# Background mutability shapes observed mutational spectrum in cancer and improves driver mutation prediction

Minghui Li[1,#], Anna-Leigh Brown[2,#], Alexander Goncearenco[2*] and Anna R. Panchenko[2*]

[1] School of Biology and Basic Medical Sciences, Soochow University, Suzhou 215123, China

[2] National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, United States

# - co-first author with equal contribution

* - corresponding authors:

Anna Panchenko, panch@ncbi.nlm.nih.gov

Alexander Goncearenco, alexandr.goncearenco@nih.gov

**Abstract**

Identifying driver mutations in cancer is notoriously difficult. To date, recurrence of a mutation in patients remains one of the most reliable markers of mutation driver status. However, some mutations are more likely to occur than others due to differences in background mutation rates arising from various forms of infidelity of DNA replication and repair machinery, endogenous, and exogenous mutagens.

We used cancer-type and mutagen-specific mutability to study the contribution of background processes of mutagenesis and DNA repair in shaping the observed mutational spectrum in cancer. We developed and tested probabilistic model that adjusts the number of mutation recurrences in patients by background mutability in order to find mutations which may be under selection in cancer.

We showed that observed recurrence frequency of cancer mutations scaled with the background mutability, especially for tumor suppressor genes. In oncogenes, however, highly recurring mutations were characterized by relatively low mutability, resulting in a U-shaped trend. Mutations not yet observed in any tumor had relatively low mutability values, indicating that background mutability might limit the mutation occurrence.

We compiled a dataset of missense mutations from 58 genes with experimentally validated functional and transforming impacts from different studies. We found that mutability of driver mutations was lower than the mutability of passengers and consequently adjusting mutation recurrence frequency by mutability significantly improved ranking of mutations and driver prediction. Even though no training on existing data was involved, our approach performed similar or better to the existing state-of-the-art methods.

**Availability:** https://www.ncbi.nlm.nih.gov/research/mutagene/gene

## Introduction

Cancer is driven by changes at the nucleotide, gene, chromatin, and cellular levels. Somatic cells may rapidly acquire mutations, one or two orders of magnitude faster than germline cells [1]. The majority of these mutations are largely neutral (passenger mutations) in comparison to a few driver mutations that give cells the selective advantage leading to their proliferation [2]. Such a binary driver-passenger model can be adjusted by taking into account additive pleiotropic effect of mutations [3, 4]. Mutations might have different functional consequences in various cancer types and patients, they can lead to activation or deactivation of proteins and dysregulation of a variety of cellular processes. This gives rise to high mutational, biochemical, and histological intra- and inter-tumor heterogeneity that explains the resistance of cancer to therapies and complicates the identification of driving events in cancer [5, 6].

Point DNA mutations can arise from various forms of infidelity of DNA replication and repair machinery, endogenous, and exogenous mutagens [6-9]. There is an interplay between processes leading to DNA damage and those maintaining genome integrity. The resulting mutation rate can vary throughout the genome by more than two orders of magnitude [10, 11] due to many factors operating on local and large scales [12-14]. Many studies support point mutation rate dependence on the local DNA sequence context for various types of germline and somatic mutations [9, 11, 13, 15] whereas local DNA sequence context has been identified as a dominant factor explaining the largest proportion of mutation rate variation in germline and soma [10, 16]. Additionally, differences in mutational burden between cancer types suggest tissue type and mutagen exposure as important confounding factors contributing to tumor heterogeneity.

Assessing background mutation rate is crucial in identifying significantly mutated genes [17, 18], sub-gene regions [19, 20], mutational hotspots[21, 22], or prioritizing mutations[23]. This is especially important considering that the functional impact of the majority of changes observed in cancer is poorly understood, in particular for rarely mutated genes [24]. Despite this need, there is a persistent lack of quantitative information on per-nucleotide background rates of cancer somatic mutations in different gene sites in various cancer types and tissues. In this study

we developed probabilistic models that estimate background mutability per nucleotide or codon substitution. Mutability is defined as a probability to obtain a nucleotide or codon substitution purely from the underlying background processes of mutagenesis and repair; those processes act on genome-wide scale and are devoid of cancer selection component affecting a specific genomic (or protein) site. The models (mutational profiles) were constructed under the assumption that vast majority of cancer context-dependent mutations have neutral effects, while only a negligible number of these mutations in specific sites are under selection. To assure this, we removed all recurrent mutations as these mutations might be under selection in cancer. Mutational profiles were calculated by sampling the frequency data on types of mutations and their trinucleotide (for nucleotide mutations) and pentanucleotide (for codon substitutions) contexts regardless of their genomic locations. These models in the forms of mutational profiles can be used to estimate the expected mutation rate in a given exonic site as a result of different local or long-range context-dependent mutational processes.

There are many computational methods that aim to detect driver genes and fewer methods trying to rank mutations with respect to their potential carcinogenicity. As many new approaches to address this issue have been developed [25] [26], it still remains an extremely difficult task and many driver mutations, especially in oncogenes, are not annotated as high impact or disease related [27] even though cancer mutations harbor the largest proportion of harmful variants [28]. We apply our model to decipher the contribution of background DNA mutability in the observed mutational spectrum in cancer for missense, nonsense, and silent mutations. We compiled a set of cancer driver and neutral missense mutations with experimentally validated impacts collected from multiple studies and used this set to verify our method and compare it with other existing methods. Our approach has been implemented online as part of the MutaGene web-server: https://www.ncbi.nlm.nih.gov/research/mutagene/gene.

**Results**

**Mutations not observed in cancer patients have low mutability**

We analyzed all 3,293,538 theoretically possible codon substitutions that could have occurred by single point mutations in 520 cancer census genes and found that only about one percent of them were actually observed in the surveyed 12,013 tumor samples derived from the COSMIC v85 dataset (Table S1). For codon substitutions which were not observed, the average mutability ($\mu$ = 1.29 x $10^{-6}$) was found to be three-fold lower compared to the mutability (see Methods) of codon substitutions observed in patients for all types of mutations ($\mu$ = 3.88 x $10^{-6}$), Mann-Whitney-Wilcoxon test, $p < 0.01$. This finding holds true for per-nucleotide mutability ($\mu$ = 1.04 x $10^{-6}$ versus $\mu$ = 3.36 x $10^{-6}$, Mann-Whitney-Wilcoxon test, $p < 0.01$) (Figure 1). As a validation of this finding, we also explored mutability and mutation frequency in a set of 9,228 patients who had undergone prospective sequencing of MSK-IMPACT gene panel (Figure S1). Looking at mutations in the genes which were sequenced in all patients in the MSK-IMPACT cohort, the same pattern remains that observed codons mutations had a higher mutability ($\mu$ = 3.41 x $10^{-6}$), compared to those which were theoretically possible, but did not occur ($\mu$ = 1.29 x $10^{-6}$ ,Mann-Whitney-Wilcoxon test, $p < 0.01$). Importantly, none of this subset of tumor samples were involved in deriving mutability values, thus showcasing the applicability mutability derived from one cohort to another.

Figure S2 shows cumulative and probability density distributions of nucleotide mutability values for all observed in patients and possible mutations in all cancer census genes and for two genes in particular, CASP8 and TP53, as examples. While there are many more possible mutations with low mutability values, the observed cancer spectrum is dominated by mutations with high mutability. A similar pattern is seen using cancer-specific mutability values (Figure S3). Mutations which are not observed in breast, lung carcinoma, colon adenocarcinoma, and skin melanoma cancer samples have substantially lower mutability, and as mutability increases, mutations reoccur in more patients, and this is true for both COSMIC v85 and MSK-Impact cohorts.

**Silent mutations have highest mutability**

Figures 2A,B show the distributions of codon mutability values for all possible missense, nonsense, and silent mutations accessible by single nucleotide base substitutions in the protein-coding sequences of 520 cancer census genes calculated with the pan-cancer background model. Codon mutability spans two orders of magnitude and silent mutations have significantly higher average mutability values ($\mu$ = 5.68 x $10^{-6}$) than nonsense ($\mu$ = 3.44 x $10^{-6}$), or missense mutations ($\mu$ = 3.29 x $10^{-6}$) according to Kruskal-Wallis test ($p < 0.01$) and Dunn's post hoc test ($p < 0.01$) for all comparisons. These differences in codon mutabilities could be a reflection of the degeneracy of genetic code, where multiple silent nucleotide substitutions in the same codon may increase its mutability (as illustrated in Figure S4). However, while the differences between types of mutations are less pronounced for nucleotide mutability (Figure 2C), silent mutations are still characterized by the highest mutability values.

**Background mutability may shape the observed mutational spectrum in cancer**

Under the null model of all mutations arising as a result of background mutational processes, somatic mutations should accumulate with respect to their mutation rate and one would expect a positive correlation between mutability and observed mutational frequency of individual mutations. Indeed, as Figures 2B,D show, this is clearly the case for silent and nonsense mutations. To further investigate this relationship, we calculated Spearman's rank (a non-parametric test taking into account that mutability is not normally distributed) and Pearson linear correlation coefficients between codon mutability and frequencies of mutations observed in 12,301 whole-exome (WES) and whole-genome (WGS) tumor samples in the COSMIC cohort (Figure 3). Pooling together all types of mutations in cancer census genes resulted in a very low but significant positive correlation between codon mutability and mutation frequency ($\rho = 0.13$ and $r = 0.02$ for Spearman and Pearson, respectively, $p < 0.01$), as shown in Figure S5A.

Breaking up all codon changes into silent, nonsense and missense reveals higher correlations, particularly for silent ($\rho = 0.14$, $r = 0.1$, $p < 0.01$) and nonsense ($\rho = 0.20$, $r = 0.15$, $p < 0.01$) mutations (Figure S5A). We also calculated correlation coefficients for each gene with at least ten unique mutations of each type: silent, nonsense, and missense (Figure 3). Overall, we

found 84 and 137 genes with significant ($p < 0.01$) positive Spearman and Pearson correlations, respectively, for at least one mutation type (Table S2). Among the genes with significant correlations, 41 belonged to tumor suppressor genes, 28 were oncogenes, and 15 genes were classified as either fusion genes or both oncogene and tumor suppressor. For some genes, including TP53 (second column, Figure 3B) and tumor suppressor CASP8 (third column, Figure 3B), a strong linear relationship between mutability and recurrence frequency of observed mutations ($R^2 > 0.5$) was observed.

**Relationship between mutability and observed frequency is different for tumor suppressor and oncogenes**

The effects of mutations on protein function, with respect to their cancer transforming ability, can drastically differ in tumor suppressor genes (TSG) and oncogenes, therefore we performed our analysis separately for these two categories (Figure 4). We used COSMIC gene classification separating genes into tumor suppressors and oncogenes. Genes which were annotated as both TSG and oncogenes were excluded from this analysis. Mutations in TSG can cause cancer through the inactivation of their products, whereas mutations on oncogenes may result in protein activation. In addition, we used COSMIC classification into genes with dominant or recessive mutations, but overall results were similar to the ones produced using classification into TSG and oncogenes (Figure S6). The strongest correlation between codon mutability and mutation recurrence frequency was observed for TSG ($\rho = 0.17,\ r = 0.13, p < 0.01$) while oncogenes showed a weak Spearman correlation, and no significant Pearson correlation ($\rho = 0.13, p < 0.01, r = 0, p = 0.61$) (Figure S5).

A U-shaped trend was detected for missense and silent mutations in oncogenes: highly recurring mutations (observed in three and more samples) were characterized by a low average mutability (Figure 4). In the latter case, selection may be a more important factor compared to background mutation rate explaining reoccurrence of these mutation. Functional conserved sites overall were found to be more frequently mutated in oncogenes [29], although our analysis did not find

a straightforward association between site mutability and its evolutionary conservation (data not shown).

**Functionally neutral mutations have higher mutability**

We complied a *combined* dataset of experimentally annotated missense mutations in cancer genes from several sources which were categorized as 'non-neutral' or 'neutral' based on their experimental effect on protein function, transforming effects, and other characteristics (see Methods). For our *combined* dataset, the mutability values of 'neutral' mutations were significantly higher (Mann-Whitney-Wilcoxon test, $p < 0.01$) than for 'non-neutral' mutations (Figure 5A). Binning the mutations by their reoccurrence frequency also showed differences between 'neutral' and 'non-neutral', with the frequency of 'neutral' mutations depending on their mutability, so that mutations that were not observed in patients had the lowest mutability ($\mu$ = 2.54 x $10^{-6}$), while those 'neutral' mutations that were observed in 3 or more samples had background mutability more than double ($\mu$ = 6.22 x $10^{-6}$). The differences between non-neutral mutations which were not observed in patients ($\mu$ = 1.59 x $10^{-6}$) and highly recurring mutations was considerably less pronounced ($\mu$ = 2.22 x $10^{-6}$). These effects persisted when examining the mutations binned by frequency in the MSK-IMPACT cohort (Figure S6).

