



# Functional Evolutionary Modeling Exposes Overlooked Protein-Coding Genes Involved in Cancer

Nadav Brandes<sup>1</sup>(✉), Nathan Linial<sup>1</sup>, and Michal Linial<sup>2</sup>(✉)

<sup>1</sup> School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

nadav.brandes@mail.huji.ac.il

<sup>2</sup> Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

michall@cc.huji.ac.il

**Abstract.** Numerous computational methods have been developed to screening the genome for candidate driver genes based on genomic data of somatic mutations in tumors. Compiling a catalog of cancer genes has profound implications for the understanding and treatment of the disease. Existing methods make many implicit and explicit assumptions about the distribution of random mutations. We present FABRIC, a new framework for quantifying the evolutionary selection of genes by assessing the functional effects of mutations on protein-coding genes using a pre-trained machine-learning model. The framework compares the estimated effects of observed genetic variations against all possible single-nucleotide mutations in the coding human genome. Compared to existing methods, FABRIC makes minimal assumptions about the distribution of random mutations. To demonstrate its wide applicability, we applied FABRIC on both naturally occurring human variants and somatic mutations in cancer. In the context of cancer, ~3 M somatic mutations were extracted from over 10,000 cancerous human samples. Of the entire human proteome, 593 protein-coding genes show statistically significant bias towards harmful mutations. These genes, discovered without any prior knowledge, show an overwhelming overlap with contemporary cancer gene catalogs. Notably, the majority of these genes (426) are unlisted in these catalogs, but a substantial fraction of them is supported by literature. In the context of normal human evolution, we analyzed ~5 M common and rare variants from ~60 K individuals, discovering 6,288 significant genes. Over 98% of them are dominated by negative selection, supporting the notion of a strong purifying selection during the evolution of the healthy human population. We present the FABRIC framework as an open-source project with a simple command-line interface.

**Keywords:** Driver genes · Machine learning · TCGA · Positive selection · Cancer evolution · Single nucleotide variants · ExAC

## 1 Introduction

Most arising somatic mutations in cancer 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 [1, 2]. The Cancer Genome Atlas (TCGA) is a valuable resource of genomic data from cancer patients covering >10,000 samples in over 30 cancer types [3]. An ongoing effort in cancer research is compiling a comprehensive catalog of cancer genes which have a role in tumorigenesis.

Numerous computational frameworks have been designed for the purpose of identifying suspect cancer genes [4–6]. Most of these frameworks, regarded as “frequentist”, are based on the premise that cancer genes are recurrent across samples and can be recognized by high 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 and genomic loci [7], the frequentist approach requires complex modeling of gene mutation rates as a function of the composition of samples, cancer types and specific loci in the genomes that display extreme deviation in their mutation rates [8]. The sensitivity of the frequentist approach to modeling choices leads to lingering uncertainty and controversy [4].

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, 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. Other variables, 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.

A simplistic functionalist model is based on the ratio of non-synonymous to synonymous (dN/dS) mutations [9]. This model is a common metric for studying the evolutionary selection of a gene. A richer functionalist model was recently explored by OncodriveFML [10]. It estimates the pathogenicity of mutations using CADD [11], 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. OncodriveFML still uses a complex background model that includes sample identities and cancer types. As a result of its complex background model, it requires computationally demanding permutation tests and is unable to analytically calculate probabilities.

With the goal of developing an analytical functionalist model, we introduce a new framework called FABRIC (Functional Alteration Bias Recovery In Coding-regions) [12]. 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 calculation of precise p-values per gene. As a result, FABRIC can provide a detailed ranking of all genes by significance.

The full description of FABRIC and its demonstration to cancer is available elsewhere [12]. In this report, we iterate the highlights of that work and further demonstrate the applicability of FABRIC to broad evolutionary contexts. In particular, we show its ability to detect a trend of negative selection in the context of naturally occurring human genetic variations.

