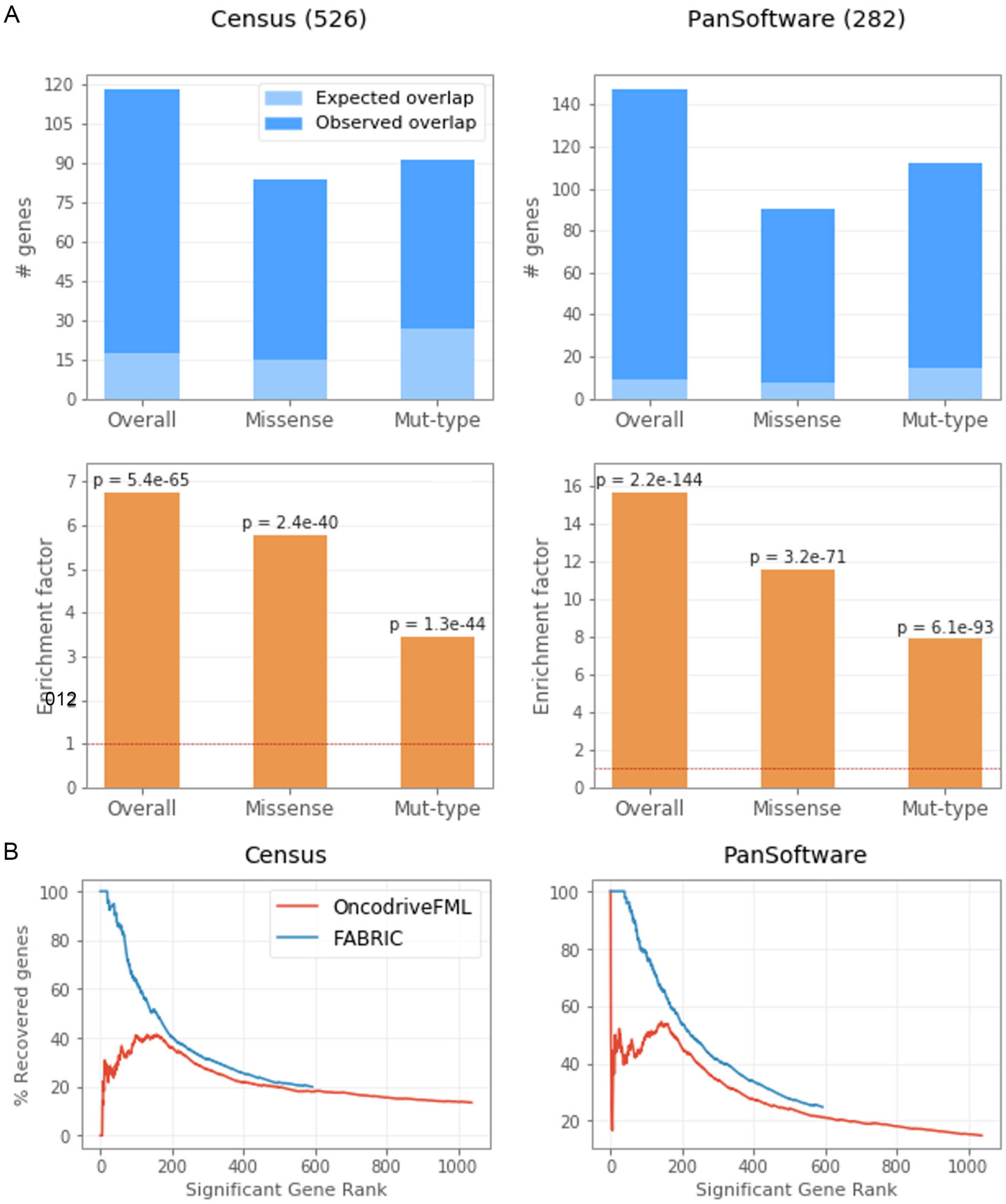


Figure 2. Overlap with known cancer genes. (**A**) We compared the lists of significant alteration promoting genes obtained by FABRIC against two resources of cancer genes: Census (526 genes) and the PanSoftware (282 genes) catalogues. The top panel (blue) shows the total number of overlapping genes between our analyses to each of the two compared resources. The numbers of genes that would be expected to overlap at random (given hyper-geometric distribution) are shown in light blue. In addition to the standard form of our framework (the left bar in each of the panels), we also explore two other variations (Missense and Mut-Type; see text). The ratio between the observed to the expected number of shared genes is defined as the enrichment factor for each pair, and is shown on the lower panel (orange). (**B**) Comparison between FABRIC and OncodriveFML by measuring the percentage of genes recognized by the external benchmark catalogs (Census and PanSoftware) from those discovered by each of the two methods, as a function of gene ranking (according to reported significance). For example, the left figure shows that 64% of the top 100 most significant genes in FABRIC are recognized by Census, compared to only 41% in OncodriveFML.