**Accounting for context-dependent mutability in ranking of mutations**

In the previous sections we explored the contribution of background mutational processes in understanding the observed mutational patterns in cancer, with the ultimate goal to facilitate the detection of cancer driver mutations or provide a reasonable ranking in terms of their potential carcinogenic effects. Therefore, we tested how well our method could differentiate between experimentally annotated neutral, or putatively passenger mutations, and non-neutral driver mutations. For this purpose, we compiled a dataset from several experimental studies which included 5,276 annotated mutations in 58 different genes: 4,137 passenger mutations and 1,139 driver mutations (Table S3). We developed two measures which utilize background mutability. The first measure, log- ratio (LR, equation 4), was calculated as a logarithm of the ratio of the number of observed and expected mutations. The second measure (B-score, equation 5)

was calculated from the binomial distribution as a probability to find a given mutation in a certain number of tumor samples or higher given the background context-dependent mutation rate. We compared the performance of these two scores to five other widely used computational methods, CHASM [30], VEST[31], REVEL[32], CanDrAplus[33]and FatHMM[34].

Table 1 and Figure S9 show the performance of the various computational predictors at classifying the *combined* dataset. The best classifier on this dataset is REVEL, an ensemble method which uses the scores of 13 different tools. Intriguingly, despite being based only on mutational frequency, background mutability, and cohort size, and having no training involved, B-Score performs comparably well, with a Matthews correlation coefficient (MCC) of 0.53, just below that of REVEL's at 0.54 (Table 1). B-Score ties with FatHMM on the MCC, with FatHMM having higher AUC-ROC than B-Score (0.85 to 0.79), while having a lower sensitivity at 10% FPR (0.40 to 0.54). Observed frequency of the mutation recurrence in the COSMIC v85 cohort alone outperformed two of the trained computational predictors, CHASM and VEST. By integrating cohort size, B-Score is able to provide a correction to observed frequency using codon mutability and yields a much better performance than frequency alone, mutability, or a ratio of the two (LR). Mutability alone performed better than random with an MCC of 0.17 emphasizing the fundamental quality of non-neutral mutations in cancer: mutability of driver mutations is lower than the mutability of passengers (Figure 5). Because B-Score relies on frequency of recurrence, we also explored performance on mutations that were not observed or observed only once in the COSMIC v85 cohort (Table S5). When mutation frequency is held constant, B-Score and mutability give the same rank order. For mutations which are not observed in the COSMIC cohort (3,886 passenger mutations and 621 driver mutations) B-Score gives a non-random classification of MCC = 0.17. Intriguingly, on mutations which are observed in only one cancer sample in the cohort (207 passenger and 157 driver mutations), B-Score outperforms some of the methods.

Our approach also allows to break ties for mutations observed in the same number of patients. For example in the TP53 gene, mutations p.Arg181Cys, p.Arg282Gln, and p.Arg282Pro have been observed in two patients in a pan-cancer cohort on the MutaGene server. However, p.Arg181Cys

(mutability of 1.00 x $10^{-5}$) is considered a passenger mutation, p.Arg282Gln (mutability of 8.11 x $10^{-6}$) – as a potential driver, and p.Arg282Pro (mutability of 4.90 x $10^{-7}$) – as a driver mutation. Indeed, p.Arg282Pro is annotated as driver mutation in the experimental dataset, and despite the low recurrence frequency, it is correctly classified as a driver by our method. While there are methods based on hotspot identification, our ranking of mutations based on our method cannot be directly compared to the list of hotspots proposed previously[21]. Hotspots are defined for sites, whereas our method assesses specific mutations, and different mutations from the same hotspot can be drivers or passengers. For instance, TP53 Tyr236 site is annotated as a hotspot in [21, 35], however p.Tyr236Phe mutation is experimentally characterized as neutral in the IARC database.

**Variability of mutation rates across genes**

Even though our probabilistic model indirectly incorporates different factors, we checked explicitly if large-scale factors, allowing mutations of the same type to have different mutational probabilities in different genes, affected retrieval performance on our *combined* test set. Several methods have been developed to estimate gene weights (see Methods), which use the overall number of mutations, number of silent mutations affecting a gene, or other factors. We implemented and tested these approaches (see Methods) and in addition estimated the gene weights based on the number of SNPs in the vicinity of a gene. We used gene weights to adjust mutability values and explored whether any of the gene weight models were helpful in distinguishing between experimentally neutral and driver non-neutral mutations. We examined the effects of several large-scale confounding factors such as gene expression levels, replication timing, and chromatin accessibility (provided in the gene covariates files for MutSigCV[36]) on gene weights. We found that "no-outlier"-based weight and "silent mutation"-based weight significantly correlated with the gene expression levels ($r = 0.66, p = 0.004$ and $r = 0.65, p = 0.004$, respectively). Overall, using gene weight as an adjustment for varying background mutational rates across genes did not improve classification performance. Only a SNP-based weight affected the AUC-ROC, but the gain was very minimal, and no gene weight affected MCC (Table S6).

**Ranking of cancer point mutations in MutaGene**

MutaGene webserver provides a collection of cancer-specific context-dependent mutational profiles and allows to calculate nucleotide and codon mutability and B-Score for missense, nonsense and silent mutations for any given protein coding DNA sequence and background mutagenesis model using the "Analyze gene" option. Following the analysis presented in this study, we added options to provide a ranking of mutations observed in cancer samples based on the B-Score or the multiple-testing adjusted q-values. Using the *combined* dataset as a performance benchmark (Table 1, Figure S9), we calibrated two thresholds: the first corresponds to the maximum MCC, and the second corresponds to 10% false positive rate. Mutations with B-Score below the first threshold are predicted to be "cancer drivers", whereas mutations with scores in between two thresholds are predicted to be "potential drivers". All mutations with scores above the second threshold are predicted as "passengers". Importantly, calculations are not limited to pan-caner and can be performed using a mutational profile for any particular cancer type, the latter would result in a cancer-specific ranking of mutations and could be useful for identification of driver mutations in a particular type of cancer. An example of prediction of driver mutations status for EGFR is shown in Figure 6.

**Discussion**

To understand what processes drive point mutation accumulation in cancer, we used DNA context-dependent probabilistic models to estimate the baseline mutability for nucleotide mutation or codon substitution in specific genomic sites. Passenger mutations, constituting the majority of all observed mutations, may have largely neutral functional impacts and are unlikely to be under selection pressure. For passenger mutations one would expect that mutations with lower DNA mutability would have lower observed mutational frequency and vice versa. In accordance with this expectation, we detected a significant positive correlation between background mutability, which is an estimate of per site mutation rate, and observed frequencies of mutations in cancer patients. In a recent study the fraction of sites harboring SNPs in the human genome was found indeed to correlate very well with the mutability although the later was estimated differently from our study [37]. We also found that cancer mutations not so far

observed in cancer patients had much lower expected background nucleotide and codon mutability compared to the observed mutations  For some genes, the observed frequency of occurrence of mutations can be predicted purely from their mutability. Outliers of this trend are important for inferring mutations under selection. For instance, if a mutation with low expected mutation rate is observed in multiple patients, it is suggestive of it being potentially important in carcinogenesis. In this respect, reoccurring synonymous mutations with low mutability may represent interesting cases for further investigation of potential synonymous drivers.

Mutability of synonymous mutations was found to be the highest among other types of mutations; and observed mutational frequency of these mutations scaled very well with their mutability. Overall, our method predicted 102 synonymous driver mutations in 64 out of 520 cancer-associated genes. Indeed, it has been previously shown that some synonymous mutations might be under selection, and can affect the speed and accuracy of transcription and translation, protein folding rate, and splicing [38]. Since observed mutational frequency of synonymous mutations scales with their mutability (Figures 1,4), it is important to correct for mutability while ranking these mutations with respect to their driver status. Some recurrent highly mutable synonymous mutations might not represent relevant candidates of drivers, whereas some rare mutations with relatively low mutability are predicted to be drivers by our method (e.g. KDR gene p.Leu355=, NTRK1 gene p.Asn270=).

We developed and tested a probabilistic model to adjust the number of reoccurrences of a mutation by its expected background mutability in order to find those sites and mutations which may be under selection in cancer. We find that while including background mutability improves performance over reoccurrence alone in distinguishing between driver and passenger mutations, the choice of how to account for mutability is also important. Our B-Score integrates information about observed rate and total cohort size, something not captured by an odds ratio alone. The advantages of this model are that: (i) it is intuitive (ii) does not rely on many parameters and (iii) does not involve explicit training on driver and passenger mutation sets. One of the disadvantages of this model is that it requires the knowledge of a total number of patients tested.

We found that B-Score exhibited a considerably improved performance of driver mutation prediction compared to the conventional way of using mutational frequency of occurrence. It performed comparably or better to state-of-the-art methods even for rare mutations observed in cancer patients; many of the methods we used for comparison were trained on existing datasets of mutations and relied upon multiple features. However, the performance of B-Score was detrimental for mutations not observed in cancer patients. Moreover, for highly heterogeneous cancer types, the background mutational processes may differ between cohorts of cancer patients, thus an appropriate cancer-specific model for background mutability should be applied. Taken together, our model provides means to explore mutation rates and enables an understanding of the differential roles that background mutation rate and selection play in shaping the observed cancer spectrum.

For our analysis of methods' performance using a dataset with experimentally determined effects of mutations combined from different studies, we had to apply a pan-cancer mutability model. While mutational processes vary widely among cancer types, and different drivers mutations have been shown to be preferentially associate with specific mutational signatures[39, 40] there remains a lack of cancer-specific driver/passenger datasets. It may be insightful for researchers to apply cancer-specific B-Score ranking of mutations using the models available on the MutaGene website.

Some efforts have been focused so far on developing a comprehensive set of cancer driver mutations verified at the levels of functional assays or animal models [26, 41, 42]. However, existing sets often contain predictions and very few neutral passenger mutations. The vast majority of computational prediction methods rely on machine learning algorithms trained on mutations from a few genes and/or on recurrent mutations as estimates of driver events or use germline SNPs or silent mutations as the presumed "neutral" set. In many cases, the performance is evaluated on similarly generated synthetic benchmarks. As a result, methods can be trained on incorrectly labeled data, predicting cancer driver mutations worse than their recurrence frequency (Table 1) or background mutability alone (Table S5). In this study, we

restricted our dataset to only missense mutations that have been experimentally assessed, with overall 4,137 passenger and 1,139 driver mutations from 58 genes. Intriguingly, we found that experimentally annotated driver mutations have a lower background mutability than neutral mutations, suggesting possible action of context-dependent codon bias towards less mutable codons at critical sites for these genes, although more studies would have to be conducted to further investigate this observation.

## Methods

### *Defining drivers and passengers – datasets of experimental functional assays*

Missense mutations for TP53 gene with experimentally determined functional transactivation activities were obtained from IARC P53 database where they were classified as functional, partially-functional, and non-functional[43].

The second dataset, referred hereinafter to as "Martelotto *et al.*", contains experimental evidence collected from literature[44]. The experimental evidence of impact of mutations included changes in enzymatic activity, response to ligand binding, impacts on downstream pathways, an ability to transform human or murine cells, tumor induction *in vivo*, or changes in the rates of progression-free or overall survival in pre-clinical models. In "Martelotto *et al.*" dataset mutations were considered "damaging" if there was literature evidence to support their impact on at least one of the above-mentioned categories. Mutations with no significant impacts on the wild-type protein function were classified as "neutral". Mutations with no reliable functional evidence were regarded as "uncertain" and were not used in this study.

The third dataset included experimentally verified BRCA1 mutations and was originally collected by using deep mutational scanning to measure the effects of missense mutations on the ability of BRCA1 to participate in homology-directed repair. Missense mutations were categorized as either "neutral" or "damaging" [45, 46]. Noteworthy, BRCA1 set contained inherited germline as well as somatic mutations.

The fourth dataset explored over 81,000 tumors to identify drivers of hypermutation in DNA polymerase epsilon and delta genes (POLE/POLD1). “Drivers of hypermutation” were variants which occurred in a minimum of two hypermutant tumors, which were never found in lowly mutated tumors, and did not co-occur with an existing known driver mutation in the same tumor. Other variants in these genes were considered “passengers” with respect to hypermutation[25].

The fifth dataset consisted of missense mutations annotated based on their effects on cell-viability in Ba/FC and MCF10A models[47]. “Activating mutations” were mutations where the cell viability was higher than the wild-type gene, and “neutral mutations” were those mutations where cell-viability was similar to the wild-type. Ng *et al.* used these consensus functional annotations to compare the performance of 21 different computational tools in classifying between activating and neutral mutations using ROC analysis, with activating mutations acting as the positive set and neutral as the negative set. This dataset contained 743 mutations (488 neutral and 255 activating) from 50 genes.

Finally, we assembled a *combined dataset* that included mutations from these five datasets described above. We removed redundant and conflicting entries when mutations annotated as non-functional or neutral in one dataset were also annotated as damaging or benign in another. All mutations were categorized as “non-neutral” (affecting function, binding or transforming) and “neutral” (other mutations). We treated “functional” and “partially -functional” mutations in IARC TP53 dataset as “neutral”, and “non-functional” as “non-neutral”. We used missense mutations in order to compare with cancer FatHMM scores. Overall, the *combined dataset* contains 5,276 mutations (4,137neutral and 1,139 non-neutral) from 58 genes (Table S3, S7).