## 2 Methods

### Framework Overview

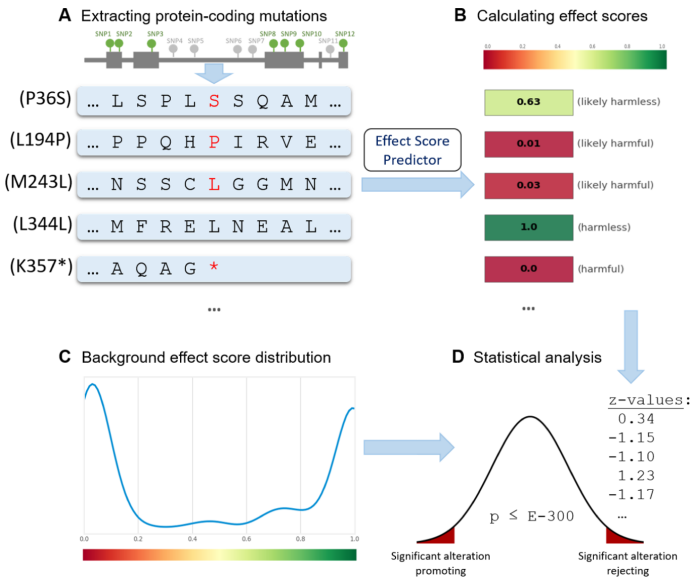
FABRIC analyzes each protein-coding gene independently, extracting all the single nucleotide variations (SNVs) observed within the coding regions of that gene (Fig. 1A). It then uses FIRM, a machine-learning model to assign functional effect scores to each SNV, which measure the predicted effects of those variants explicitly on the protein function (Fig. 1B). 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 and loss of function (LoF) mutations are assigned a 0 score. Missense mutations are processed through FIRM [12] to obtain a score between 0 to 1. Notably, FIRM 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 (Fig. 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 (Fig. 1D). Z-values measure the strengths of deviations between observed to expected scores, and used to derive exact p-values. If a gene's average z-value is significantly negative, it means that mutations 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. An observed score that is significantly higher than expected indicates a gene that is more constrained than expected. We refer to these genes as "alteration rejecting".

We illustrate FABRIC's background model by analyzing *TP53*, one of the most studied cancer gene (Fig. 1E-H; see details in [12]). Importantly, we derive 12 background distributions, corresponding to the 12 possible single-nucleotide substitutions (Fig. 1F). These distributions are gene specific and are independent of the input data (Fig. 1E). Hence, the background model accounts for the exact number of mutations and their SNV frequencies as observed for the studied gene. The mixed model background distribution is gene specific (Fig. 1G). Note that we only considered SNVs (and ignored in-frame indels and splicing variants) at coding regions. Following such simplification, 93% of the somatic mutations in the analyzed dataset is considered. Note that by ignoring complex variations and effects, FABRIC underestimates the damage to gene function.

### Effect Score Prediction Mode

A key component of FABRIC is a pre-trained machine-learning model for predicting the effects of missense genetic variants on protein function. This machine-learning model is called FIRM (Functional Impact Rating at the Molecular-level; Fig. 2). Different from many mutation prediction tools (e.g. CADD [11], Polyphen2 [13]) that predict clinical pathogenicity scores, FIRM seeks positive selection at the biochemical, functional level. Importantly, FIRM was pre-trained on ClinVar [14], which is independent to the datasets used in this work.



◀**Fig. 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) 3,167 SNVs were observed in coding regions of *TP53* from which a  $4 \times 4$  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 3,167 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 3,167 SNVs observed in *TP53* which are categorized as follows: 92 synonymous mutations (effect scores of 1), 512 nonsense mutations (effect scores of 0), and 2,563 missense mutations with an average score of 0.02.