Table 2. Alteration promoting genes across cancer types

TCGA project	Disease	# Samples	# Mutations	Avg. (std) mutations per sample	# Significant Alteration Promoting Genes	# Significant Diff Genes[a]
BRCA	Breast Invasive Carcinoma	986	120 988	122.7 (347.8)	12	6
LUAD	Lung Adenocarcinoma	567	208 180	367.2 (386.3)	14	5
UCEC	Uterine Corpus Endometrial Carcinoma	530	886 377	1 672.4 (4159.7)	146	24
HNSC	Head and Neck Squamous Cell Carcinoma	508	102 309	201.4 (288.1)	15	7
LGG	Brain Lower Grade Glioma	508	35 556	70.0 (635.7)	9	3
PRAD	Prostate Adenocarcinoma	495	29 286	59.2 (408.3)	3	0
LUSC	Lung Squamous Cell Carcinoma	492	181 116	368.1 (317.6)	12	2
THCA	Thyroid Carcinoma	492	10 899	22.2 (52.3)	4	1
SKCM	Skin Cutaneous Melanoma	467	392 571	840.6 (1423.4)	16	11
STAD	Stomach Adenocarcinoma	437	213 144	487.7 (929.3)	10	1
OV	Ovarian Serous Cystadenocarcinoma	436	75 168	172.4 (178.7)	3	0
BLCA	Bladder Urothelial Carcinoma	412	134 513	326.5 (378.2)	36	15
COAD	Colon Adenocarcinoma	399	264 786	663.6 (1360.4)	19	5
GBM	Glioblastoma Multiforme	393	82 765	210.6 (964.8)	9	1
LIHC	Liver Hepatocellular Carcinoma	364	54 238	149.0 (161.5)	5	0
KIRC	Kidney Renal Clear Cell Carcinoma	336	26 693	79.4 (123.3)	5	3
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	289	103 405	357.8 (1 157.9)	12	3
KIRP	Kidney Renal Papillary Cell Carcinoma	281	23 765	84.6 (40.1)	1	0
SARC	Sarcoma	237	28 159	118.8 (281.7)	2	0
ESCA	Esophageal Carcinoma	184	45 313	246.3 (317.1)	4	0
PCPG	Pheochromocytoma and Paraganglioma	179	2411	13.5 (7.4)	2	0
PAAD	Pancreatic Adenocarcinoma	178	29 959	168.3 (1 534.7)	5	0
TGCT	Testicular Germ Cell Tumors	144	3198	22.2 (12.1)	2	0
LAML	Acute Myeloid Leukemia	143	9905	69.3 (271.7)	6	0
READ	Rectum Adenocarcinoma	137	64 804	473.0 (1783.4)	10	2
THYM	Thymoma	123	4737	38.5 (120.6)	2	1
ACC	Adrenocortical Carcinoma	92	10 747	116.8 (316.1)	0	0
MESO	Mesothelioma	82	3827	46.7 (43.8)	3	0
UVM	Uveal Melanoma	80	1856	23.2 (58.6)	3	2
KICH	Kidney Chromophobe	66	2896	43.9 (116.5)	1	0
UCS	Uterine Carcinosarcoma	57	10 449	183.3 (681.7)	6	0
CHOL	Cholangiocarcinoma	51	5503	107.9 (220.0)	2	0
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	37	6406	173.1 (106.1)	1	0

[a] Diff Genes, genes with significantly different alteration bias compared to other cancer types (Figure 4C).

cally deriving exact *P*-values, as it is not limited by permutation tests. OncodriveFML, on the other hand, gave the exact same *P*-value (E–06) to the top 160 genes, meaning that the ranking among the most significant cancer genes is arbitrary in the OncodriveFML platform. It should be noted that the rate of agreement between OncodriveFML and the PanSoftware catalogue (Figure 2B) is likely inflated, as OncodriveFML is one of the eight softwares used to derive that catalogue (11).

Alteration bias across cancer types

The primary analysis in this work is presented from a pan-cancer perspective. Namely, all the somatic mutations extracted from TCGA were combined into a single pool (per gene), disregarding from which samples or cancer types they originated. An important benefit of this pan-cancer setting was the acquiring of the needed statistical power for the analysis, obtained by maximizing the number of samples.

However, a notable heterogeneity exists among cancer types in the dominance of cancer genes (40). To highlight such differences, we conducted similar analyses, separately within each cancer type. By merely changing its input data, FABRIC automatically recalculated the specific background model for each combination of gene and cancer type, based on the observed mutations in each combination. As a result of each cancer type having its own unique background model, based only on observed mutations within that cancer, FABRIC remained insensitive to differences that exist between cancer types, such as cancer-specific nucleotide substitution frequencies (26). We analyzed 33 cancer types, ranging from ∼40 to ∼1000 samples and ∼2000 to ∼900 000 somatic mutations in each (Table 2). In total, we found 380 cancer-type specific alteration promoting genes, involving 231 unique genes. The summary statistics of each analyzed gene in each cancer type is available in Supplementary Table S1.