***Datasets of cancer mutations***

The Catalogue of Somatic Mutations in Cancer (COSMIC) database stores data on somatic cancer mutations and integrates the experimental data from the full-genome (WGS) and whole-exome (WES) sequencing studies[48]. Cancer census genes (520 genes) were defined according to COSMIC release v84. For analyses comparing oncogenes and TSG, genes classified as only fusion

genes or those with both oncogenic and TSG activities were not used. This resulted in 205 oncogenes and 167 TSG (Table S2). A subset of 12,013 tumors from COSMIC release v85 was used to extract observed frequency data, among them 98% of all samples contained less than 1000 mutations so were not hypermutated. COSMIC v85 samples which came from cell-lines, xenografts, or organoid cultures were excluded. Only mutations with somatic status of "Confirmed somatic variant" were included and mutations which were flagged as SNPs were excluded. For each cancer patient, a single sample from a single tumor was used. Additionally, it is possible that the same patient may be assigned different unique identifiers in different papers, and duplicate tumor samples may be erroneously added to COSMIC database during manual curation. These samples may affect the recurrence counts of mutations. We applied clustering method in order to detect and remove any redundant tumor samples. Each sample was represented as a binary vector with 1 if a sample had a mutation in a particular genomic location and 0 otherwise. The binary vectors were compared with Jaccard distance metric, $J = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$, where identical samples have $J = 0$, followed by agglomerative clustering with complete linkage. Non-singleton clusters with pairwise distance cutoff of $J = 0.3$ were extracted and only one representative of each cluster was used, whereas other samples were discarded. Because of this relatively stringent criteria for inclusion, it is likely that some small number of non-duplicate samples were discarded in this process.

MSK-IMPACT cohort was obtained from cBioPortal[49]. We ensured that no mutations were counted multiple times for each patient; if there were multiple tumor samples per patient, primary and metastatic, the primary tumor was kept, and the metastatic discarded. Only those tumors which were sequenced against a "matched" normal sample were kept to ensure validity of somatic mutations.

In the 520 genes we explored, we found that 250 genes contained at least one mutation annotated as a driver by our method. We investigated if these genes were expressed in cancer cell lines from multiple tissue types using RNAseq data from the Cancer Cell Line Encyclopedia [50]. Using RNAseq data of 1,076 unique cancer cell lines from 26 different tissue types and a cut

off for expression at 0.5 RPKM (Reads Per Kilobase of transcript, per Million mapped reads), we found that genes with driver mutations were expressed in 5,646 out of 6,506 gene-tissue comparisons.

***Calculation of context-dependent DNA background mutability***

Context-dependent mutational profiles were constructed from the pools of mutations from different cancer samples by counting mutations observed in a specific trinucleotide context. Altogether, there are 64 different types of trinucleotides and three types of mutations $x \gg y$ (for example C≫A, C≫T, C≫G and so on) in the central position of each trinucleotide which results in 192 trinucleotide context-dependent mutation types. In a mutated double-stranded DNA both complementary purine-pyrimidine nucleotides are substituted, and therefore we considered only substitutions in pyrimidine (C or T) bases, resulting in 96 possible context-dependent mutation types: $m$ = {N [$x \gg y$] N}, where N = [41 T]. Thus, mutational profile can be expressed as a vector of a number of mutations of certain type $(f_1, \dots, f_{96})$. Profiles were constructed under the assumption that vast majority of cancer context-dependent mutations have neutral effects, while only a negligible number of these mutations in specific sites are under selection. To assure this, we removed recurrent mutations (observed twice or more times in the same site) as these mutations might be under selection in cancer. In the current study we used pan-cancer and cancer-specific mutational profiles for breast, lung adenocarcinoma, and skin adenocarcinoma cancer derived from MutaGene [51].

We applied mutational profiles to build DNA context-dependent probabilistic models that described baseline DNA mutagenesis per nucleotide or per codon. ***Mutability*** was defined as a probability to obtain a context-dependent nucleotide mutation purely from the baseline mutagenic processes operating in a given group of samples used to derive the mutational profile. Mutability is proportional to the expected mutation rate of a certain type of mutation (96 altogether) regardless of the genomic site it occurs. It was calculated using the total number of mutations of type $x \gg y$ in a certain local context, which can be obtained from the context-dependent mutational profile, $f_m$ . Given the number of cancer samples used to construct

mutational profile, *N*, and the number of different trinucleotides of type *t* in a diploid human exome, $n_t$ (calculated from the reference genome), the ***nucleotide mutation mutability*** is calculated as:

$$p_m^{nuc} = \frac{f_m}{N\, n_t} \quad (1)$$

In protein-coding sequences it is practical to calculate mutation probability for a codon in its local pentanucleotide context, given trinucleotide contexts of each nucleotide in the codon. For a given transcript of a protein, at exon boundaries the local context of the nucleotides was taken from the genomic context. The COSMIC consensus transcript was chosen for the transcript for each protein. Changes in codons can lead to amino acid substitutions, synonymous and nonsense mutations. Therefore, we calculate ***codon mutability*** as the probability to observe a specific type of codon change which can be realized by single nucleotide mutations at each codon position *i* as:

$$p_M^{codon} = 1 - \prod_i^3 \left(1 - \sum_j^k p_{ij}^{nuc}\right) \quad (2)$$

Where *k* denotes a number of mutually exclusive mutations at codon position *i*. For example, for Phe codon "TTT" in a given context 5'-A and G-3': three single nucleotide mutations can lead to Phe->Leu substitution (to codons "TTG", "TTA" and "CTT" for Leu): A[T≫C]TTG in the first codon position or mutually exclusive ATT[T≫G]G and ATT[T>>A]G in the third codon position (Figure S4). In this case the probability of Phe→ Leu substitution in the $ATTTG$ context can be calculated as $p_{Phe\to Leu}^{codon} = 1 - (1 - p_{A[T\to C]T})(1 - p_{T[T\to A]G} - p_{T[T\to G]G})$ where trinucleotide probabilities were taken from the mutational profile. Amino acid substitutions corresponding to each missense mutation are calculated by translating the mutated and wild type codons using a standard codon table. Codon mutability strongly depends on the neighboring codons as illustrated in Figure S4.

***Gene-weight adjusted mutability***

Gene weights estimate a relative probability of a gene compared to other genes to be mutated in cancer through somatic mutagenesis. There are multiple ways the gene weights can be calculated.

*SNP-based weight* was calculated using the number of SNPs in the vicinity of the gene of interest. We used the "EnsDb.Hsapiens.v86" database to find genomic coordinates of a gene, including introns, and extended the range in both 3' and 5' directions according to the window size (Table S7). We then counted the number of common SNPs from dbSNP database[52] within the genomic region. Gene weight was calculated as: $\omega_g^{SNP} = \frac{n_{SNP}}{L_{window}}$, where $n_{SNP}$ is the number of SNPs and $L_{window}$ is the length of the genomic region in base pairs. We tested several window sizes for defining the genomic regions around the gene of interest (Table S6).

*Mutation-based weight* was calculated using the number of nucleotide sites with reoccurring mutations counted only once to avoid the bias that may be present due to selection on individual sites: $\omega_g^{mut\_sites} = \frac{k_g}{n_k}$. Here $k_g$ is the number of mutated sites and $n_k$ is the number of base pairs in the gene transcript.

*Silent mutation-based weight* was introduced previously and was shown to be superior in assessment of significant non-synonymous mutations across genes[53]. This weight can be calculated by taking into account only silent somatic mutations: $\omega_g^{silent} = \frac{s_g}{N\,L_g}$. Here $s_g$ is the total number of somatic silent mutations within the gene, $N$ is the number of tumor samples and $L_g$ is the number of codons in the gene transcript.

*No-outlier-based weight* introduced previously[21] takes into account the number of all codon mutations within a gene, $C_g$, excluding mutations in outlier codon sites bearing more than the 99$^{th}$ percentile of mutations of the gene: $\omega_g^{out} = \frac{C_g}{N\,L_g}$, normalized by gene length $L_g$ in amino acids and the total number of samples $N$.

Using gene weights, an adjusted probability per codon can be then expressed as:

$$p'_M = \omega_g p_M \quad (3)$$

Similarly, per nucleotide probability can be calculated adjusted by gene weight.

***Identification of significant mutations***

We introduced two measures which take into account the background DNA mutability. The first measure (LR) is calculated as a ratio of the numbers of observed and expected mutations in a given nucleotide site or in a given codon:

$$LR = \log(\frac{n_{obs}}{n_{exp}}) \quad (4)$$

where the expected number of mutations, $n_{exp} = Np_m^{nuc}$ or $n_{exp} = Np_M^{codon}$ where *N* is the number of tumor samples and $p_m^{nuc}$ or $p_M^{codon}$ are calculated using equations (1) or (2) and the observed number of mutations $n_{obs}$ in a given site is taken from COSMIC v85 with a pseudo count correction ($n_{obs} + 1$) to account for mutations that have not been observed due to a limited tumor sample collection.

The second measure uses the binomial model to calculate the probability of observing a certain type of mutation in a given site more frequently than *k*:

$$B_{score} = \sum_{k=n+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} \quad (5)$$

where $p = p'_M$ or $p = p'_m$ and $k$ is the number of observed mutations of a given type at a particular nucleotide or codon. Depending on the dataset chosen or a particular cohort of patients (for instance, corresponding to one cancer type), the total number of samples $N$ and the numbers of observed mutations $k$ will change. While ranking mutations in a given gene, $B_{score}$ can further be adjusted for multiple-testing with Benjamini-Hochberg correction as implemented on the MutaGene website.

***Computational Predictions***

CanDrAplus[34] program was downloaded and ran using default specifications with the "Cancer-in-General" annotation data file. REVEL[33] predictions were obtained from dbNSFP database[54]. CHASM [31] and VEST[32] were obtained using CRAVAT[55]. Several versions of CHASM [31] are available, and we used the version which performed the best on the *combined dataset*. FatHMM[35] cancer-associated scores were obtained from their webserver.

***Statistical and ROC analyses***

Differences between various groups were tested with the Kruskal-Wallis, Dunn test, and Mann-Whitney-Wilcoxon tests implemented in *R* software. Dunn's test is a non-parametric pairwise multiple comparisons procedure based on rank sums; it is used to infer difference between means in multiple groups and was used because it is relatively conservative post-hoc test for Kruskal-Wallis. Associations between mutability and observed frequency (the number of individuals with a mutation in whole-exome/genome studies from COSMIC), was tested using Pearson as well as Spearman correlation tests since the variables were not normally distributed. Where *R* reports the calculated p-value below $2.2 \times 10^{-16}$, the value has been shown as the alpha level, $p < 0.01$.

To quantify the performance of scores, we performed Receiver Operating Characteristics (ROC) and precision-recall analyses. Sensitivity or true positive rate was defined as TPR=TP/(TP + FN) and specificity was defined as SPC=TN/(FP+TN). Additionally, in order to account for imbalances in the labeled dataset, the quality of the predictions was described by Matthews correlation coefficient:

$$\mathrm{MCC} = \frac{TP * TN + \mathrm{FP} * \mathrm{FN}}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

**ACKNOWLEDGEMENT**

We thank Yuri Wolf, Alejandro Schaffer and Igor Rogozin for helpful discussions. The work was supported by Intramural Research Programs of the National Library of Medicine, National Institutes of Health. Minghui Li was supported by the National Natural Science Foundation of

China (Grant No. 31701136) and Natural Science Foundation of Jiangsu Province, China (Grant No. BK20170335)

## Figures and Tables

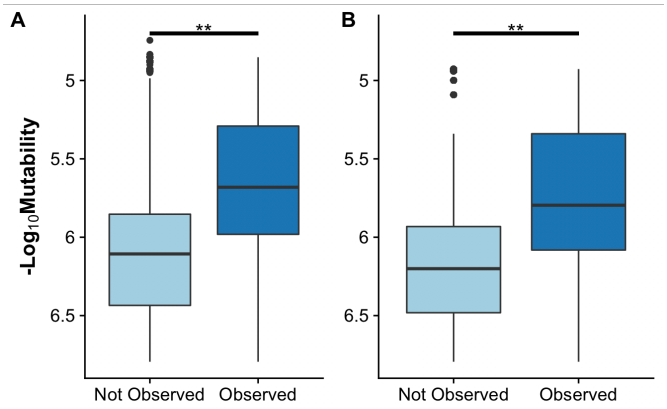



**Figure 1.** Mutability for observed mutations and all theoretically possible mutations that were not observed in COSMIC v85 pan-cancer cohort: **(A)** Codon mutability; **(B)** Nucleotide mutability. Differences on Mann-Whitney-Wilcoxon test significant at $p < 0.01$.

