Nucleotide and codon background mutability shape cancer mutational spectrum and advance driver mutation identification

Minghui Li[1,#], Anna-Leigh Brown[2,#], Alexander Goncearenco[2*], 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

To decipher the contribution of background DNA mutability in shaping observed mutational spectrum in cancer, we developed probabilistic models to estimate DNA context-dependent background mutability per nucleotide and codon substitution as a result of mutagenic processes. We showed that observed frequency of cancer mutations follows expected mutability, especially for tumor suppressor genes. In oncogenes, highly recurring mutations were characterized by a lower mutability, showing a U-shape trend. Mutations not occurring in tumors were shown to have lower mutability than observed mutations, indicating that DNA mutability could be a limiting step in mutation occurrence. To check whether mutability helps to discriminate driver mutations from passengers, we compiled a set of missense mutations with experimentally validated functional and transforming impacts. Accounting for background mutability significantly improved performance of classification into cancer driver and passenger mutations compared to mutation recurrence alone, performing similar to state-of-the-art machine-learning methods, even though no training was involved.

Introduction

Cancer is driven by changes at the 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]. 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 [4,5].

Point DNA mutations can arise from various forms of infidelity of DNA replication and repair machinery, endogenous, and exogenous mutagens [6-8]. There is an intricate 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 [9,10] due to many causes including timing of replication, nucleosome positioning, histone modifications levels, and other factors operating on local and large scales [11-14]. Many studies support point mutation rate dependence on the local DNA sequence context for various types of germline and somatic mutations [8,10,12,15]. Moreover, local DNA sequence context has been identified as a dominant factor explaining the largest proportion of mutation rate variation in germline and soma [9,16]. 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

59 persistent lack of quantitative information on per-nucleotide background rates of cancer somatic
60 mutations in different gene sites in various cancer types and tissues. In this study we developed
61 context-dependent probabilistic models that estimate baseline mutability per nucleotide or
62 codon substitution in specific genomic sites. These models were applied to decipher the
63 contribution of background DNA mutability in the observed mutational spectrum in cancer for
64 missense, nonsense, and silent mutations. To complement this approach, we also analyzed
65 background mutability for mutations that were not observed in cancer patients. Finally, for
66 practical purposes, we compiled a manually curated set of cancer driver and neutral missense
67 mutations with experimentally validated impacts collected from multiple studies. This data set
68 along with the recently published data set of experimentally characterized cancer mutations[27]
69 was used to verify our conclusions about the properties of cancer driver and passenger
70 mutations. The method has been implemented and can be found online as part of the MutaGene
71 web-server: https://www.ncbi.nlm.nih.gov/research/mutagene/gene.
72
73 **Results**
74 **Mutations not observed in cancer patients have low mutability**
75 We analyzed all 3,241,017 theoretically possible codon substitutions that could have occurred in
76 520 cancer census genes and found that only about one percent of them were actually observed
77 in the surveyed 12,013 tumor samples (Table S1). As illustrated in Figure 1B and D, the average
78 mutability of codon substitutions not detected in cancer patients (μ = 1.30 x 10^{-6}) was found to
79 be three-fold lower compared to the mutability of observed codon substitutions for all types of
80 mutations (μ = 3.88 x 10^{-6}), Mann-Whitney-Wilcoxon test, $p < 0.01$. This finding holds true for
81 per-nucleotide mutability (μ = 1.03 x 10^{-6} versus μ = 3.34 x 10^{-6}, Mann-Whitney-Wilcoxon test,
82 $p < 0.01$) (Figures 1D and S1). Figure 2 shows cumulative and probability density distributions
83 of nucleotide mutability values for all observed and possible mutations in all cancer census genes
84 and for two genes in particular, CASP8 and TP53, as examples. While there are many more
85 possible mutations with low mutability values, mutations at the higher range of the mutability
86 spectrum tend to occur more often in cancer samples.
87

88 **Silent mutations have highest mutability**

89 Figures 1A,B show the distributions of codon mutability values for all possible missense,
90 nonsense, and silent mutations accessible by single nucleotide base substitutions in the protein-
91 coding sequences of 520 cancer census genes calculated with the pan-cancer background model.
92 As can be seen on this figure, codon mutability spans two orders of magnitude and silent
93 mutations have significantly higher average mutability values (μ = 5.68 x 10^{-6}) than nonsense (μ
94 = 3.44 x 10^{-6}), or missense mutations (μ = 3.29 x 10^{-6}) according to Kruskal-Wallis test ($p < 0.01$)
95 and Dunn's post hoc test ($p < 0.01$) for all comparisons. These differences in codon mutabilities
96 could be a reflection of the degeneracy of genetic code, where multiple silent nucleotide
97 substitutions in the same codon may increase its mutability (as illustrated in Figure S5).
98 Consistent with this explanation, for nucleotide mutability shown in Figure 1C the differences
99 between types of mutations became less pronounced. However, silent mutations were still
100 characterized by the highest mutability values. Since observed mutational frequency of silent
101 mutations depends on their mutability (Figures 1B,D), it is important to correct for mutability
102 while ranking silent mutations with respect to their driver status, which is difficult to assess
103 comprehensively. Some recurrent highly mutable silent mutations might not represent relevant
104 candidates of drivers whereas some low mutable mutations are predicted to be drivers by our
105 method (KDR gene p.Leu355=, NTRK1 gene p.Asn270= and others).