The results of all the analyses with at least 30 000 observations are also shown as quantile-quantile (QQ) plots in Figure 3. Evidently, the total number of mutations is a crucial factor in the obtaining of significant results across cancer-type projects. For reference, a similar QQ plot for OncodriveFML (showing a similar pattern) is available at Supplementary Figure S2.

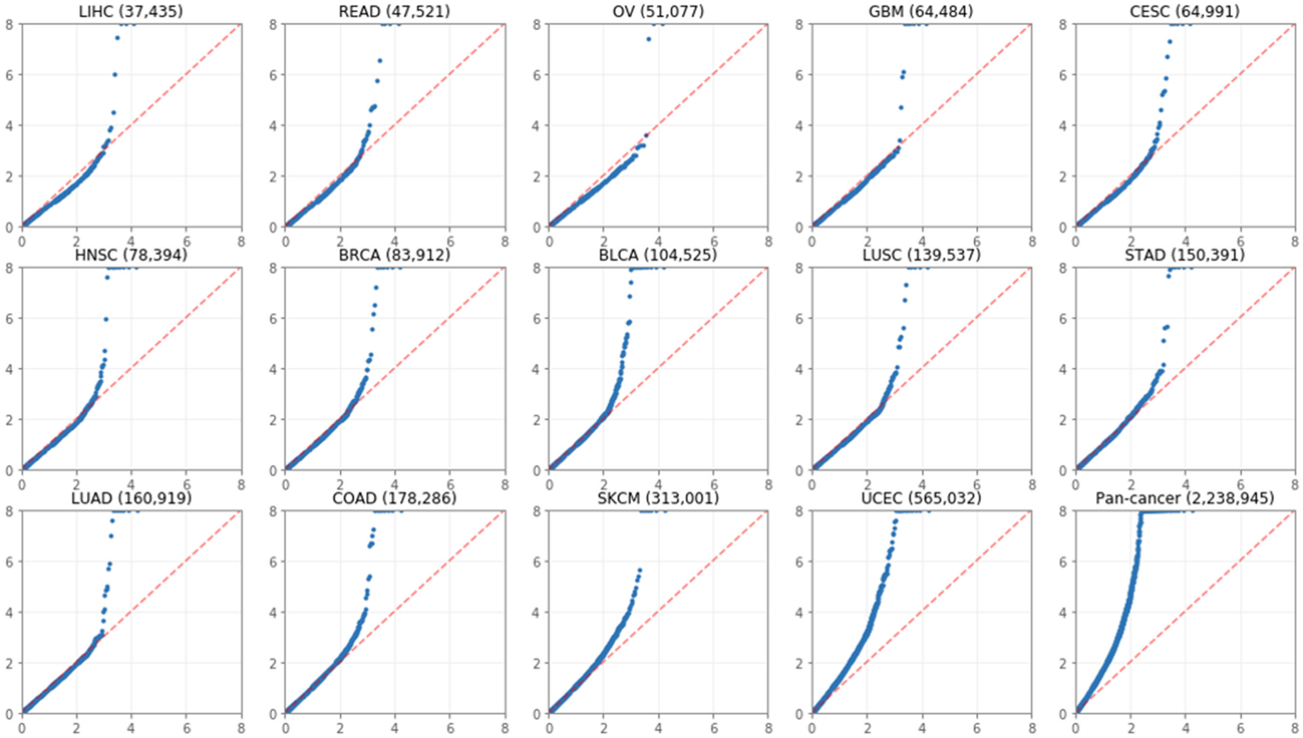


Figure 3. Quantile–quantile (QQ) plots. Quantile–quantile (QQ) plots comparing the significance of FABRIC's results (y-axis) to uniform distribution (x-axis) across 14 different cancer types and the pan-cancer analysis (bottom-right corner). The total number of observations (effect scores of somatic mutations) in each analysis is shown in parentheses. For visibility, very significant genes are truncated and shown as $P = \text{E–08}$.

The mild superiority of FABRIC over OncodriveFML, which has been demonstrated over the pan-cancer TCGA dataset (Figure 2B), is also observed across most cancer types. For example, a similar trend is demonstrated in Uterine Corpus Endometrial Carcinoma (UCEC; Figure 4A), the TCGA project with the highest number of observed mutations. Similar comparisons across other cancer types are shown in Supplementary Figure S1.

To further highlight cancer-type patterns, the magnitude of alteration bias across cancer types is shown for selected genes (Figure 4B). We also present genes with significant differences among cancer types (Figure 4C; see Supplementary Methods and Supplementary Table S1-TCGA_diff), marking genes within specific cancer types that show a significant alteration bias compared to the same genes in all other cancer types. Note that alteration bias compared to the background model (Figure 4B) is not the same as alteration bias compared to other cancer types (Figure 4C). For example, *TP53* is ranked at the top of the pan-cancer list of alteration promoting genes (Figure 4B) but shows only a weak difference in alteration bias across cancer types (not among the top genes in Figure 4C), confirming its universal role across many cancer types.