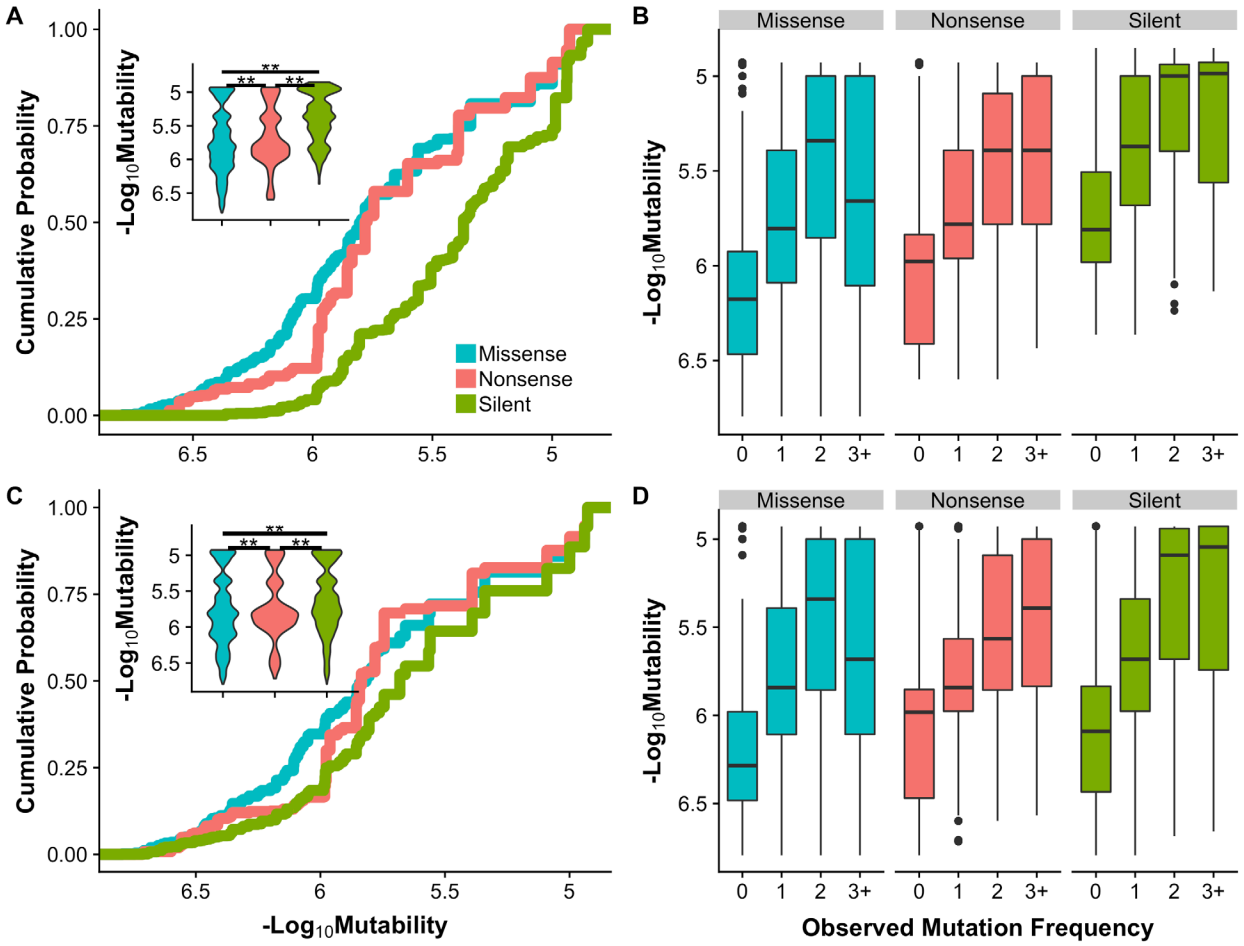



**Figure 2.** Mutability distributions by mutation type and gene role in cancer. **(A)** Cumulative distribution of codon mutability of silent (green), nonsense (red) and missense (blue) mutations. **(C)** Cumulative distribution of nucleotide mutability for silent, nonsense and missense mutations. Inset shows the probability density distributions of mutability by mutation type. Significance was determined by Dunn's test; difference with $p < 0.01$. is marked with a double asterisk. **(B)** and **(D)** are codon and nucleotide mutability respectively binned by frequency in the COSMIC cohort. '0', '1', '2' and '3+' refer to mutations that were not observed (including all possible point mutations), observed once, twice, or in three or more cancer samples.

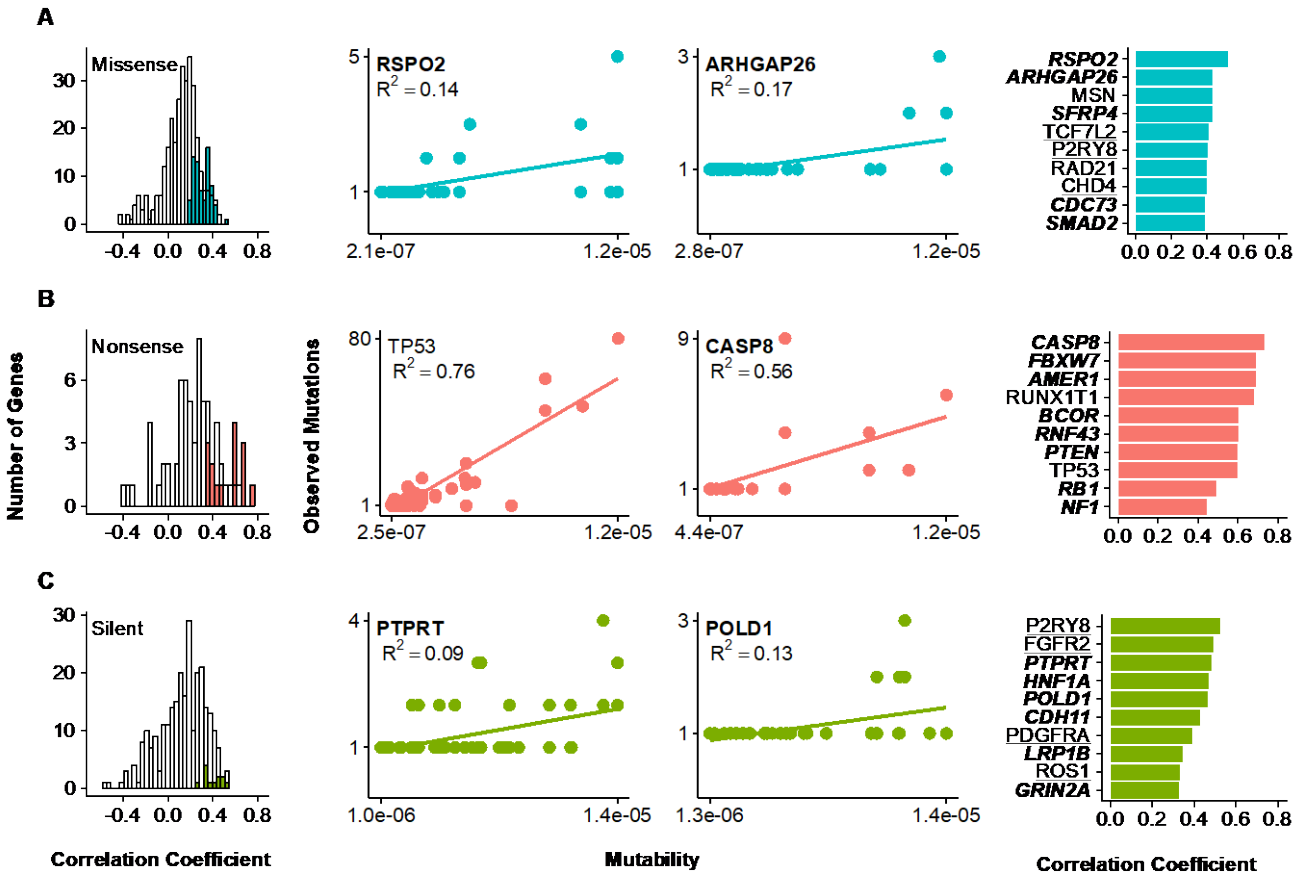



**Figure 3.** Relationship between codon mutability and frequency of mutations. Histograms show the Spearman rank correlation coefficients between observed mutations and mutability across cancer genes with at least 10 observed mutations of each type: **(A)** missense (blue), **(B)** nonsense (red) and **(C)** silent (green). Filled bars in the left column denote genes with significant correlation at $p < 0.01$. Scatterplots with regression lines and confidence intervals show the linear relationship between mutability and observed frequency of each type of mutation for several representative genes. Adjusted $R^2$ shown to convey goodness of fit. Bar graphs show Spearman correlation coefficient for genes with significant correlation at $p < 0.01$. Genes with bold font are tumor suppressors (TSG), underlined genes are oncogenes, and plain font were either categorized as both TSG and oncogene or fusion genes. Mutation frequencies were taken from the pan-cancer COSMIC cohort.

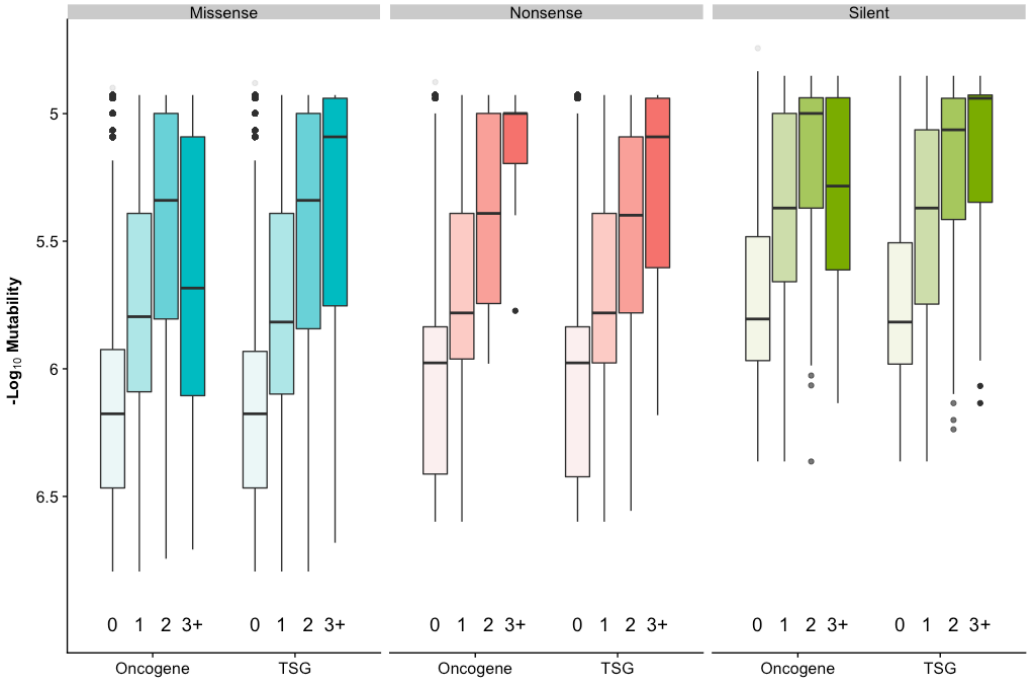



**Figure 4.** Mutability and frequency of mutation by mutation type and gene's role in cancer. Pooled mutations in cancer census genes were grouped for oncogene and tumor suppressor (TSG) genes. Boxplots show codon mutability calculated with pan-cancer model. See Table S1 for counts.

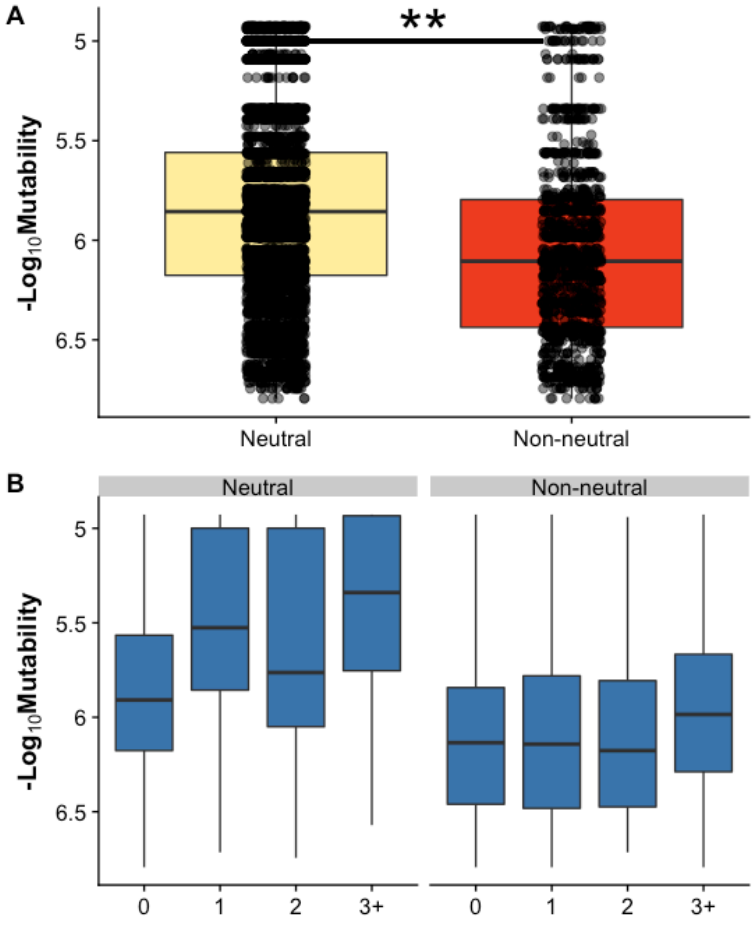



**Figure 5.** Codon mutability of missense mutations grouped by the effect on protein function. **(A)** Mutations from the combined dataset were categorized as neutral and non-neutral. Significant differences with $p < 0.01$ are marked with a double asterisk. Mutability was calculated with pan-cancer background model; see Supplementary Figure S5 for analysis with of individual datasets on different background models. **(B)** Mutations binned by the frequency in the COSMIC v85 cohort; see Supplementary Figure S6 for binning by frequency in MSK-IMPACT cohort.