106

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

108 Under the null model of all mutations arising as a result of background mutational processes,
109 somatic mutations should accumulate with respect to their mutation rate and one would expect
110 a positive correlation between mutability and observed mutational frequency of individual
111 mutations. Indeed, as Figure 1 shows, this is clearly the case for silent and nonsense mutations.
112 To further investigate this relationship, we calculated Spearman's rank (a non-parametric test
113 taking into account that mutability is not normally distributed) and Pearson linear correlation
114 coefficients between codon mutability and frequencies of mutations observed in 12,301 whole-
115 exome (WES) and whole-genome (WGS) tumor samples. Pooling together all types of mutations
116 in cancer census genes resulted in a very low but significant positive correlation between codon

117 mutability and mutation frequency ($\rho = 0.13$ and $r = 0.02$ for Spearman and Pearson
118 respectively $p < 0.01$), as shown in Figure S2A.
119
120 Breaking up all codon changes into silent, nonsense and missense reveals higher correlations,
121 particularly for silent ($\rho = 0.14$, $r = 0.1$, $p < 0.01$) and nonsense ($\rho = 0.20$, $r = 0.15$, $p <$
122 0.01) mutations (Figure S2A). We calculated correlation coefficients for each gene with at least
123 ten unique mutations of each type: silent, nonsense, and missense (Figures 3). Overall, we found
124 84 and 137 genes with significant ($p < 0.01$) positive Spearman and Pearson correlations,
125 respectively, for at least one mutation type (Table S2). Among the genes with significant
126 correlations, 41 belonged to tumor suppressor genes, 28 were oncogenes, and 15 genes were
127 classified as either fusion genes or both oncogene and tumor suppressor. For some genes,
128 including TP53 (second column, Figure 3B) and tumor suppressor CASP8 (third column, Figure
129 3B), a strong linear relationship between mutability and observed mutations frequency ($R^2 >$
130 0.5) was observed.
131
132 **Relationship between mutability and observed frequency is different for tumor suppressor and**
133 **oncogenes**
134 The effects of cancer mutations on protein function with respect to their cancer transforming
135 ability, can drastically differ in tumor suppressor genes (TSG) and oncogenes, therefore we
136 performed our analysis separately for these two categories (Figure 1E). We used COSMIC gene
137 classification separating genes into tumor suppressors and oncogenes. Mutations in tumor
138 suppressors can cause cancer through the inactivation of their products, whereas mutations on
139 oncogenes may result in protein activation. In addition, we used COSMIC classification into genes
140 with dominant or recessive mutations but overall results were similar to the ones produced using
141 classification into TSG and oncogenes (Figure S3). The strongest correlation between codon
142 mutability and mutation frequency was observed for TSG ($\rho = 0.17$, $r = 0.13$, $p < 0.01$) while
143 oncogenes showed a weak Spearman correlation, and no significant Pearson correlation ($\rho =$
144 0.13, $p < 0.01$, $r = 0$, $p = 0.61$) (Figure S2).

145 A U-shape trend was detected for missense and silent mutations in oncogenes: mutations with
146 low observed frequency (one or two samples) had the highest codon mutability on average, but
147 highly recurring mutations (observed in three and more samples) were characterized by a lower
148 average mutability (Figure 1E). In the latter case, selection may be a more important factor than
149 background mutation rate explaining reoccurrence of these mutation. Indeed, oncogenes are
150 characterized by a few gain-of-function missense mutations that are rather specific events
151 occurring on particular sites, compared to deactivating events in TSG that can affect a much
152 broader range of sites. Functional conserved sites overall were found to be more frequently
153 mutated in oncogenes [25], although our analysis did not find a straightforward association
154 between site mutability and its evolutionary conservation (data not shown).
155
156 **Functional effects of cancer mutations are related to codon mutability**
157 We calculated mutability for missense mutations which in turn were categorized based on their
158 experimental effect on protein function, transforming effects, and other characteristics (see
159 Methods). In TP53 transactivation dataset, mutability was found to differ with respect to
160 experimental effects with "functional" and "partially functional" mutations having significantly
161 higher mutability than "non-functional" (damaging) mutations (Dunn's post hoc test, $p < 0.01$
162 for both comparisons) (Figure 4A). No significant difference was found between functional and
163 partially functional mutations. Consistent with this finding, "neutral" mutations in the Martelotto
164 *et al.*[26] dataset had significantly higher mutability than "non-neutral" mutations (Mann-Whitney-
165 Wilcoxon test, $p < 0.01$)(Figure 4B). The same but much weaker trend was observed for BRCA1
166 missense mutations which were classified according to their ability to engage in homology-
167 directed repair (HDR) (Mann-Whitney-Wilcoxon test, $p = 0.036$)(Figure 4C). For our combined
168 dataset, the mutability values of "neutral" mutations ($\mu = 1.35 \times 10^{-6}$) were significantly higher
169 (Mann-Whitney-Wilcoxon test, $p < 0.01$) than for non-neutral mutations ($\mu = 2.59 \times 10^{-6}$)
170 (Figure 4D). This is consistent with our finding of high mutability of silent mutations, if we assume
171 that vast majority of them are benign (Figure 1). These results are also corroborated (with the
172 exception of skin cancers) if we use cancer-specific context-dependent background mutation
173 models for TP53 on Martelotto *et al.* set, although for BRCA1 set the difference between

damaging and neutral mutations is only observed using pan-cancer model (Figure S4, Table S4). Similar to our combined dataset, activating mutations from the validation set (see methods) had a much lower background codon mutability ($\mu = 1.35 \times 10^{-6}$) compared to neutral mutations ($\mu = 3.77 \times 10^{-6}$) (Mann-Whitney-Wilcoxon test, $p < 0.01$).