ARID1A, a well-studied cancer driver that belongs to the growing set of cancer drivers found to play a role in chromatin remodeling (41), is a highly significant alteration promoting gene across many cancer types (Figure 4B). However, in Skin Cutaneous Melanoma (SKCM) it is significantly less damaged (Figure 4C), suggesting that its role in oncogenesis within this cancer type is not as important compared to other cancer types. *FAT1* (FAT atypical cadherin 1) and CIC (Capicua transcriptional repressor), both well-known cancer drivers, seem to be particularly dominant in the Head and Neck Squamous Cell Carcinoma (HSNC) and the Brain Lower Grade Glioma (LGG) cancer types, respectively. *FAT1* encodes a cadherin-like protein that binds β-catenin, antagonizing its nuclear localization. Damaging mutations to *FAT1* that suppress its binding capacity lead to activation of the Wnt signaling, which is fundamental in tumorigenesis (42). *APC*, another tumor suppressor that binds β-catenin (43), seems especially dominant in the Colon Adenocarcinoma (COAD) and the Rectum Adenocarcinoma (READ) cancer types, two cancer types sharing high degree of molecular communality. Additional genes that are especially dominant in specific cancer types are: *SETD2* in Kidney Renal Clear Cell Carcinoma (KIRC), *KMT2C* in Breast Invasive Carcinoma (BRCA) and Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), and *GTF2I* in Thymoma (THYM).

Alteration promoting genes unlisted in contemporary cancer gene catalogues

Of the 593 significant pan-cancer alteration promoting genes, we found an outstandingly large subset to overlap with known cancer genes (Figure 2A), yet 426 of the reported genes are not listed in either Census or the PanSoftware gene catalogues; we denote them 'unlisted genes'. The full collection of all 426 unlisted genes is available in Supplementary Table S1-TCGA_combined_unlisted; the top 10 are shown in Table 3.