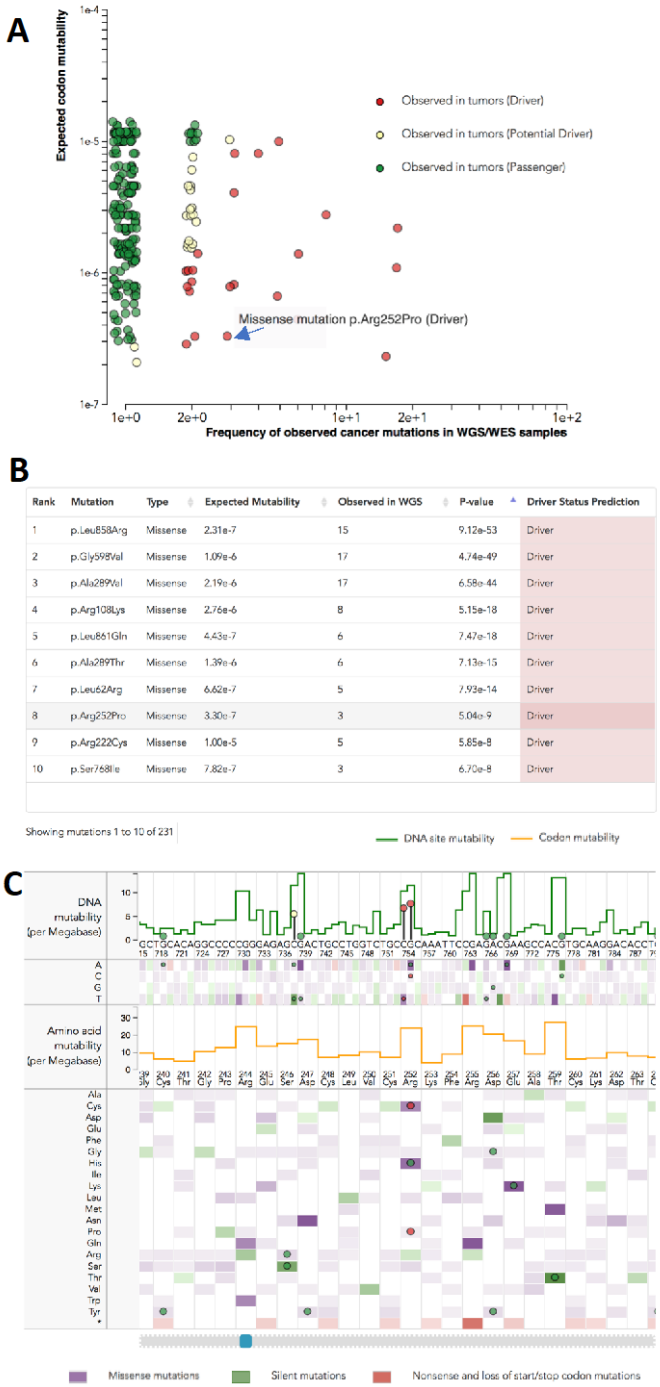



**Figure 6.** Ranking of mutations and prediction of driver mutations based on mutational frequencies adjusted by mutability. Snapshots from MutaGene server show the results of analysis of EGFR gene with a Pan-cancer mutability model. **(A)** Scatterplot with expected mutability versus observed mutational frequencies. **(B)** Top list of mutations ranked by their B-Scores. **(C)** EGFR nucleotide and translated protein sequence shows per nucleotide site mutability (green line), per codon mutability (orange line), as well as mutabilities of nucleotide and codon substitutions (heatmaps). Mutations observed in tumors from ICGC repository are shown as circles colored by their prediction status: Driver, Potential driver, and Passenger.

**Table 1.** Comparison of different methods to distinguish neutral from non-neutral mutations using *combined* experimental dataset. Scores developed in this study (B-Score and LR) are underlined. Performance of mutability is listed as a reference point. AUC-ROC and AUC-PR values for observed frequency counts were extrapolated since some experimentally validated mutations were not observed in tumor samples. Since several versions of CHASM are available, we used the version which performed the best on the *combined dataset*. FatHMM cancer-associated scores were obtained from its webserver.

| Measure | AUC-ROC | AUC-PR | Matthews correlation | Sensitivity at 10% FPR |
|---|---|---|---|---|
| **REVEL** | 0.85 | 0.67 | 0.54 | 0.63 |
| **FatHMM** | 0.85 | 0.59 | 0.53 | 0.40 |
| **B-Score** | 0.79 | 0.65 | 0.53 | 0.54 |
| **CanDrAplus** | 0.83 | 0.52 | 0.52 | 0.41 |
| **LR** | 0.71 | 0.60 | 0.51 | 0.48 |
| **Recurrence frequency** | 0.71 | 0.58 | 0.47 | 0.48 |
| **VEST** | 0.74 | 0.46 | 0.31 | 0.35 |
| **CHASM** | 0.74 | 0.43 | 0.30 | 0.28 |
| **Mutability** | 0.63 | 0.33 | 0.17 | 0.21 |

## Supplementary Figures and Tables

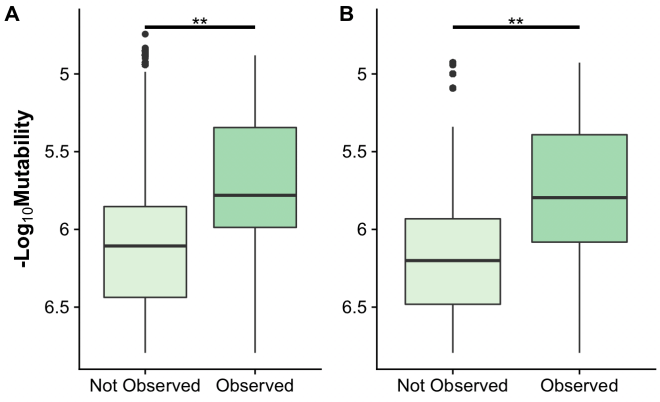



**Figure S1.** Mutability values for observed mutations in MSK-IMPACT subset cohort and all theoretically possible mutations that were not observed in cancer patients: **(A)** Codon mutability; **(B)** Nucleotide mutability. Differences on Mann-Whitney-Wilcoxon test significant at $p < 0.01$.

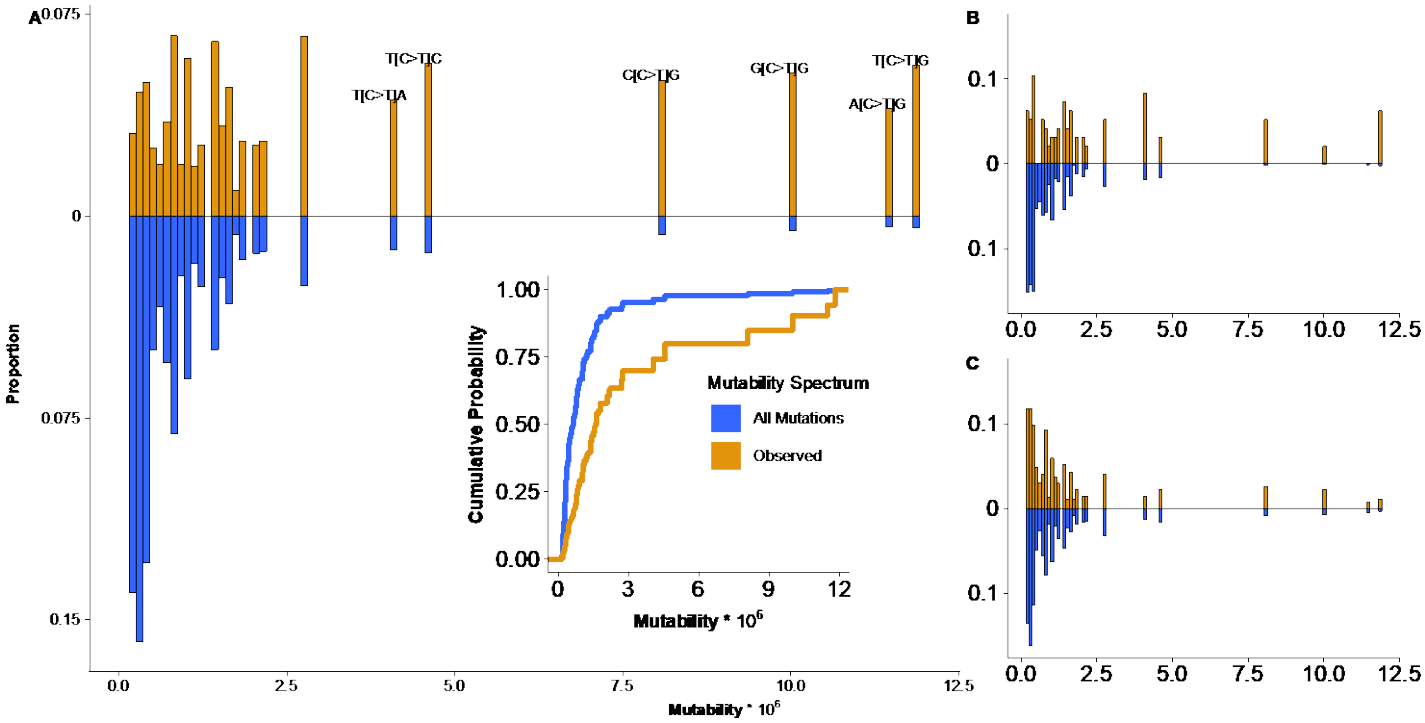



**Figure S2.** Comparison between nucleotide mutability (expected) spectrum of all possible mutations (blue) and mutations which were observed in cancer patients (brown) in the Cosmic 85 cohort. **(A)** Mutations from 520 cancer-related genes; **(B)** CASP8 and **(C)** TP53 genes. Inset shows the cumulative distribution functions for both spectra. Annotations in **(A)** show nucleotide mutation types.

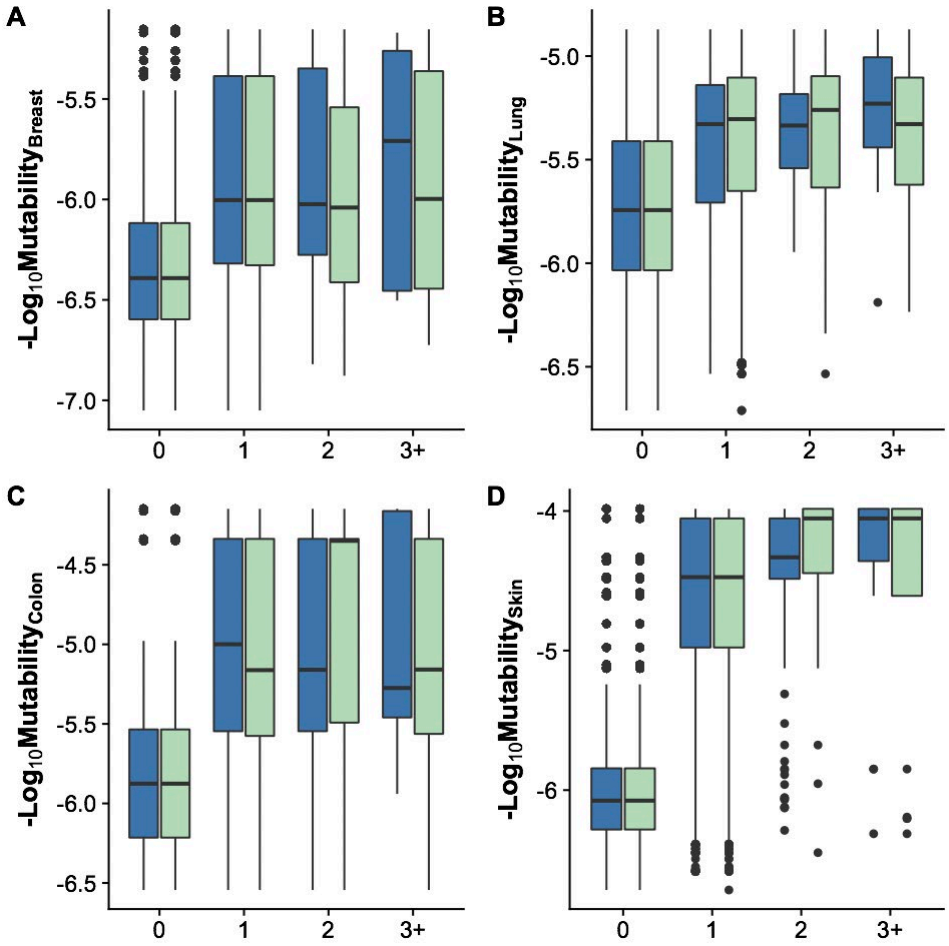
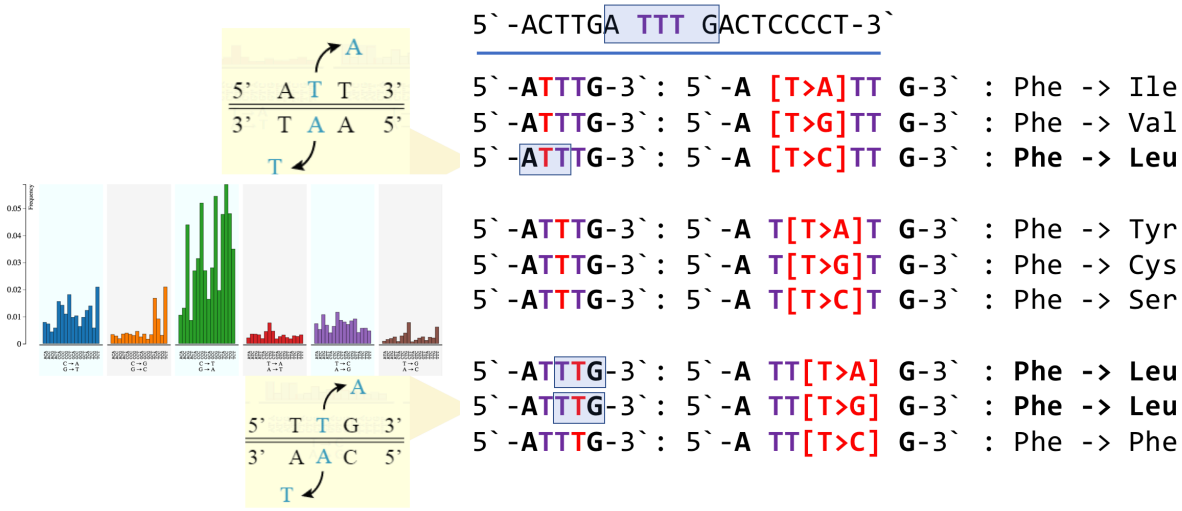