Accounting for context-dependent mutability in distinguishing driver from passenger mutations

In the previous sections we explored the contribution of background mutational processes to understand 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 carcinogenic effects. Therefore, to discover driver mutations, we tested how well different methods differentiate between experimentally verified neutral and driver non-neutral mutations. For this purpose, we compiled a dataset from several experimental studies which included ten oncogenes, five tumor suppressor genes, and two genes which were classified as both oncogenes and tumor suppressors (Table S7). In addition, we validated our results on a recently published dataset of mutations in 50 genes [27].

We developed two measures which utilize background mutation rate. The first measure, log-ratio (LR, equation 4), was calculated as a logarithm of a 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 (mutability). We compared the performance of these two scores to Condel[28], a meta-predictor which incorporates several state-of-the-art computational approaches, and FatHMM[29], a method explicitly developed to predict cancer driver mutations which performed very well in the most recent assessment [30]. Furthermore, the performance was indirectly compared to 21 different methods on a validation set (see Methods).

202 Figure 5 shows the ROC and precision-recall performance on the combined dataset. As can be
203 seen in the inset of Figure 5A and Table 1, B-score had the best performance in the region of
204 driver mutation prediction with low false positive rate and highest Matthews correlation
205 coefficient (MCC), a performance measure known to be more robust on unbalanced datasets.
206 Condel and FatHMM achieved highest AUC-ROC and a spike in true positive rate was observed in
207 FatHMM score, as it approached TP53 true positives, due to an extreme weight assigned by
208 FatHMM to TP53 gene. Importantly, correction of observed frequency by codon mutability
209 provided a much better performance than frequency or mutability alone. Mutability alone
210 performed better than random with an AUC=0.67 emphasizing the fundamental quality of non-
211 neutral mutations in cancer: mutability of driver mutations is lower than the mutability of
212 passengers (Figure 4D). As an additional validation, we compared B-Score performance to a
213 recent functional dataset [27] on 255 "activating" mutations and 488 "neutral" mutations. A ROC
214 analysis showed that B-Score performed with an AUC-ROC of 0.76 putting its performance among
215 the top three methods tested by Ng *et al.* Importantly to note, observed frequency performed
216 unexpectedly well on this test set with an AUC-ROC of 0.72, outperforming 18 of the 21 tested
217 computational methods tested.
218
219 **Variability of mutation rates across genes**
220 We checked if incorporating large-scale factors, allowing mutations of the same type to have
221 different mutational probabilities in different genes, affected retrieval performance. Several
222 methods have been developed to estimate gene weights, which use the overall number of
223 mutations, number of silent mutations affecting a gene, or other factors. We implemented and
224 tested all these approaches (see Methods) and in addition estimated the gene weights based on
225 the number of SNPs in the vicinity of a gene. We used gene weights to adjust codon mutability
226 values and explored whether any of the gene weight models were helpful in distinguishing
227 between experimentally neutral and driver non-neutral mutations. Unexpectedly, using gene
228 weight as an adjustment for varying background mutational rates across genes did not affect
229 classification performance. Only a SNP based weight affected the AUC-ROC, but the gain was very
230 minimal and no gene weight affected MCC (Table S6). This result held true on the validation

231 dataset with mutations sampled from 50 genes. Importantly, certain gene weights decreased
232 classification accuracy; in Figure 5 results with no gene weights are reported. In addition, we
233 examined the effects of several large-scale confounding factors such as gene expression levels,
234 replication timing, and chromatin accessibility provided in the gene covariates files for
235 MutSigCV[31] and found that only "no-outlier"-based weight and "silent mutation"-based weight
236 significantly correlated with the gene expression levels ($r = 0.66, p = 0.004$ and $r = 0.65, p =$
237 0.004, respectively).
238
239 **Ranking of cancer point mutations in MutaGene**
240 MutaGene webserver provides a collection of cancer specific context-dependent mutational
241 profiles and allows to calculate nucleotide and codon mutability for missense, nonsense and
242 silent mutations for a given protein coding sequence and background mutagenesis model using
243 the "Analyze gene" option. Following the analysis presented in this study, we added options to
244 provide a ranking of mutations observed in cancer samples based on B-score or the multiple-
245 testing adjusted q-values. Using our combined benchmark (Figure 5), we calibrated two
246 thresholds: the first corresponds to the maximum MCC coefficient, and the second corresponds
247 to 10% false positive rate (FPR). Mutations with B-score below the first threshold are predicted
248 to be "cancer drivers", whereas mutations with scores in between two thresholds are predicted
249 to be "potential drivers". All mutations with scores above the second threshold are predicted as
250 "passengers". Importantly, calculations can be performed using pan-cancer cohort or any
251 particular cancer type, the latter would result in a cancer-specific ranking of mutations and could
252 be useful for identification of driver mutations in a particular type of cancer. An example of
253 prediction of driver mutations status for EGFR is shown in Figure 6.
254
255 **Discussion**
256 To understand what processes drive point mutation accumulation in cancer, we used DNA
257 context-dependent probabilistic models to estimate the baseline mutability for nucleotide
258 mutation or codon substitution in specific genomic sites. Passenger mutations, constituting the
259 majority of all observed mutations, may have largely neutral functional impact and are unlikely
260 to be under selection pressure. For passenger mutations one would expect that mutations with