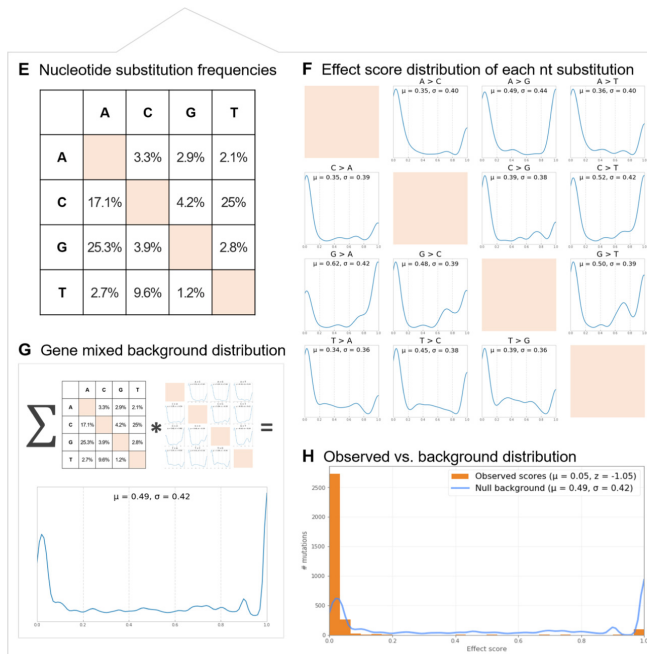
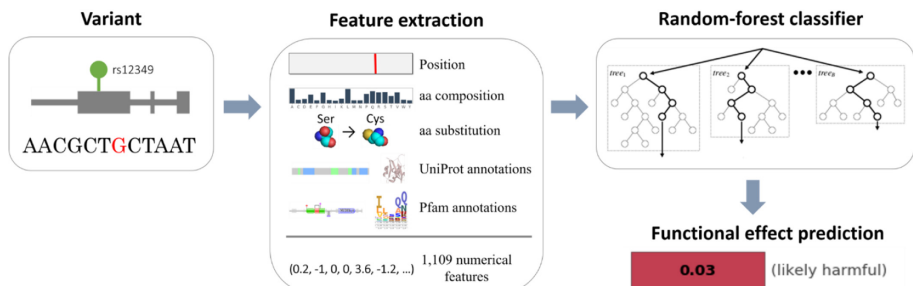


Fig. 1. (continued)

In order to ensure that FIRM does not capture any clinical or evolutionary information, we restricted its used features to purely biochemical properties. FIRM extracts an immense set of features (1,109 in total), aimed at capturing the rich proteomic context of each missense variant. The main classes of features include: 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 UniProtKB, v) amino-acid scales (i.e.



**Fig. 2. Overview of FIRM.** FIRM is the underlying machine-learning model used to predict the functional effects of variants, which is used by the FABRIC framework. By exploiting a rich proteomic knowledgebase, FIRM extracts features representing variants in a 1,109-dimensional space. A random-forest classifier then assigns each variant a predicted functional effect score. FIRM was pre-trained on the ClinVar dataset (which is independent of the datasets examined by FABRIC in this work).

various numeric values assigned to amino-acids [15, 16]), vi) Pfam domains and Pfam clans. More details on FIRM, including performance analysis, are described in [12].