**Figure S3.** Relationship between cancer-specific nucleotide mutability and observed mutation frequency in the COSMIC 85 cohort and in the MSK-IMPACT in a subset of cancer genes. Blue boxes show mutations with the given frequency in the Cosmic 85 cohort and green is MSK-IMPACT. Counts are binned as in Figure 2 and refer to how many times a particular mutation was observed in the given cancer type. '0', '1', '2' **(A)** Breast ($n_{COSMIC}$ = 1,667 $n_{COSMIC}$ =783 samples) **(B)** Lung carcinoma ($n_{COSMIC}$ = 301; $n_{MSK}$ = 1,203) **(C)** Colon adenocarcinoma ($n_{COSMIC}$ = 369; $n_{MSK}$ = 688) **(D)** Skin malignant melanoma ($n_{COSMIC}$ = 376; $n_{MSK}$ =182)




$$p_M^{codon} = 1 - \prod_i^3 \left(1 - \sum_j^k p_{ij}^{nuc}\right)$$

$$p_{Phe \to Leu}^{codon} = 1 - (1 - p_{A[T \to C]T})(1 - p_{T[T \to A]G} - p_{T[T \to G]G})$$

**Figure S4.** Example of calculation of codon mutability from nucleotide mutabilities of all possible Phe → Leu missense mutations in the Phe codon TTT with a specific pentanucleotide context.

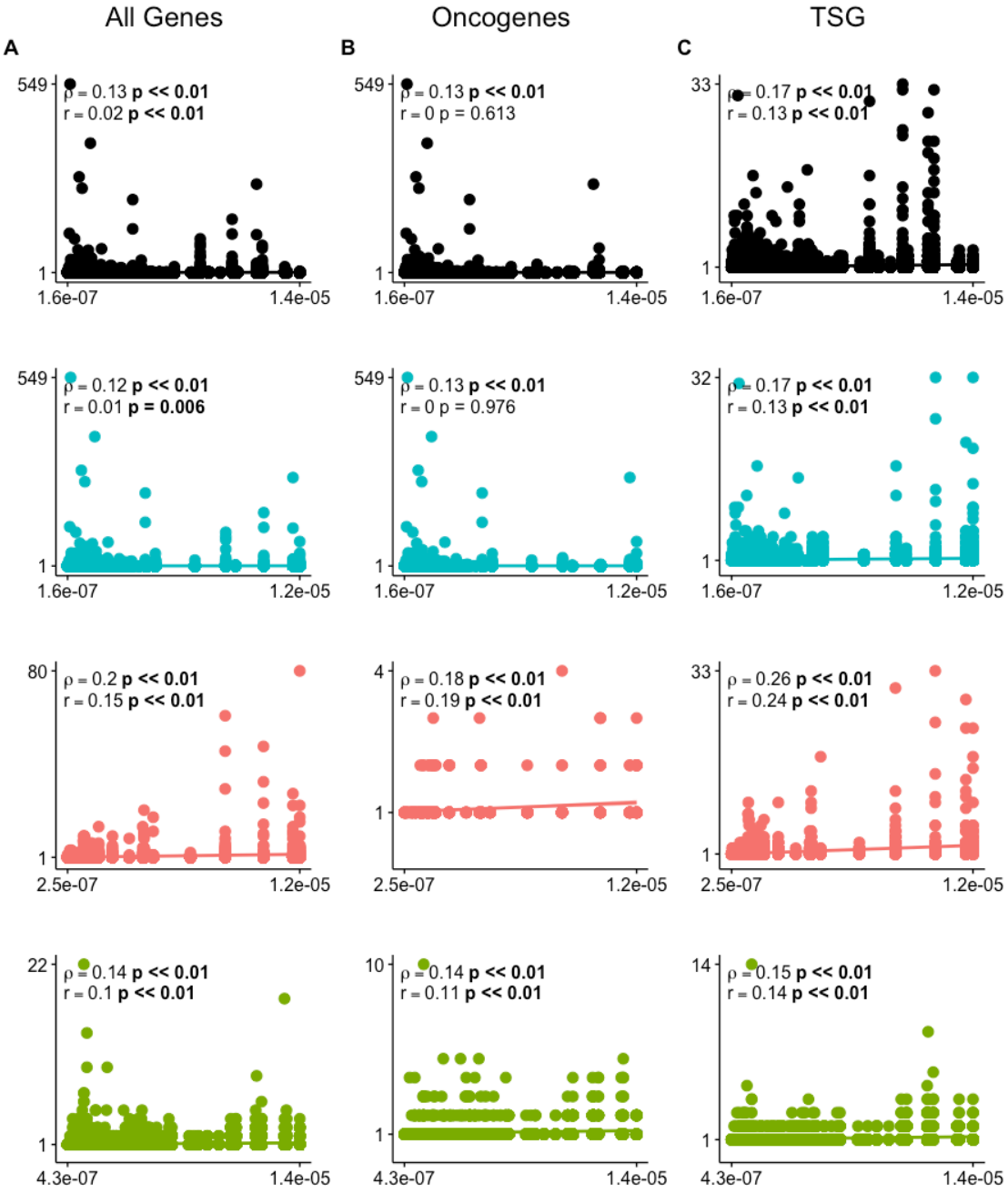



**Figure S5.** Relationship between codon mutability and observed mutation frequency by mutation type and role in cancer. **(A)** Scatterplot showing the observed mutation frequency for all genes and mutation types. **(B)** Scatterplots for oncogenes (n = 202) and **(C)** TSG (n = 166) for all mutation types. Different colors show scatterplots broken down by mutation type: missense (blue), nonsense (red) and silent (green). Spearman and Pearson correlation coefficient with respective p-values shown in all, significant at $p < 0.01$ in bold.

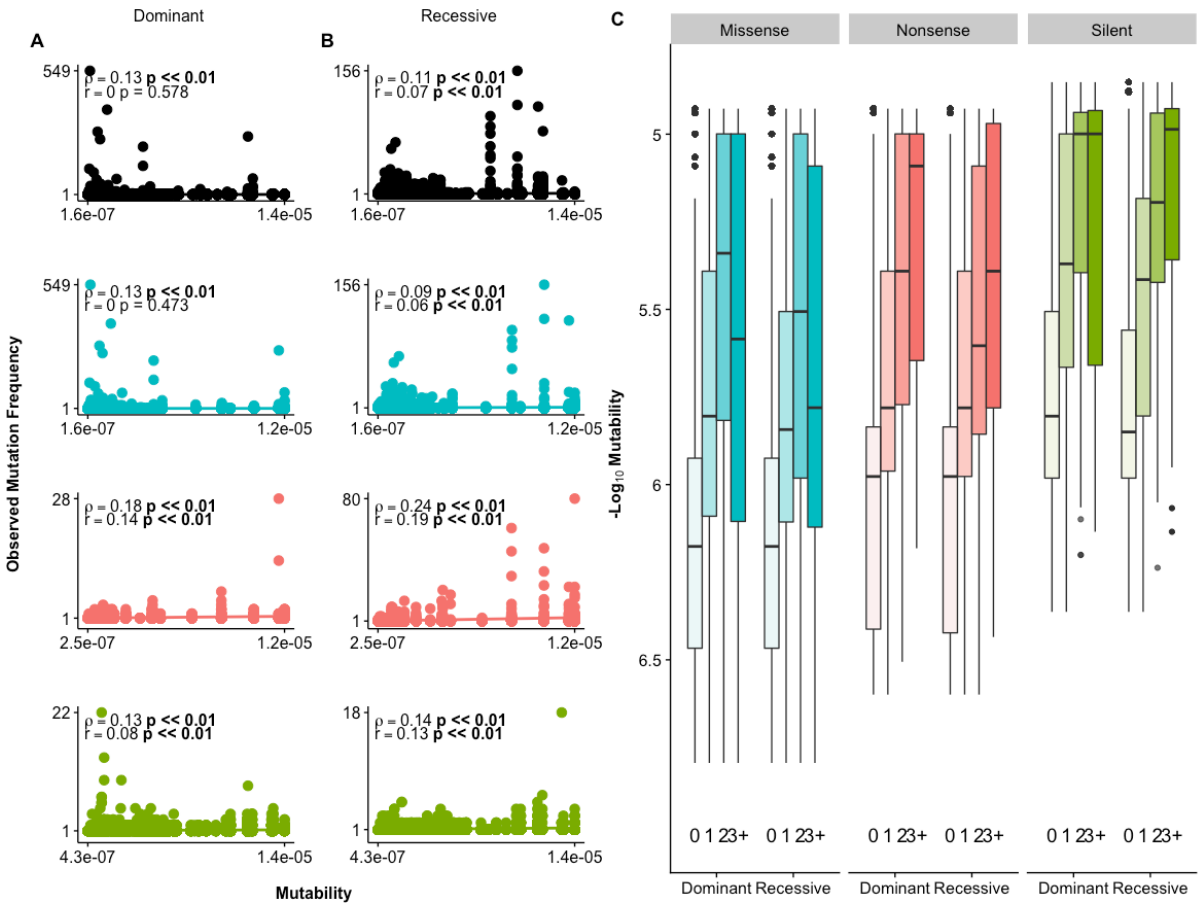



**Figure S6.** Relationship between codon mutability and observed mutation frequency by mutation type and molecular genetics. **(A)** Genes with only dominant mutations, **(B)** Genes with only recessive mutations. Different colors show scatterplots broken down by mutation type: missense (blue), nonsense (red) and silent (green). **(C)** As in Figure 4, pooled mutations in cancer census genes grouped by Dominant and Recessive mutations. Spearman and Pearson correlation coefficient with respective p-values shown in all, significant at $p < 0.01$ in bold. Counts summarized in Table S1.