261 lower DNA mutability would have lower observed mutational frequency and vice versa. In
262 accordance with this expectation, we detected a significant positive correlation between baseline
263 mutability, which is an estimate of per site mutation rate, and observed frequencies of mutations
264 in cancer patients. In accordance with this expectation, we also found that mutations not so far
265 observed in cancer patients had much lower expected background nucleotide and codon
266 mutability compared to the observed mutations. For some genes, the observed frequency of
267 occurrence of mutations can be predicted purely from their mutability. Outliers of this trend are
268 important for inferring mutations under selection. For instance, if a mutation with low expected
269 mutation rate is observed in multiple patients, it is suggestive of it being potentially important in
270 carcinogenesis.
271
272 In this respect, reoccurring synonymous mutations with low mutability may represent interesting
273 cases for further investigation of potential silent drivers. Mutability of synonymous mutations
274 was found to be the highest among other types of mutations; and observed mutational frequency
275 of silent mutations scales well with their mutability. However, silent mutations that reoccur in
276 three or more samples in oncogenes have lower mutability than those occurring in only two
277 samples, suggesting that selection may be driving the high reoccurrence of these mutations.
278 Overall, our method predicted 102 silent mutations in 64 out of 520 cancer-associated genes to
279 be drivers. Indeed, it has been previously shown that some synonymous mutations might be
280 under selection, and synonymous codon mutations can affect the speed and accuracy of
281 transcription and translation, protein folding rate, and splicing [32].
282
283 There are many computational methods that aim to detect driver genes and fewer methods
284 trying to rank mutations with respect to their potential carcinogenicity. As many new approaches
285 to address this issue have been developed [33] [30], it still remains an extremely difficult task and
286 many driver mutations, especially in oncogenes, are not annotated as high impact or disease
287 related [34].
288

289 This can be explained by the scarcity of per site mutational data in WGS and WES studies for
290 specific cancers, difficulty to distinguish selection signal from background, and limited data on
291 experimental verification of driver mutational status. Increasing data in WGS and WES studies for
292 specific cancers thus should help to resolve the lack of cancer-specific benchmarks. While our
293 current attempts at including large-scale factors affecting mutation rates into gene weight
294 models did not improve prediction significantly in a pan-cancer benchmark, these factors are
295 likely tissue and cancer type specific, and should be tested in cancer-specific benchmarks.
296
297 We provide here a method to rank mutations with respect to their likelihood of being drivers.
298 This allows us to break ties for mutations which have been observed in the same number of
299 patients. For example in the TP53 gene, p.Arg181Cys, p.Arg282Gln, and p.Arg282Pro all have
300 been observed in two out of 9,450 pan-cancer patients. However, p.Arg181Cys (mutability 1.00
301 x 10^{-5}) is considered a passenger mutation, p.Arg282Gln (mutability 8.11 x 10^{-6}) as a potential
302 driver, and p.Arg282Pro (mutability 4.90 x 10^{-7}) as a driver mutation. Indeed, p.Arg282Pro is
303 annotated as driver mutation in the combined dataset, and despite the low mutation frequency,
304 it is correctly classified as a driver by its B-Score. While there are methods based on hotspot
305 identification, our ranking of mutations based on B-Score cannot be directly compared to the list
306 of hotspots proposed previously[21]. Hotspots are defined for sites, whereas B-Score assesses
307 specific mutations, and different mutations from the same hotspot can be drivers or passengers.
308 For instance, TP53 Tyr236 site is annotated as a hotspot in [21,35], however p.Tyr236Phe mutation
309 is experimentally characterized as neutral in the IARC database.
310
311 Some efforts have been focused so far on developing a comprehensive set of cancer driver
312 mutations verified at the levels of functional assays or animal models [36-38]. However, existing sets
313 often contain predictions and very few neutral passenger mutations. The vast majority of
314 computational prediction methods rely on machine learning algorithms trained on mutations
315 from a few genes and/or on recurrent mutations as estimates of driver events, in many cases the
316 performance is evaluated on synthetic benchmarks. In this study we collected experimental
317 evidence on cancer mutations, overall 3,787 neutral and 880 driver mutations from 17 genes and

488 neutral and 255 activating mutations from 50 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 hypothesis.

We developed and tested two models, LR and B-Score, to adjust the number of reoccurrences of a mutation by its expected background mutability in order to find sites 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. B-Score integrates information about observed rate and total sample size, something not captured by an odds ratio alone. B-Score implements a binomial test as a probabilistic model to estimate the number of mutations expected purely from the context-dependent background nucleotide or codon mutability, given a certain sample size. 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 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, many of which were trained on existing datasets of mutations and included multiple features. 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.