### 3 Results and Discussion

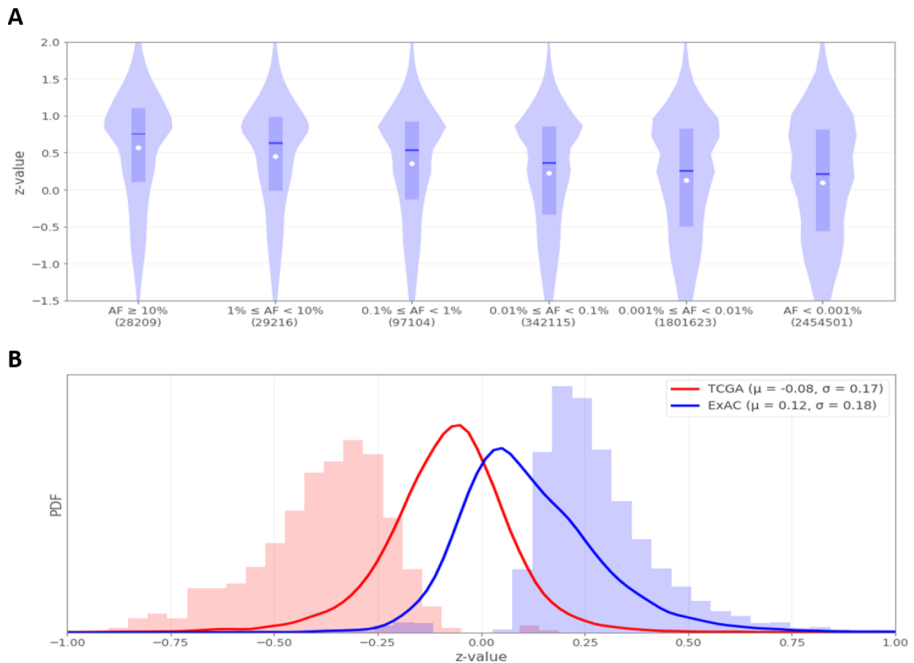
**Alteration Bias in Cancer** We applied FABRIC on  $\sim 3$  M somatic mutations from over 10,000 cancerous human samples extracted from the TCGA database [3]. Of the entire human proteome, we discovered 593 alteration promoting protein-coding genes, namely genes showing statistically significant bias towards harmful mutations [12]. To verify our results and check for new discoveries, we compared our results against prominent resources of cancer genes: COSMIC-Census catalogue [17], and the recently compiled PanSoftware catalogue of 299 cancer driver genes [6]. We found a very strong and significant overlap between the discovered alteration promoting genes and those catalogues [12], although the majority of the genes (426 of 593) were not listed in them.

#### Alteration Bias in the Healthy Human Population

We tested the evolutionary signal that can be extracted from germline variants in healthy human population. We used the ExAC dataset [18], one of the largest and most complete contemporary catalogs of genetic variation in the healthy human population. The full dataset of ExAC contained 10,089,609 variants. We filtered out 1,054,475 low-quality variants, and among the remaining 9,035,134 variants we found 8,538,742 SNVs. Of these, 4,747,096 were found to be in coding regions, contributing to a final dataset of 4,752,768 gene effects. Applying FABRIC on this dataset, the effect scores of the variants in each gene were compared against the background distribution derived from the nucleotide substitution frequencies of the same observed variants.

We observed that variants with lower allele frequencies have lower z-values, i.e. are generally more damaging than expected (Fig. 3A). As expected, more harmful variants (with lower z-values) are less likely to become fixed in the population. We also found expected correlations between the effect score biases of genes (mean z-values) to other popular scoring techniques that measure evolutionary selection. We report Spearman's correlation of  $\rho = -0.4$  (p-value  $< E-300$ ) between the Residual Variation Intolerance Score (RVIS) [19] to the mean effect score z-values of genes. Similarly, we report Spearman's correlation of  $\rho = -0.28$  (p-value  $< E-300$ ) to the Gene Damage Index (GDI) [20]. Both metrics give higher scores to genes that are damaged more than expected, while we give lower scores to such genes, hence the expectation for negative correlation. This further confirms the evolutionary constraints reflected by the effect score biases.

Considering all  $\sim 20$ K protein coding genes, we discovered 6,141 significant alteration rejecting genes, and only 147 significant alteration promoting genes. In other words, almost all of the significant results (97.7%) are alteration rejecting genes, meaning that in the case of the healthy human population, most genes are under negative selection. This is the exact opposite to the trend observed in cancer (Fig. 3B). Whereas cancer is dominated by positive selection, germline variants that have undergone selective pressure throughout long-term human evolution are dominated by negative selection.