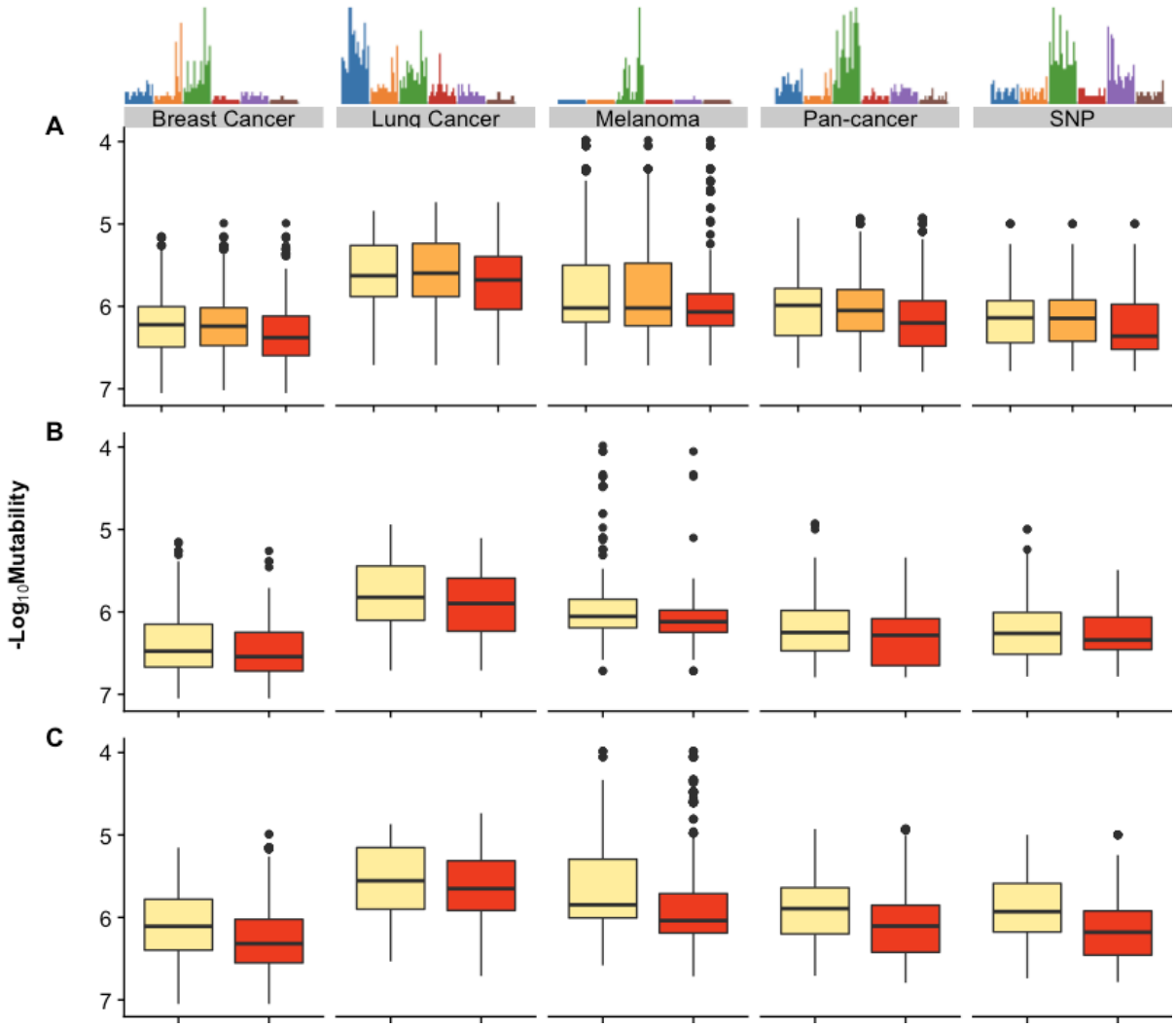



**Figure S7.** Mutability of missense mutations calculated with different background models. **(A)** TP53 **(B)** BRCA1 **(C)** Martelotto et al. datasets shown in Figure 4. Colors refer to functional annotation, yellow – Function/Neutral/Benign, orange – partially-functional, red – Non-functional/Non-neutral/Deleterious. Mutability values were calculated with background models of breast cancer, lung adenocarcinoma cancer, skin melanoma, and pan-cancer. Additionally, a background model for common non-pathogenic SNPs in general human population representing germline mutability was used. All models are available on the MutaGene website (https://www.ncbi.nlm.nih.gov/projects/mutagene/). Mutational profiles for each background model shown on the top. Significance tests summarized in Table S4.

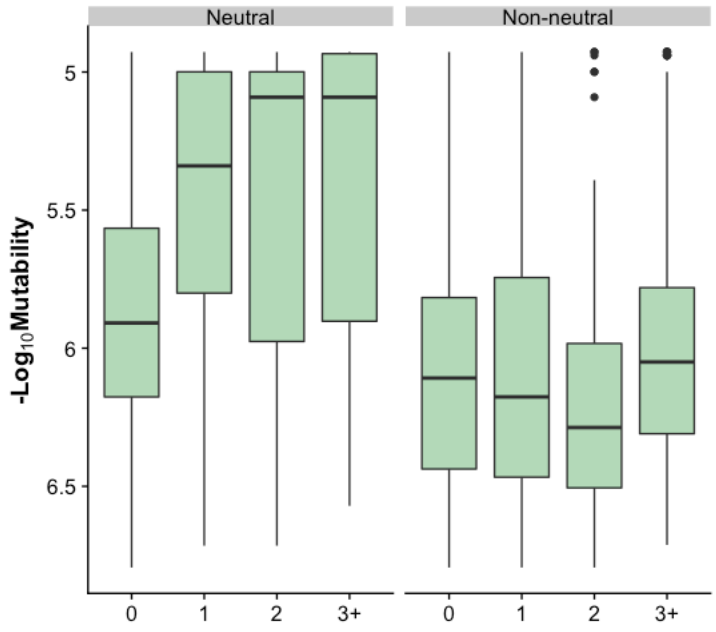



**Figure S8.** Codon mutability of missense mutations grouped by the effect on protein function and binned by mutation frequency in the MSK-IMPACT cohort.

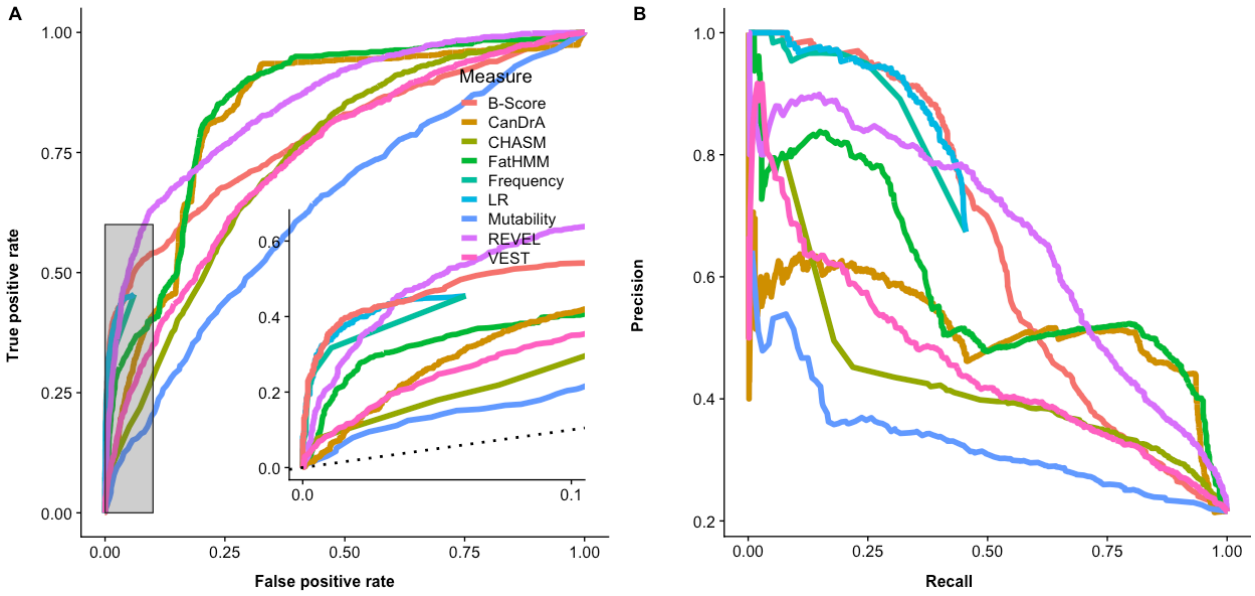



**Figure S9.** Assessment of classification performance between neutral and non-neutral mutations in a combined dataset. **(A)** ROC curves for B-Score, LR, and observed mutational frequency based on mutation frequency in COSMIC v85 cohort. Inset shows the performance of highlighted area corresponding to up to 10% FPR. **(B)** Precision-recall curves for the same benchmark set. The ROC for observed frequency and LR cannot be calculated for all mutations because some experimentally validated mutations were not observed in the COSMIC v85 cohort.

**Table S1.** Counts for boxplots in Figure 1 and Figure S2.

| | | Count group | | | |
|---|---|---|---|---|---|
| | | 0 (not observed) | 1 | 2 | 3+ |
| | | Codon mutations | | | |
| Mutation Type | Missense | 2660636 | 36019 | 2088 | 896 |
| | Nonsense | 142526 | 3667 | 362 | 181 |
| | Silent | 429864 | 12523 | 873 | 147 |
| | | Nucleotide mutations | | | |
| | Missense | 2978365 | 36131 | 2082 | 882 |
| | Nonsense | 168773 | 3719 | 350 | 178 |
| | Silent | 920579 | 12854 | 736 | 132 |
| | | Codon mutations – Oncogene | | | |
| | Missense | 811980 | 11884 | 738 | 344 |
| | Nonsense | 42575 | 793 | 53 | 9 |
| | Silent | 130163 | 4378 | 349 | 47 |
| | | Codon mutations – TSG | | | |
| | Missense | 1058490 | 14290 | 815 | 263 |
| | Nonsense | 56636 | 1987 | 232 | 125 |
| | Silent | 171368 | 4731 | 338 | 71 |
| | | Codon mutations – Dominant | | | |
| | Missense | 1743551 | 23492 | 1333 | 467 |
| | Nonsense | 94451 | 1750 | 115 | 31 |
| | Silent | 281044 | 8486 | 597 | 84 |
| | | Codon mutations – Recessive | | | |
| | Missense | 697206 | 9497 | 564 | 379 |
| | Nonsense | 37310 | 1615 | 223 | 146 |
| | Silent | 113365 | 2903 | 185 | 51 |
| | | Codon mutations – Dom/Rec | | | |
| | Missense | 57061 | 802 | 48 | 10 |
| | Nonsense | 2814 | 99 | 8 | 1 |
| | Silent | 9151 | 308 | 18 | 1 |
| | | Codon mutations – Rec/X | | | |
| | Missense | 5503 | 0 | 0 | 0 |
| | Nonsense | 304 | 0 | 0 | 0 |
| | Silent | 904 | 0 | 0 | 0 |
| | | Codon mutations – No Molecular Genetics Information Provided | | | |
| | Missense | 157315 | 2228 | 143 | 40 |
| | Nonsense | 7647 | 203 | 16 | 3 |
| | Silent | 25400 | 826 | 73 | 11 |

**Table S2.** Correlation between mutability and recurrence of mutations in cancer-associated genes
https://www.ncbi.nlm.nih.gov/research/mutagene/static/data/TableS2.xlsx

**Table S3.** Combined dataset with experimentally annotated neutral and non-neutral mutations in 58 genes
https://www.ncbi.nlm.nih.gov/research/mutagene/static/data/TableS3.xlsx

**Table S4.** Comparison of mutability on three experimental datasets with different cancer-specific background mutation models

| | **Model** | **P Value – Dunn Test** | **Comparison** |
|---|---|---|---|
| TP53 IARC | Pan-cancer | $2.17 \times 10^{-9}$ | functional – non-functional |
| | | 0.36 | functional – partially-functional |
| | | $1.41 \times 10^{-8}$ | non-functional – partially-functional |
| | Breast cancer | $1.84 \times 10^{-9}$ | functional – non-functional |
| | | 0.47 | functional – partially-functional |
| | | $6.11 \times 10^{-8}$ | non-functional – partially-functional |
| | Lung adenocarcinoma | $9.82 \times 10^{-5}$ | functional – non-functional |
| | | 0.23 | functional – partially-functional |
| | | $3.13 \times 10^{-5}$ | non-functional – partially-functional |
| | Skin melanoma | 0.005 | functional – non-functional |
| | | 0.51 | functional – partially-functional |
| | | 0.04 | non-functional – partially-functional |
| | SNP | $4.90 \times 10^{-5}$ | functional – non-functional |
| | | 0.42 | functional – partially-functional |
| | | 0.00013 | non-functional – partially-functional |
| Martelotto et. al | | **P-Value – Mann-Whitney-Wilcoxon** | |
| | Pan-cancer | $6.24 \times 10^{-8}$ | |
| | Breast cancer | $9.06 \times 10^{-7}$ | |
| | Lung adenocarcinoma | 0.02 | |
| | Skin melanoma | $5.45 \times 10^{-10}$ | |
| | SNP | $8.90 \times 10^{-12}$ | |
| BRCA1 – DMS | Pan-cancer | 0.04 | |
| | Breast cancer | 0.07 | |
| | Lung adenocarcinoma | 0.12 | |
| | Skin melanoma | 0.29 | |
| | SNP | 0.009 | |

**Table S5.** Performance of different classifiers on rarely observed and unobserved mutations. Classifiers are sorted by their maximum Matthews correlation.

| Measure | AUC-ROC | AUC-PR | Matthews correlation | Sensitivity at 10% FPR | Observed recurrence frequency in COSMIC Cohort |
|---|---|---|---|---|---|
| REVEL | 0.82 | 0.50 | 0.45 | 0.56 | 0 |
| CanDrAplus | 0.81 | 0.35 | 0.43 | 0.32 | |
| FatHMM | 0.81 | 0.38 | 0.42 | 0.29 | |
| VEST | 0.69 | 0.28 | 0.21 | 0.28 | |
| CHASM | 0.70 | 0.28 | 0.21 | 0.26 | |
| B-Score | 0.65 | 0.24 | 0.17 | 0.22 | |
| REVEL | 0.82 | 0.79 | 0.55 | 0.61 | 1 |
| CanDrAplus | 0.78 | 0.73 | 0.53 | 0.41 | |
| FatHMM | 0.81 | 0.75 | 0.51 | 0.44 | |
| B-Score | 0.79 | 0.73 | 0.46 | 0.45 | |
| VEST | 0.75 | 0.69 | 0.42 | 0.41 | |
| CHASM | 0.72 | 0.64 | 0.36 | 0.24 | |

**Table S6.** Performance metrics for the binomial model on combined and validation datasets using different values as a gene weight.