Methods

Datasets of experimental functional assays and cancer mutations

347 Missense mutations for TP53 gene with experimentally determined functional transactivation
348 activities were obtained from IARC P53 database where they were classified as functional,
349 partially-functional, and non-functional[39]. The second dataset, referred hereinafter to as
350 "Martelotto *et al.*", contains experimental evidence collected from literature [26]. The
351 experimental evidence of impact of mutations included changes in enzymatic activity, response
352 to ligand binding, impacts on downstream pathways, an ability to transform human or murine
353 cells, tumor induction in vivo, or changes in the rates of progression-free or overall survival in
354 pre-clinical models. In "Martelotto *et al.*" dataset mutations were considered "damaging" if there
355 was literature evidence to support their impact on at least one of the above-mentioned
356 categories. Mutations with no significant impacts on the wild-type protein function were
357 classified as "neutral". Mutations with no reliable functional evidence were regarded as
358 "uncertain" and were not used in this study. The third dataset included experimentally verified
359 BRCA1 mutations and was originally collected by using deep mutational scanning to measure the
360 effects of missense mutations on the ability of BRCA1 to participate in homology-directed repair.
361 Missense mutations were categorized as either neutral or damaging [40,41]. Noteworthy, BRCA1 set
362 contained inherited germline as well as somatic mutations. The fourth dataset explored over
363 81,000 tumors to identify drivers of hypermutation in DNA polymerase epsilon and delta genes
364 (POLE/POLD1). Drivers of hypermutation were variants which occurred in a minimum of two
365 hypermutant tumors, which were never found in lowly mutated tumors, and did not co-occur
366 with an existing known driver mutation in the same tumor. Other variants in these genes were
367 considered passengers with respect to hypermutation [42].
368
369 Finally, we assembled a *combined dataset* that included mutations from these four datasets
370 described above. We removed redundant and conflicting entries when mutations annotated as
371 non-functional or neutral in one dataset were also annotated as damaging or benign in another.
372 All mutations were categorized as "non-neutral" (affecting function, binding or transforming) and
373 "neutral" (other mutations). We treated "functional" and "partially -functional" mutations in
374 IARC TP53 dataset as "neutral", and "non-functional" as "non-neutral". We used missense
375 mutations in order to compare with Condel and cancer FatHMM scores. Overall, the combined

376 dataset contains 4,667 mutations (3,787 neutral and 880 non-neutral) from 17 genes (Table S3,
377 S7).
378
379 As an additional validation on driver prediction performance, we used a recently obtained
380 dataset containing missense mutations which were annotated based on their effects on cell-
381 viability in Ba/FC and MCF10A models[27]. Activating mutations were annotated as those
382 mutations where the cell viability was higher than the wild-type gene, and neutral mutations
383 were those mutations where cell-viability was similar to the wild-type. Ng *et al.* used these
384 consensus functional annotations to compare the performance of 21 different computational
385 tools in classifying between activating and neutral mutations using ROC analysis, with activating
386 mutations acting as the positive set and neutral as the negative set. Validation dataset contained
387 743 mutations (488 neutral and 255 activating) from 50 genes.
388
389 The Catalogue of Somatic Mutations in Cancer (COSMIC) database stores data on somatic cancer
390 mutations and integrates the experimental data from the full-genome (WGS) and whole-exome
391 (WES) sequencing studies[43]. Cancer census genes were defined according to COSMIC release v84.
392 Genes were binned into those with oncogene and tumor suppressor (TSG) functions. For analyses
393 comparing oncogenes and TSG, genes classified as only fusion genes or those with both
394 oncogenic and TSG activities were not used. This resulted in 205 oncogenes and 167 TSG (Table
395 S2). A subset of 12,013 tumors from COSMIC release v85 was used to extract observed frequency
396 data. COSMIC v85 samples which came from cell-lines, xenografts, or organoid cultures were
397 excluded. Only mutations with somatic status of "Confirmed somatic variant" were included and
398 mutations which were flagged as SNPs were excluded. For each cancer patient, a single sample
399 from a single tumor was used. Additionally, it is possible that the same patient may be assigned
400 different unique identifiers in different papers, and duplicate tumor samples may be erroneously
401 added to COSMIC database during manual curation. These samples may affect the recurrence
402 counts of mutations and, consequently, the B-score. We applied clustering in order to detect and
403 remove any redundant tumor samples. Each sample was represented as a binary vector with 1 if

404 a sample had a mutation in a particular genomic location and 0 otherwise. The binary vectors
405 were compared with Jaccard distance metric,
406 $J = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$
407 where identical samples have $J = 0$, followed by agglomerative clustering with complete linkage.
408 Non-singleton clusters with pairwise distance cutoff of $J = 0.3$ were extracted and only one
409 representative of each cluster was used, whereas other samples were discarded.
410
411 ***Calculation of context-dependent DNA background mutability***
412 Context-dependent mutational profiles were constructed from the pools of mutations from
413 different cancer samples by counting mutations observed in a specific trinucleotide context.
414 Altogether, there are 64 different types of trinucleotides and three types of mutations $x \gg y$ (for
415 example C≫A, C≫T, C≫G and so on) in the central position of each trinucleotide which results
416 in 192 trinucleotide context-dependent mutation types. In a mutated double-stranded DNA both
417 complementary purine-pyrimidine nucleotides are substituted, and therefore we considered
418 only substitutions in pyrimidine (C or T) bases, resulting in 96 possible context-dependent
419 mutation types: m = {N [$x \gg y$] N}, where N = {A, C, G, T}. Thus, mutational profile can be
420 expressed as a vector of a number of mutations of certain type $(f_1, \dots, f_{96})$. Profiles were
421 constructed under the assumption that vast majority of cancer context-dependent mutations
422 have neutral effects, while only a negligible number of these mutations in specific sites are under
423 selection. To assure this, we removed recurrent mutations (observed twice or more times in the
424 same site) as these mutations might be under selection in cancer. In the current study we used
425 pan-cancer and cancer-specific mutational profiles for breast, lung adenocarcinoma, and skin
426 adenocarcinoma cancer derived from MutaGene [44].
427
428 We applied mutational profiles to build DNA context-dependent probabilistic models that
429 described baseline DNA mutagenesis per nucleotide or per codon. *Mutability* was defined as a
430 probability to observe a context-dependent nucleotide mutation purely from the baseline
431 mutagenic processes operating in a given group of samples used to derive the mutational profile.
432 Mutability is proportional to the expected mutation rate of a certain type of mutation (96