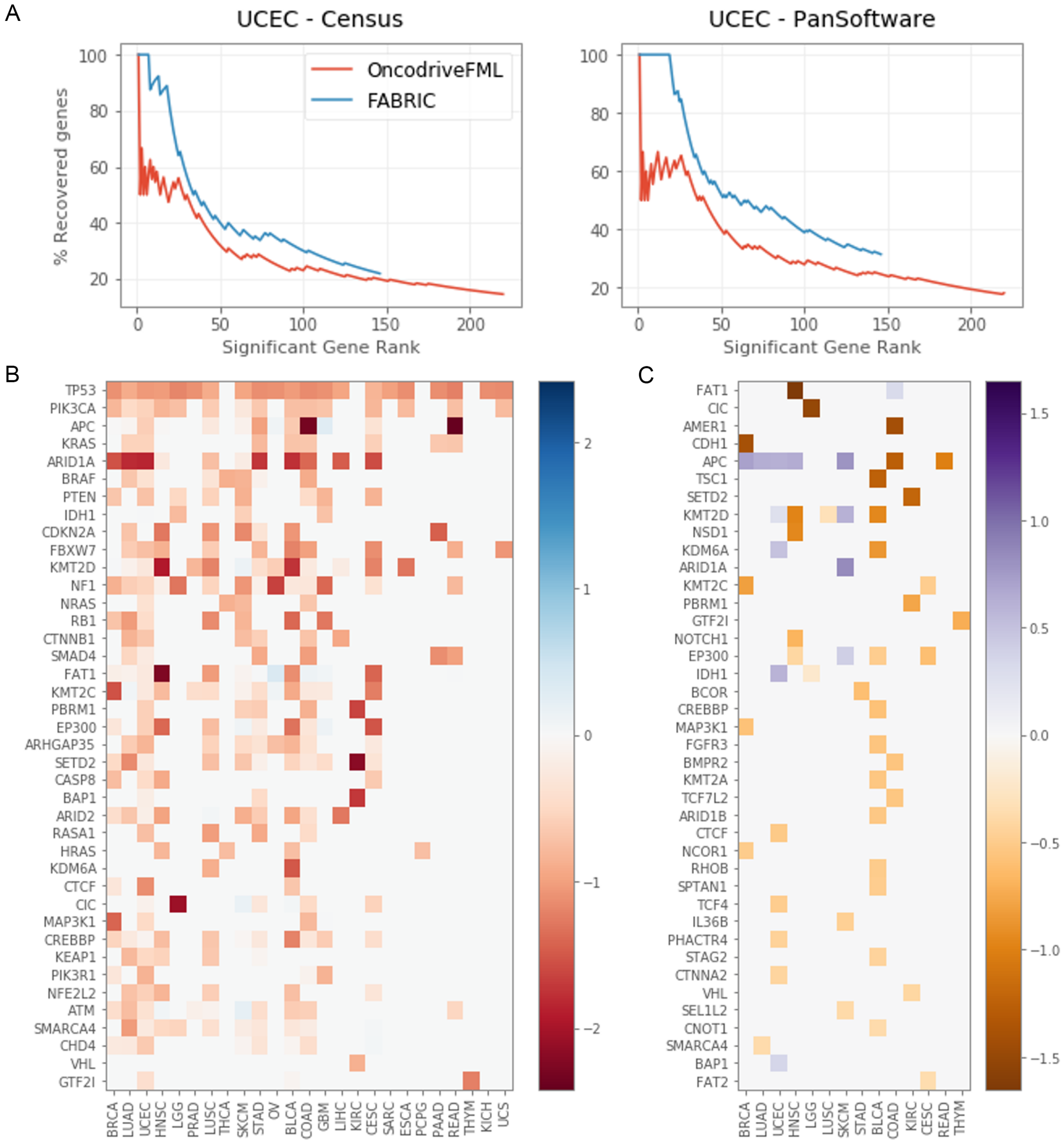


Figure 4. Alteration bias across cancer types. (**A**) Comparison between OncodriveFML and FABRIC over Uterine Corpus Endometrial Carcinoma (UCEC). Comparisons over other cancer types are shown in Supplementary Figure S1. (**B**) Average z-values of mutation effect scores (compared to the expected backgrounds) across cancer types for the top 40 alteration promoting genes detected by FABRIC (sorted by significance). More negative values (red) indicate genes that are more biased towards harmful mutations. Entries with less than 15 observed mutations were filtered out (gray color). (**C**) Top 40 genes (of 68) found to have significant differences in alteration bias across cancer types. Each value indicates the mean z-value difference between the relevant cancer type to all other cancer types. Negative values (orange) indicate genes that are more damaged in the relevant cancer types; positive values (purple) indicate genes that are less damaged. Non-significant values (after FDR) are not shown (gray color).

Table 3. Top 10 alteration promoting genes unlisted in cancer gene catalogues

Rank	Gene symbol	Gene name	Chr	# Observed mutations	Mutations per nt	Score *z*-value	FDR *q*-value	Literature evidence
41	ZC3H13	zinc finger CCCH-type containing 13	13	344	0.07	−0.45	4.1E-12	**Weak** [proliferation]
46	ZNRF3	zinc and ring finger 3	22	153	0.05	−0.65	1.5E-09	**Strong** [TSG, Wnt signaling]
54	GRM5	glutamate metabotropic receptor 5	11	441	0.12	−0.31	2.8E-08	**None**
61	USP28	ubiquitin specific peptidase 28	11	240	0.07	−0.43	1.5E-07	**Strong** [DNA damage response]
63	MICU3	mitochondrial calcium uptake family member 3	8	88	0.06	−0.64	3.9E-07	**None**
70	CNOT1	CCR4-NOT transcription complex subunit 1	16	467	0.07	−0.27	2.3E-06	**Weak** [OG, Hedgehog signaling]
72	ZNF14	zinc finger protein 14	19	160	0.08	−0.47	2.9E-06	**Weak** [Methylation]
73	MAP2K7	mitogen-activated protein kinase kinase 7	19	119	0.09	−0.51	3.5E-06	**Weak** [Motility]
75	LSM11	LSM11, U7 small nuclear RNA associated	5	52	0.05	−0.81	3.9E-06	**None**
76	ATAD2	ATPase family, AAA domain containing 2	8	319	0.08	−0.34	4.5E-06	**Strong** [OG, proliferation]

Of the 426 unlisted genes, 51 are very significant (FDR *q*-value < 1E–03). The most significant is *ZC3H13* (*q*-value = 4.1E–12, Table 3), which is ranked 41 in the list of 593 significant genes. In other words, all 40 highest ranking genes found by FABRIC are listed as cancer genes in either of the two external catalogues. Among the significant alteration promoting genes, those listed in Census and PanSoftware show similar statistical properties, in terms of significance (*q*-value) and effect size (*z*-value), to those unlisted in those catalogues (Figure 5A), suggesting that the unlisted genes could be genuine cancer genes.

To systematically examine whether the 426 unlisted genes are supported in literature, we consider two databases curating the literature evidence of cancer genes: the Candidate Cancer Gene Database, CCGD (44) and DisGeNET (45). CCGD is a manually curated resource for genes implicated in cancer by transposon mutagenesis in mice. DisGeNET is the largest gene-disease association dataset. We queried DisGeNET for neoplasm-associated genes. The 426 alteration promoting genes unlisted in Census and PanSoftware are supported by a significantly high number of studies according to CCGD, and have a significantly high neoplasm score according to DisGeNET (Figure 5B; see Supplementary Methods).

The alteration biases (*z*-values) of top unlisted genes are also shown across cancer types (Figure 5C). For example, ZNRF3 and MAP2K7 are dominant in Stomach adenocarcinoma (STAD), CNOT1 in Bladder urothelial carcinoma (BLCA), and ATAD2 in Colon adenocarcinoma (COAD).