**Fig. 3. Alteration rejection in the healthy human population. (A)** Alteration bias (measured by z-value) of germline variants from ExAC across ranges of Allele Frequency (AF). The boxes represent the Q1–Q3 ranges, the middle lines the medians (Q2), and the white dots the means. Since there are ~60 k samples in the dataset, the last range (AF < 0.001%) captures only the 2,454,501 effect scores of singleton variants. **(B)** Distribution of alteration bias (measured by mean z-value) of the 17,828 and 17,946 analyzed genes in TCGA (red) and ExAC (blue), respectively. The density plots show the distribution of all analyzed genes, while the shaded histograms only the 599 and 6,288 genes with significant alteration bias in each dataset (comprised of both alteration promoting and alteration rejecting genes in both datasets). (Color figure online)

It is also interesting to note a mild overlap between the alteration promoting genes in cancer, found in the analysis of somatic mutations in TCGA, to alteration rejecting genes in the healthy human population, found in the analysis of germline variants in the ExAC dataset. Of the 17,313 genes that are shared to both analyses, 584 are significant alteration promoters in cancer, 5,995 are significant alteration rejecters in the human population, and 350 are both. According to random hyper-geometric distribution, we would expect only 202 overlapping genes ( $\times 1.73$  enrichment,  $p\text{-value} = 1.17\text{E}-36$ ). This supports the notion that cancer driver genes, which undergo positive selection during tumor evolution, are subjected to negative selection during normal human evolution.

**Funding.** This work was supported by the European Research Council’s grant on High Dimensional Combinatorics (N.B. fellowship) [N.L. grant #339096] and a grant from Yad Hanadiv (M.L. #9660)

## References

1. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., et al.: Cancer genome landscapes. *Science* **339**(80), 1546–1558 (2013)
2. Marx, V.: Cancer genomes: discerning drivers from passengers (2014)
3. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol* **19**, A68 (2015)
4. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., et al.: Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci.* **113**, 14330–14335 (2016). 201616440
5. Gonzalez-Perez, A., Deu-Pons, J., Lopez-Bigas, N.: Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.* **4**, 89 (2012)
6. Bailey, M.H., Tokheim, C., Porta-Pardo, E., et al.: Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018)
7. Lawrence, M.S., Stojanov, P., Mermel, C.H., et al.: Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014)
8. Zhang, J., Liu, J., Sun, J., et al.: Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. *Brief. Bioinform.* **15**, 244–255 (2014)
9. Greenman, C., Stephens, P., Smith, R., et al.: Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007)
10. Mularoni, L., Sabarinathan, R., Deu-Pons, J., et al.: OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016)
11. Kircher, M., Witten, D.M., Jain, P., et al.: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310 (2014)
12. Brandes, N., Linal, N., Linal, M.: Quantifying gene selection in cancer through protein functional alteration bias. *Nucleic Acids Res.* **47**, 6642–6655 (2019)
13. Adzhubei, I., Jordan, D.M., Sunyaev, S.R.: Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7–20 (2013)
14. Landrum, M.J., Lee, J.M., Benson, M., et al.: ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2015)
15. Brandes, N., Ofer, D., Linal, M.: ASAP: A machine learning framework for local protein properties. *Database* (2016). <https://doi.org/10.1093/database/baw133>
16. Ofer, D., Linal, M.: ProFET: feature engineering captures high-level protein functions. *Bioinformatics* **31**, 3429–3436 (2015)
17. Santarius, T., Shipley, J., Brewer, D., et al.: A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010)
18. Karczewski, K.J., Weisburd, B., Thomas, B., et al.: The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017)
19. Petrovski, S., Wang, Q., Heinzen, E.L., et al.: Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013)
20. Itan, Y., Shang, L., Boisson, B., et al.: The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci.* **112**, 13615–13620 (2015)