433 altogether) regardless of the genomic site it occurs. It was calculated using the total number
434 of mutations of type $x \gg y$ in a certain local context, which can be obtained from the context-
435 dependent mutational profile, f_m . Given the number of cancer samples used to construct
436 mutational profile, *N*, and the number of different trinucleotides of type *t* in a diploid human
437 exome, n_t (calculated from the reference genome), the mutability of a nucleotide mutation is
438 calculated as:

439 $$p_m^{nuc} = \frac{f_m}{N\, n_t} \quad (1)$$

440 In protein-coding sequences it is practical to calculate mutation probability for a codon in its local
441 pentanucleotide context, given trinucleotide contexts of each nucleotide in the codon. For a
442 given transcript of a protein, at exon boundaries the local context of the nucleotides was taken
443 from the genomic context. The COSMIC consensus transcript was chosen for the transcript for
444 each protein. Changes in codons can lead to amino acid substitutions, synonymous and nonsense
445 mutations. Therefore, we calculate a probability to observe a specific type of codon change which
446 can be realized by single nucleotide mutations at each codon position *i* as:
447

448 $$p_M^{codon} = 1 - \prod_i^3 \left(1 - \sum_j^k p_{ij}^{nuc}\right) \quad (2)$$

449

450 Where *k* denotes a number of mutually exclusive mutations at codon position *i*. For example, for
451 Phe codon "TTT" in a given context 5'-A and G-3': three single nucleotide mutations can lead to
452 Phe->Leu substitution (to codons "TTG", "TTA" and "CTT" for Leu): A[T≫C]TTG in the first codon
453 position or mutually exclusive ATT[T≫G]G and ATT[T>>A]G in the third codon position (Figure
454 S5). In this case the probability of Phe→ Leu substitution in the $ATTTG$ context can be calculated
455 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
456 were taken from the mutational profile. Amino acid substitutions corresponding to each
457 missense mutation are calculated by translating the mutated and wild type codons using a
458 standard codon table. Codon mutability strongly depends on the neighboring codons as
459 illustrated in Figure S5.
460

461 ***Gene-weight adjusted mutability***

462 Gene weights estimate a relative probability of a gene compared to other genes to be mutated
463 in cancer through somatic mutagenesis. There are multiple ways the gene weights can be
464 calculated.
465
466 *SNP-based weight* was calculated using the number of SNPs in the vicinity of the gene of interest.
467 We used the "EnsDb.Hsapiens.v86" database to find genomic coordinates of a gene, including
468 introns, and extended the range in both 3' and 5' directions according to the window size (Table
469 S6). We then counted the number of common SNPs from dbSNP database[45] within the genomic
470 region. Gene weight was calculated as: $\omega_g^{SNP} = \frac{n_{SNP}}{L_{window}}$, where n_{SNP} is the number of SNPs and
471 L_{window} is the length of the genomic region in base pairs. We tested several window sizes for
472 defining the genomic regions around the gene of interest (Table S6).
473
474 *Mutation-based weight* was calculated using the number of nucleotide sites with reoccurring
475 mutations counted only once to avoid the bias that may be present due to selection on individual
476 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
477 in the gene transcript.
478
479 *Silent mutation-based weight* was introduced previously and was shown to be superior in
480 assessment of significant non-synonymous mutations across genes[46]. This weight can be
481 calculated by taking into account only silent somatic mutations: $\omega_g^{silent} = \frac{s_g}{N\,L_g}$. Here s_g is the
482 total number of somatic silent mutations within the gene, N is the number of tumor samples
483 and L_g is the number of codons in the gene transcript.
484
485 *No-outlier-based weight* introduced previously[21] takes into account the number of all codon
486 mutations within a gene, C_g, excluding mutations in outlier codon sites bearing more than the
487 99th percentile of mutations of the gene: $\omega_g^{out} = \frac{C_g}{N\,L_g}$, normalized by gene length L_g in amino
488 acids and the total number of samples N.
489

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

491 $$p'_M = \omega_g p_M \quad (3)$$

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

493

494 ***Identification of significant mutations***

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

498 $$LR = \log\left(\frac{n_{obs}}{n_{exp}}\right) \quad (4)$$

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

504

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

507

508 $$B_{score} = \sum_{k=n+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} \quad (5)$$

509

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

516

517 ***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 2.2×10^{-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 (MCC):

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

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 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)

References

1. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).
2. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-8 (2007).