Among the ten most significant unlisted genes, three (*GRM5*, *MICU3* and *LSM11*) show almost no record in literature for involvement in cancer (Table 3). Four genes (*ZC3H13*, *CNOT1*, *ZNF14* and *MAP2K7*) have weak support, mostly by in-vitro assays. These genes have been manipulated in cell-lines and demonstrated cancer related properties such as migration and cell division. The other three genes (*ZNRF3*, *USP28* and *ATAD2*) have strong evidence for having a role in cancer. *ZNRF3*, a cell-surface transmembrane E3 ubiquitin ligase, was implicated in regulating the Wnt pathway in colorectal neoplasia (46). *USP28*, a ubiquitin specific protease, acts as a tumor-promoting factor. Its high mRNA and protein levels correlate with low survival rate. It stabilizes cell cycle genes via *TP53* in response to DNA damage (47). *ATAD2* is an oncogene leading to enhanced cell proliferation and resistance to apoptosis. It is part of the Myc signaling pathway, and was implicated in cervical cancer and other aggressive tumors (48,49). The level of support demonstrated for these three unlisted genes, independently implicated by FABRIC, emphasizes the incompleteness of contemporary cancer gene catalogues.

We argue that the 426 unlisted genes are good candidate for further research. Their significance by FABRIC provides strong evidence that they undergo positive selection and play a role in tumor.

In summary, we developed FABRIC, a novel framework for the detection of genes undergoing selection in cancer. As a purely functionalist framework, it makes minimal assumptions about the data. Its utilized signal (the functional effects of mutations) is completely orthogonal to the signal exploited by traditional frequentist approaches (the number of mutations), allowing straightforward meta-analysis combining the two approaches. Through an unbiased systematic analysis of ∼3M somatic mutations from ∼10K cancer samples, we detected 593 alteration promoting genes. 426 of these genes are unlisted in the prominent cancer gene catalogues. We have presented initial evidence for their relevance to cancer, marking them attractive targets for further research and consideration in cancer catalogues. We provide the full analysis results as a comprehensive resource with the quantified selection of all human coding genes, ranked by statistical significance.