| Gene Weight | AUC-ROC | AUC-PR | MCC | Sensitivity at 10% error |
|---|---|---|---|---|
| SNP – 10,000 bp | 0.803 | 0.658 | 0.524 | 0.542 |
| SNP – 20,000 bp | 0.802 | 0.657 | 0.524 | 0.543 |
| SNP – 100,000 bp | 0.804 | 0.659 | 0.523 | 0.549 |
| SNP – 200,000 bp | 0.805 | 0.659 | 0.523 | 0.550 |
| No-outlier based weight | 0.677 | 0.577 | 0.479 | 0.489 |
| N Mutated Sites - based weight | 0.706 | 0.59 | 0.502 | 0.487 |
| No Weight | 0.785 | 0.653 | 0.527 | 0.541 |
| Silent mutation-based weight | 0.803 | 0.61 | 0.512 | 0.539 |

SNP, gene weighted by the number of SNPs in a window of various base pairs, window size indicated; No-outlier-based weight, gene weight as described by Chang et. al, see Methods for description. N Mutated Sites, gene weighted by the number of mutated sites in the gene over the total number of sites in the gene; No Weight, genes given no weighting; Silent mutations, gene weighted by number of silent mutations

## References

1. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A. 2010;107(3):961-8. Epub 2010/01/19. doi: 10.1073/pnas.0912629107. PubMed PMID: 20080596; PubMed Central PMCID: PMC2824313.
2. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446(7132):153-8. Epub 2007/03/09. doi: 10.1038/nature05610. PubMed PMID: 17344846; PubMed Central PMCID: PMC2712719.
3. Leedham S, Tomlinson I. The continuum model of selection in human tumors: general paradigm or niche product? Cancer research. 2012;72(13):3131-4. Epub 2012/05/04. doi: 10.1158/0008-5472.CAN-12-1052. PubMed PMID: 22552286.
4. Nussinov R, Tsai CJ. 'Latent drivers' expand the cancer mutational landscape. Curr Opin Struct Biol. 2015;32:25-32. Epub 2015/02/11. doi: 10.1016/j.sbi.2015.01.004. PubMed PMID: 25661093.
5. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018;15(2):81-94. Epub 2017/11/09. doi: 10.1038/nrclinonc.2017.166. PubMed PMID: 29115304.
6. Hiley C, de Bruin EC, McGranahan N, Swanton C. Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine. Genome Biol. 2014;15(8):453. Epub 2014/09/16. doi: 10.1186/s13059-014-0453-8. PubMed PMID: 25222836; PubMed Central PMCID: PMCPMC4281956.
7. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979-93. Epub 2012/05/23. doi: 10.1016/j.cell.2012.04.024. PubMed PMID: 22608084; PubMed Central PMCID: PMCPMC3414841.
8. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21. Epub 2013/08/16. doi: 10.1038/nature12477. PubMed PMID: 23945592; PubMed Central PMCID: PMC3776390.
9. Rogozin IB, Pavlov YI, Goncearenco A, De S, Lada AG, Poliakov E, et al. Mutational signatures and mutable motifs in cancer genomes. Briefings in bioinformatics. 2017. Epub 2017/05/13. doi: 10.1093/bib/bbx049. PubMed PMID: 28498882.
10. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151(7):1431-42. Epub 2012/12/25. doi: 10.1016/j.cell.2012.11.019. PubMed PMID: 23260136; PubMed Central PMCID: PMC3712641.
11. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nature reviews Genetics. 2011;12(11):756-66. Epub 2011/10/05. doi: 10.1038/nrg3098. PubMed PMID: 21969038.
12. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nature reviews Genetics. 2016;17(11):704-14. Epub 2016/10/16. doi: 10.1038/nrg.2016.104. PubMed PMID: 27739533.
13. Stratton MR. Exploring the genomes of cancer cells: progress and promise. Science. 2011;331(6024):1553-8. Epub 2011/03/26. doi: 10.1126/science.1204040. PubMed PMID: 21436442.
14. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence M, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518(7539):360-4. Epub 2015/02/20. doi: 10.1038/nature14221. PubMed PMID: 25693567; PubMed Central PMCID: PMC4405175.
15. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochimica et biophysica acta. 1992;1171(1):11-8. Epub 1992/11/15. PubMed PMID: 1420357.

16. Chen C, Qi H, Shen Y, Pickrell J, Przeworski M. Contrasting Determinants of Mutation Rates in Germline and Soma. Genetics. 2017;207(1):255-67. Epub 2017/07/25. doi: 10.1534/genetics.117.1114. PubMed PMID: 28733365; PubMed Central PMCID: PMC5586376.
17. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic acids research. 2012;40(21):e169. Epub 2012/08/21. doi: 10.1093/nar/gks743. PubMed PMID: 22904074; PubMed Central PMCID: PMC3505979.
18. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171(5):1029-41 e21. Epub 2017/10/24. doi: 10.1016/j.cell.2017.09.042. PubMed PMID: 29056346; PubMed Central PMCID: PMC5720395.
19. Araya CL, Cenik C, Reuter JA, Kiss G, Pande VS, Snyder MP, et al. Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nature genetics. 2016;48(2):117-25. Epub 2015/12/23. doi: 10.1038/ng.3471. PubMed PMID: 26691984; PubMed Central PMCID: PMC4731297.
20. Peterson TA, Gauran IIM, Park J, Park D, Kann MG. Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. PLoS computational biology. 2017;13(4):e1005428. Epub 2017/04/21. doi: 10.1371/journal.pcbi.1005428. PubMed PMID: 28426665; PubMed Central PMCID: PMC5398485.
21. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nature biotechnology. 2016;34(2):155-63. Epub 2015/12/01. doi: 10.1038/nbt.3391. PubMed PMID: 26619011; PubMed Central PMCID: PMC4744099.
22. Porta-Pardo E, Kamburov A, Tamborero D, Pons T, Grases D, Valencia A, et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nature methods. 2017;14(8):782-8. Epub 2017/07/18. doi: 10.1038/nmeth.4364. PubMed PMID: 28714987.
23. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27(15):2147-8. Epub 2011/06/21. doi: 10.1093/bioinformatics/btr357. PubMed PMID: 21685053; PubMed Central PMCID: PMC3137226.
24. Li M, Kales SC, Ma K, Shoemaker BA, Crespo-Barreto J, Cangelosi AL, et al. Balancing Protein Stability and Activity in Cancer: A New Approach for Identifying Driver Mutations Affecting CBL Ubiquitin Ligase Activation. Cancer research. 2016;76(3):561-71. Epub 2015/12/18. doi: 10.1158/0008-5472.CAN-14-3812. PubMed PMID: 26676746; PubMed Central PMCID: PMCPMC4738050.
25. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell. 2017;171(5):1042-56 e10. Epub 2017/10/24. doi: 10.1016/j.cell.2017.09.048. PubMed PMID: 29056344; PubMed Central PMCID: PMCPMC5849393.
26. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173(2):371-85 e18. Epub 2018/04/07. doi: 10.1016/j.cell.2018.02.060. PubMed PMID: 29625053.
27. Molina-Vila MA, Nabau-Moreto N, Tornador C, Sabnis AJ, Rosell R, Estivill X, et al. Activating mutations cluster in the "molecular brake" regions of protein kinases and do not associate with conserved or catalytic residues. Hum Mutat. 2014;35(3):318-28. Epub 2013/12/11. doi: 10.1002/humu.22493. PubMed PMID: 24323975.
28. Schaafsma GCP, Vihinen M. Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. Hum Mutat. 2017;38(7):839-48. Epub 2017/04/27. doi: 10.1002/humu.23236. PubMed PMID: 28444810.
29. Stehr H, Jang SH, Duarte JM, Wierling C, Lehrach H, Lappe M, et al. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. Mol Cancer. 2011;10:54. Epub

2011/05/18. doi: 10.1186/1476-4598-10-54. PubMed PMID: 21575214; PubMed Central PMCID: PMCPMC3123651.

30. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer research. 2009;69(16):6660-7. Epub 2009/08/06. doi: 10.1158/0008-5472.CAN-09-1133. PubMed PMID: 19654296; PubMed Central PMCID: PMCPMC2763410.

31. Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, et al. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat. 2016;37(1):28-35. Epub 2015/10/08. doi: 10.1002/humu.22911. PubMed PMID: 26442818; PubMed Central PMCID: PMCPMC5057310.

32. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-85. Epub 2016/09/27. doi: 10.1016/j.ajhg.2016.08.016. PubMed PMID: 27666373; PubMed Central PMCID: PMCPMC5065685.

33. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLoS One. 2013;8(10):e77945. Epub 2013/11/10. doi: 10.1371/journal.pone.0077945. PubMed PMID: 24205039; PubMed Central PMCID: PMCPMC3813554.

34. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65. Epub 2012/10/04. doi: 10.1002/humu.22225. PubMed PMID: 23033316; PubMed Central PMCID: PMC3558800.

35. Chang MT, Bhattarai TS, Schram AM, Bielski CM, Donoghue MTA, Jonsson P, et al. Accelerating Discovery of Functional Mutant Alleles in Cancer. Cancer Discov. 2018;8(2):174-83. Epub 2017/12/17. doi: 10.1158/2159-8290.CD-17-0321. PubMed PMID: 29247016; PubMed Central PMCID: PMCPMC5809279.

36. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8. Epub 2013/06/19. doi: 10.1038/nature12213. PubMed PMID: 23770567; PubMed Central PMCID: PMCPMC3919509.

37. Gorlov IP, Gorlova OY, Amos CI. Relative effects of mutability and selection on single nucleotide polymorphisms in transcribed regions of the human genome. BMC Genomics. 2008;9:292. Epub 2008/06/19. doi: 10.1186/1471-2164-9-292. PubMed PMID: 18559102; PubMed Central PMCID: PMCPMC2442617.

38. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014;156(6):1324-35. Epub 2014/03/19. doi: 10.1016/j.cell.2014.01.051. PubMed PMID: 24630730.

39. Temko D, Tomlinson IPM, Severini S, Schuster-Bockler B, Graham TA. The effects of mutational processes and selection on driver mutations across cancer types. Nat Commun. 2018;9(1):1857. Epub 2018/05/12. doi: 10.1038/s41467-018-04208-6. PubMed PMID: 29748584; PubMed Central PMCID: PMCPMC5945620.

40. Poulos RC, Wong YT, Ryan R, Pang H, Wong JWH. Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. PLoS Genet. 2018;14(11):e1007779. Epub 2018/11/10. doi: 10.1371/journal.pgen.1007779. PubMed PMID: 30412573.

41. Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nat Methods. 2016;13(10):806-7. Epub 2016/09/30. doi: 10.1038/nmeth.4000. PubMed PMID: 27684579; PubMed Central PMCID: PMCPMC5317181.

42. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database

issue):D980-5. Epub 2013/11/16. doi: 10.1093/nar/gkt1113. PubMed PMID: 24234437; PubMed Central PMCID: PMCPMC3965032.

43. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat. 2002;19(6):607-14. Epub 2002/05/15. doi: 10.1002/humu.10081. PubMed PMID: 12007217.

44. Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol. 2014;15(10):484. Epub 2014/10/29. doi: 10.1186/s13059-014-0484-1. PubMed PMID: 25348012; PubMed Central PMCID: PMC4232638.

45. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics. 2015;200(2):413-22. Epub 2015/04/01. doi: 10.1534/genetics.115.175802. PubMed PMID: 25823446; PubMed Central PMCID: PMC4492368.

46. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics. 2017;11(1):10. Epub 2017/05/18. doi: 10.1186/s40246-017-0104-8. PubMed PMID: 28511696; PubMed Central PMCID: PMC5433009.

47. Ng PK, Li J, Jeong KJ, Shao S, Chen H, Tsang YH, et al. Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell. 2018;33(3):450-62 e10. Epub 2018/03/14. doi: 10.1016/j.ccell.2018.01.021. PubMed PMID: 29533785; PubMed Central PMCID: PMCPMC5926201.

48. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1):D777-D83. Epub 2016/12/03. doi: 10.1093/nar/gkw1121. PubMed PMID: 27899578; PubMed Central PMCID: PMC5210583.

49. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1. Epub 2013/04/04. doi: 10.1126/scisignal.2004088. PubMed PMID: 23550210; PubMed Central PMCID: PMCPMC4160307.

50. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. Epub 2012/03/31. doi: 10.1038/nature11003. PubMed PMID: 22460905; PubMed Central PMCID: PMCPMC3320027.

51. Goncearenco A, Rager SL, Li M, Sang QX, Rogozin IB, Panchenko AR. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic acids research. 2017;45(W1):W514-W22. Epub 2017/05/05. doi: 10.1093/nar/gkx367. PubMed PMID: 28472504; PubMed Central PMCID: PMCPMC5793731.

52. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11. Epub 2000/01/11. PubMed PMID: 11125122; PubMed Central PMCID: PMC29783.

53. Evans P, Avey S, Kong Y, Krauthammer M. Adjusting for background mutation frequency biases improves the identification of cancer driver genes. IEEE Trans Nanobioscience. 2013;12(3):150-7. Epub 2013/05/23. doi: 10.1109/TNB.2013.2263391. PubMed PMID: 23694700; PubMed Central PMCID: PMC3989533.

54. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235-41. Epub 2015/11/12. doi: 10.1002/humu.22932. PubMed PMID: 26555599; PubMed Central PMCID: PMCPMC4752381.

55. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, et al. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics. 2013;29(5):647-8. Epub 2013/01/18. doi: 10.1093/bioinformatics/btt017. PubMed PMID: 23325621; PubMed Central PMCID: PMCPMC3582272.