547 3. Leedham, S. & Tomlinson, I. The continuum model of selection in human tumors: general
548 paradigm or niche product? *Cancer Res* **72**, 3131-4 (2012).
549 4. Dagogo-Jack, I. & Shaw, A.T. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev*
550 *Clin Oncol* **15**, 81-94 (2018).
551 5. Hiley, C., de Bruin, E.C., McGranahan, N. & Swanton, C. Deciphering intratumor heterogeneity and
552 temporal acquisition of driver events to refine precision medicine. *Genome Biol* **15**, 453 (2014).
553 6. Salk, J.J., Fox, E.J. & Loeb, L.A. Mutational heterogeneity in human cancers: origin and
554 consequences. *Annu Rev Pathol* **5**, 51-75 (2010).
555 7. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21
556 (2013).
557 8. Rogozin, I.B. *et al.* Mutational signatures and mutable motifs in cancer genomes. *Brief Bioinform*
558 (2017).
559 9. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo
560 germline mutation. *Cell* **151**, 1431-42 (2012).
561 10. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat*
562 *Rev Genet* **12**, 756-66 (2011).
563 11. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**,
564 704-714 (2016).
565 12. Stratton, M.R. Exploring the genomes of cancer cells: progress and promise. *Science* **331**, 1553-8
566 (2011).
567 13. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional
568 mutation rates in human cancer cells. *Nature* **488**, 504-7 (2012).
569 14. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer.
570 *Nature* **518**, 360-364 (2015).
571 15. Rogozin, I.B. & Kolchanov, N.A. Somatic hypermutagenesis in immunoglobulin genes. II. Influence
572 of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta* **1171**, 11-8 (1992).
573 16. Chen, C., Qi, H., Shen, Y., Pickrell, J. & Przeworski, M. Contrasting Determinants of Mutation Rates
574 in Germline and Soma. *Genetics* **207**, 255-267 (2017).
575 17. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids*
576 *Res* **40**, e169 (2012).
577 18. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**,
578 1029-1041 e21 (2017).
579 19. Araya, C.L. *et al.* Identification of significantly mutated regions across cancer types highlights a
580 rich landscape of functional molecular alterations. *Nat Genet* **48**, 117-25 (2016).
581 20. Peterson, T.A., Gauran, I.I.M., Park, J., Park, D. & Kann, M.G. Oncodomains: A protein domain-
582 centric framework for analyzing rare variants in tumor samples. *PLoS Comput Biol* **13**, e1005428
583 (2017).
584 21. Chang, M.T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity
585 and mutational specificity. *Nat Biotechnol* **34**, 155-63 (2016).
586 22. Porta-Pardo, E. *et al.* Comparison of algorithms for the detection of cancer drivers at subgene
587 resolution. *Nat Methods* **14**, 782-788 (2017).
588 23. Wong, W.C. *et al.* CHASM and SNVBox: toolkit for detecting biologically important single
589 nucleotide mutations in cancer. *Bioinformatics* **27**, 2147-8 (2011).
590 24. Li, M. *et al.* Balancing Protein Stability and Activity in Cancer: A New Approach for Identifying
591 Driver Mutations Affecting CBL Ubiquitin Ligase Activation. *Cancer Res* **76**, 561-71 (2016).
592 25. Stehr, H. *et al.* The structural impact of cancer-associated missense mutations in oncogenes and
593 tumor suppressors. *Mol Cancer* **10**, 54 (2011).

26. Martelotto, L.G. *et al.* Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol* **15**, 484 (2014).
27. Ng, P.K. *et al.* Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell* **33**, 450-462 e10 (2018).
28. Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**, 440-9 (2011).
29. Shihab, H.A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65 (2013).
30. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e18 (2018).
31. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
32. Supek, F., Minana, B., Valcarcel, J., Gabaldon, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324-1335 (2014).
33. Li, M., Goncearenco, A. & Panchenko, A.R. Annotating Mutational Effects on Proteins and Protein Interactions: Designing Novel and Revisiting Existing Protocols. *Methods Mol Biol* **1550**, 235-260 (2017).
34. Molina-Vila, M.A. *et al.* Activating mutations cluster in the "molecular brake" regions of protein kinases and do not associate with conserved or catalytic residues. *Hum Mutat* **35**, 318-28 (2014).
35. Chang, M.T. *et al.* Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov* **8**, 174-183 (2018).
36. Ainscough, B.J. *et al.* DoCM: a database of curated mutations in cancer. *Nat Methods* **13**, 806-7 (2016).
37. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-5 (2014).
38. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* **10**, 25 (2018).
39. Olivier, M. *et al.* The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat* **19**, 607-14 (2002).
40. Starita, L.M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413-22 (2015).
41. Mahmood, K. *et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum Genomics* **11**, 10 (2017).
42. Campbell, B.B. *et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* **171**, 1042-1056 e10 (2017).
43. Forbes, S.A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**, D777-D783 (2017).
44. Goncearenco, A. *et al.* Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Res* **45**, W514-W522 (2017).
45. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
46. Evans, P., Avey, S., Kong, Y. & Krauthammer, M. Adjusting for background mutation frequency biases improves the identification of cancer driver genes. *IEEE Trans Nanobioscience* **12**, 150-7 (2013).