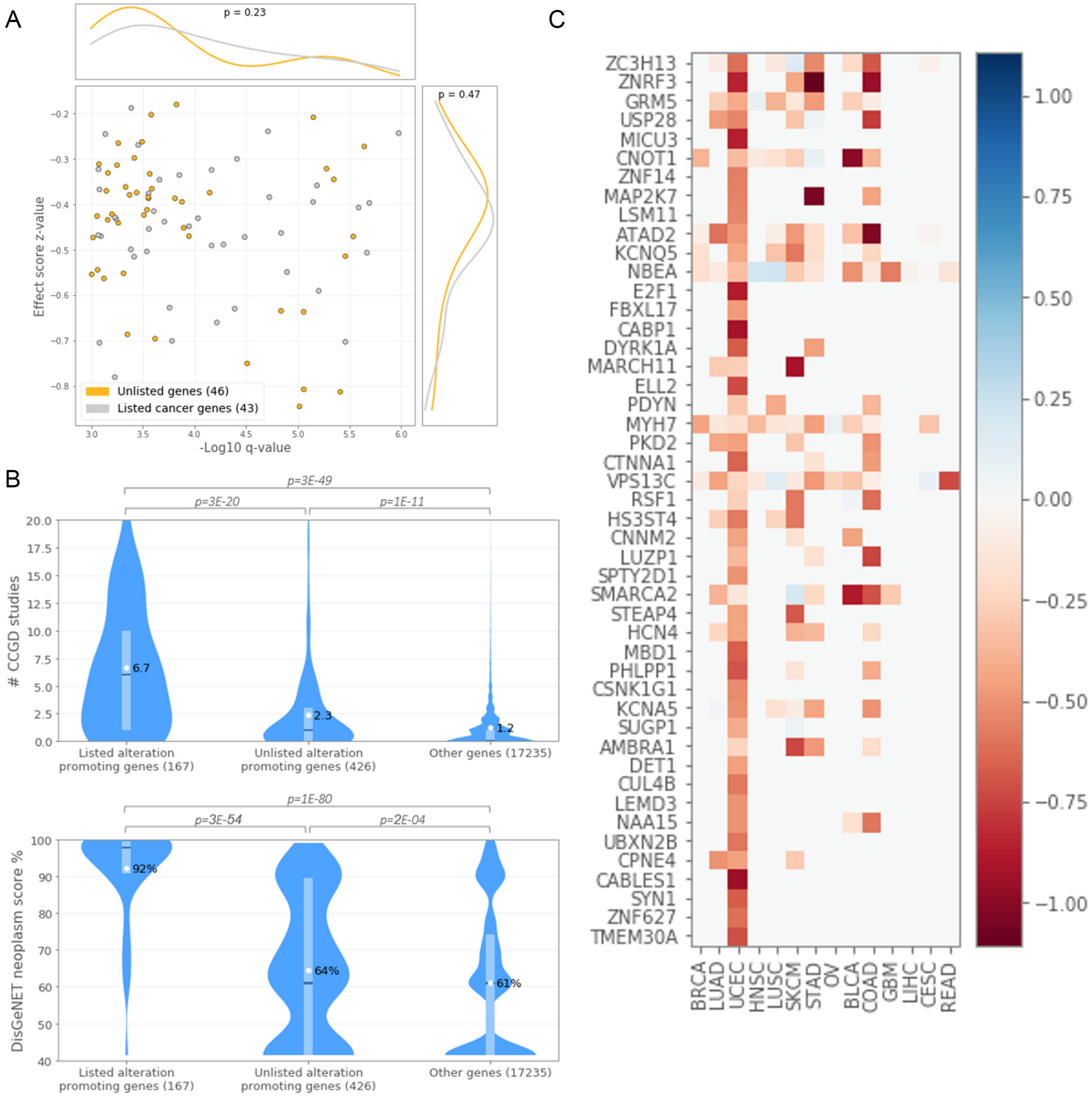


Figure 5. Evidence for 426 significant genes unlisted in cancer catalogues. (**A**) FABRIC summary statistics (*q*-value and *z*-value) of the alteration promoting genes in the significance range 1E–03 to 1E–06. Of the 89 presented genes, 43 are listed in Census or the PanSoftware cancer driver gene catalogues, while the other 46 are unlisted in those lists. The two gene groups have similar *q*-value and *z*-value distributions ($P = 0.23$ and $P = 0.47$, respectively). (**B**) Literature support, based on the CCGD and DisGeNET databases, for the three gene groups (see main text). Alteration promoting genes, according to FABRIC, are supported by more studies according to CCGD, even those unlisted in Census and PanSoftware (2.3 studies on average, compared to 1.2; $P =$ 1E-11). (**C**) *z*-values of highly significant (FDR *q*-value < 1E–03) unlisted genes across cancer types, after keeping only genes with at least 15 observations in at least one cancer type (e.g. most genes are presented for UCEC due to the high number of observations in this cancer type).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Or Zuk from the Hebrew University of Jerusalem for his helpful advice and insight on the statistical foundation of this work. We would also like to thank Roni Rasnic for her valuable review of the manuscript. We thank ELIXIR-EU for their resources used throughout this study.

FUNDING

European Research Council's grant on High Dimensional Combinatorics (N.B. fellowship) [339096]. Funding for open access charge: Lab's money.

Conflict of interest statement. None declared.

REFERENCES

1. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
2. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
3. Pleasance,E.D., Cheetham,R.K., Stephens,P.J., McBride,D.J., Humphray,S.J., Greenman,C.D., Varela,I., Lin,M.-L., Ordóñez,G.R. and Bignell,G.R. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.
4. Marx,V. (2014) Cancer genomes: discerning drivers from passengers. *Nat. Methods*, **11**, 375–379.
5. Tomczak,K., Czerwińska,P. and Wiznerowicz,M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, **19**, A68.
6. Porta-Pardo,E., Kamburov,A., Tamborero,D., Pons,T., Grases,D., Valencia,A., Lopez-Bigas,N., Getz,G. and Godzik,A. (2017) Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nature*, **201**, 7.
7. Forbes,S., Beare,D., Bindal,N., Bamford,S., Ward,S., Cole,C., Jia,M., Kok,C., Boutselakis,H. and De,T. (2016) COSMIC: high-resolution cancer genetics using the catalogue of somatic mutations in cancer. *Curr. Protoc. Hum. Genet.*, **91**, 10.11.1–10.11.37.
8. Tokheim,C.J., Papadopoulos,N., Kinzler,K.W., Vogelstein,B. and Karchin,R. (2016) Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 14330–14335.
9. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
10. Gonzalez-Perez,A., Deu-Pons,J. and Lopez-Bigas,N. (2012) Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.*, **4**, 89.
11. Bailey,M.H., Tokheim,C., Porta-Pardo,E., Sengupta,S., Bertrand,D., Weerasinghe,A., Colaprico,A., Wendl,M.C., Kim,J., Reardon,B. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **173**, 371–385.
12. Przytycki,P.F. and Singh,M. (2017) Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes. *Genome Med.*, **9**, 79.
13. Lawrence,M.S., Stojanov,P., Mermel,C.H., Robinson,J.T., Garraway,L.A., Golub,T.R., Meyerson,M., Gabriel,S.B., Lander,E.S. and Getz,G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
14. Liu,L., De,S. and Michor,F. (2013) DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.*, **4**, 1502.
15. Hodgkinson,A., Chen,Y. and Eyre-Walker,A. (2012) The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.*, **33**, 136–143.
16. Roberts,S.A. and Gordenin,D.A. (2014) Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer*, **14**, 786–800.
17. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
18. Gonzalez-Perez,A., Mustonen,V., Reva,B., Ritchie,G.R., Creixell,P., Karchin,R., Vazquez,M., Fink,J.L., Kassahn,K.S., Pearson,J.V. *et al.* (2013) Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods*, **10**, 723–729.
19. Greenman,C., Stephens,P., Smith,R., Dalgliesh,G.L., Hunter,C., Bignell,G., Davies,H., Teague,J., Butler,A., Stevens,C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.
20. Martincorena,I., Raine,K.M., Gerstung,M., Dawson,K.J., Haase,K., Van Loo,P., Davies,H., Stratton,M.R. and Campbell,P.J. (2017) Universal patterns of selection in cancer and somatic tissues. *Cell*, **171**, 1029–1041.
21. Mularoni,L., Sabarinathan,R., Deu-Pons,J., Gonzalez-Perez,A. and Lopez-Bigas,N. (2016) OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.*, **17**, 128.
22. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
23. Hansen,N.F. (2016) Variant calling from next generation sequence data. *Methods Mol. Biol.*, **1418**, 209–224.
24. Grossman,R.L., Heath,A.P., Ferretti,V., Varmus,H.E., Lowy,D.R., Kibbe,W.A. and Staudt,L.M. (2016) Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, **375**, 1109–1112.
25. Helleday,T., Eshtad,S. and Nik-Zainal,S. (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, **15**, 585–598.
26. Alexandrov,L.B. and Stratton,M.R. (2014) Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.*, **24**, 52–60.
27. Sim,N.L., Kumar,P., Hu,J., Henikoff,S., Schneider,G. and Ng,P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, **40**, W452–W457.
28. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, doi:10.1002/0471142905.hg0720s76.
29. Schwarz,J.M., Cooper,D.N., Schuelke,M. and Seelow,D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, **11**, 361–362.
30. Hecht,M., Bromberg,Y. and Rost,B. (2015) Better prediction of functional effects for sequence variants. *BMC Genomics*, **16**, S1.
31. Hopf,T.A., Ingraham,J.B., Poelwijk,F.J., Scharfe,C.P., Springer,M., Sander,C. and Marks,D.S. (2017) Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, **35**, 128–135.
32. Dong,C., Wei,P., Jian,X., Gibbs,R., Boerwinkle,E., Wang,K. and Liu,X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.
33. Brandes,N., Ofer,D. and Linial,M. (2016) ASAP: a machine learning framework for local protein properties. *Database*, **2016**, baw133.
34. Ofer,D. and Linial,M. (2015) ProFET: feature engineering captures high-level protein functions. *Bioinformatics*, **31**, 3429–3436.
35. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D. and Jang,W. (2017) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
36. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2014) OMIM. org: online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
37. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R. and Dubourg,V. (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
38. Yen,J.L., Garcia,S., Montana,A., Harris,J., Chervitz,S., Morra,M., West,J., Chen,R. and Church,D.M. (2017) A variant by any name: quantifying annotation discordance across tools and clinical databases. *Genome Med.*, **9**, 7.
39. Santarius,T., Shipley,J., Brewer,D., Stratton,M.R. and Cooper,C.S. (2010) A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer*, **10**, 59–64.
40. Kandoth,C., McLellan,M.D., Vandin,F., Ye,K., Niu,B., Lu,C., Xie,M., Zhang,Q., McMichael,J.F., Wyczalkowski,M.A. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.
41. Jones,S., Li,M., Parsons,D.W., Zhang,X., Wesseling,J., Kristel,P., Schmidt,M.K., Markowitz,S., Yan,H., Bigner,D. *et al.* (2012) Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types. *Hum. Mutat.*, **33**, 100–103.
42. Morris,L.G., Kaufman,A.M., Gong,Y., Ramaswami,D., Walsh,L.A., Turcan,S., Eng,S., Kannan,K., Zou,Y., Peng,L. *et al.* (2013) Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nat. Genet.*, **45**, 253–261.
43. Stamos,J.L. and Weis,W.I. (2013) The beta-catenin destruction complex. *Cold Spring Harb. Perspect. Biol.*, **5**, a007898.
44. Abbott,K.L., Nyre,E.T., Abrahante,J., Ho,Y.Y., Isaksson Vogel,R. and Starr,T.K. (2015) The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.*, **43**, D844–D848.

45. Pinero,J., Bravo,A., Queralt-Rosinach,N., Gutierrez-Sacristan,A., Deu-Pons,J., Centeno,E., Garcia-Garcia,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
46. Bond,C.E., McKeone,D.M., Kalimutho,M., Bettington,M.L., Pearson,S.A., Dumenil,T.D., Wockner,L.F., Burge,M., Leggett,B.A. and Whitehall,V.L. (2016) RNF43 and ZNRF3 are commonly altered in serrated pathway colorectal tumorigenesis. *Oncotarget*, **7**, 70589–70600.
47. Zhang,D., Zaugg,K., Mak,T.W. and Elledge,S.J. (2006) A role for the deubiquitinating enzyme USP28 in control of the DNA-damage response. *Cell*, **126**, 529–542.
48. Boussouar,F., Jamshidikia,M., Morozumi,Y., Rousseaux,S. and Khochbin,S. (2013) Malignant genome reprogramming by ATAD2. *Biochim. Biophys. Acta*, **1829**, 1010–1014.
49. Ciro,M., Prosperini,E., Quarto,M., Grazini,U., Walfridsson,J., McBlane,F., Nucifero,P., Pacchiana,G., Capra,M., Christensen,J. *et al.* (2009) ATAD2 is a novel cofactor for MYC, overexpressed and amplified in aggressive tumors. *Cancer Res.*, **69**, 8491–8498.