Figures and Tables

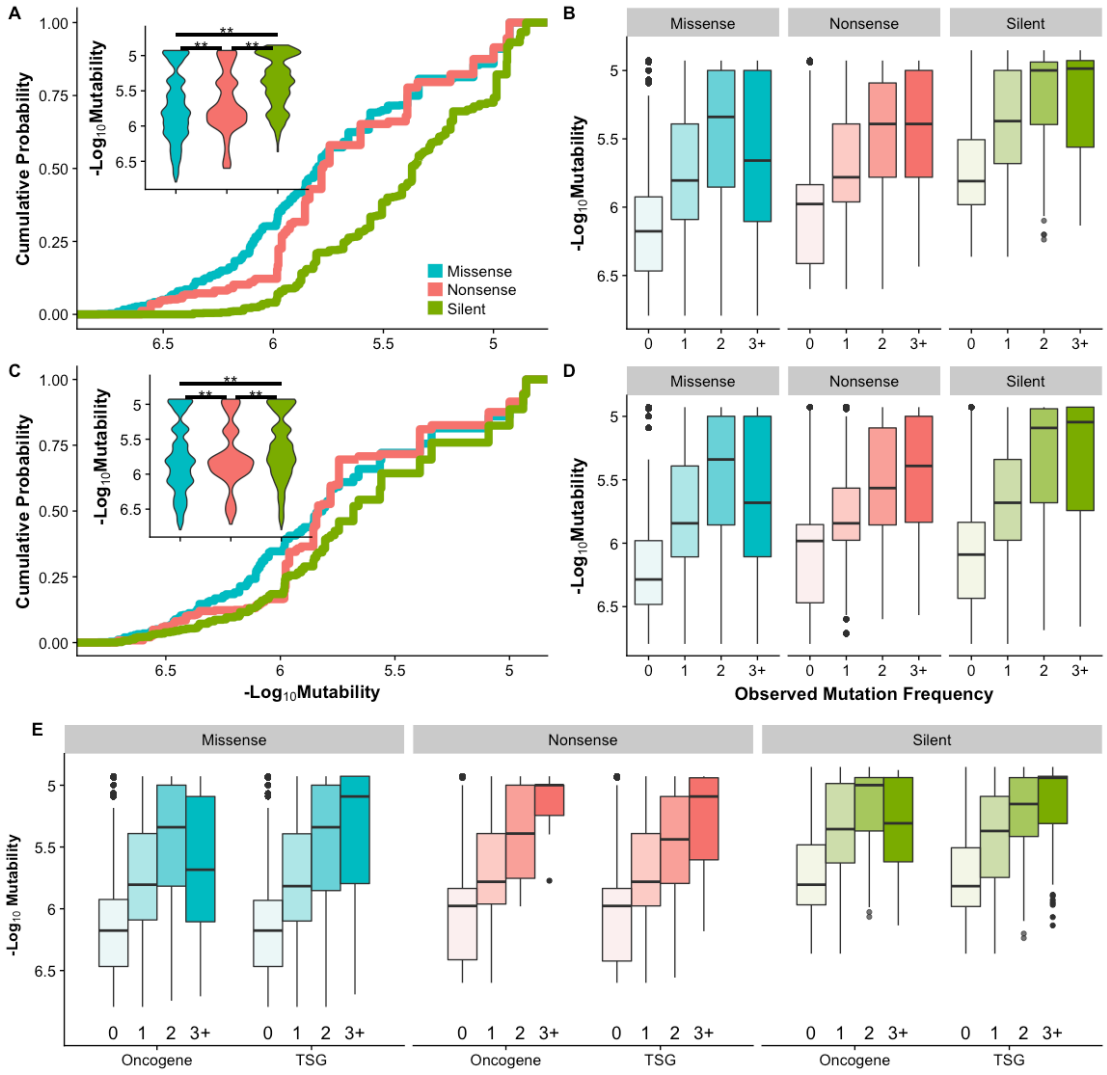


Figure 1. Mutability distributions by mutation type and gene role in cancer. **(A)** Cumulative distribution of codon mutability of silent (green color), nonsense (red color) and missense (blue color) mutations. **(C)** Cumulative distribution of nucleotide mutability for silent, nonsense and missense 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. '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. **(E)** 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.

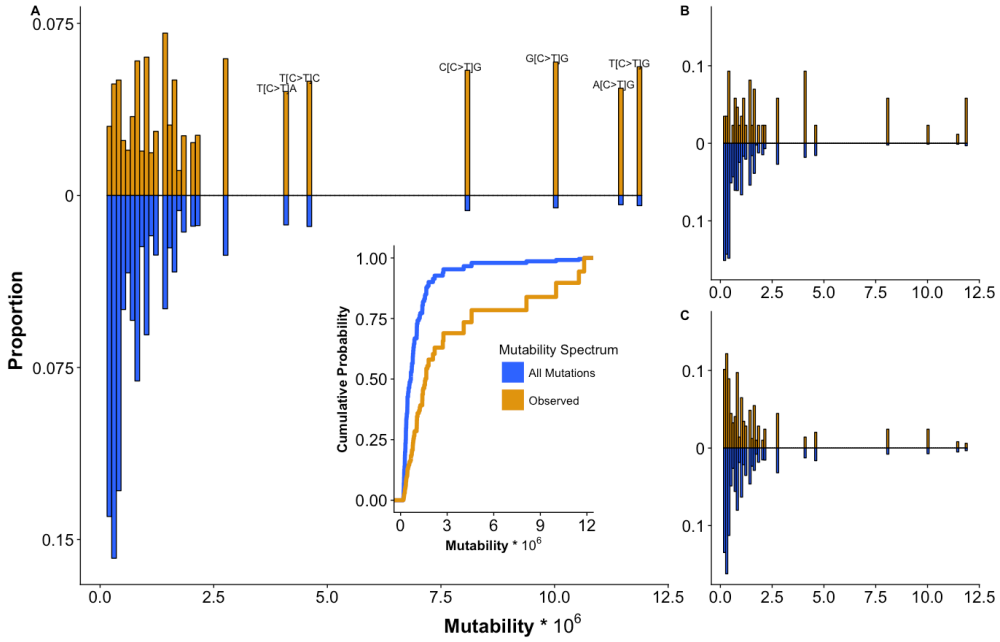


Figure 2. Comparison between nucleotide mutability (expected) spectrum of all possible mutations (blue) and mutations which were observed in cancer patients (brown). **(A)** Mutations in 520 cancer-related genes; **(B)** CASP8 and **(C)** TP53 genes. Inset shows the cumulative distribution functions for both spectra. Annotations in **(A)** show nucleotide mutation types.

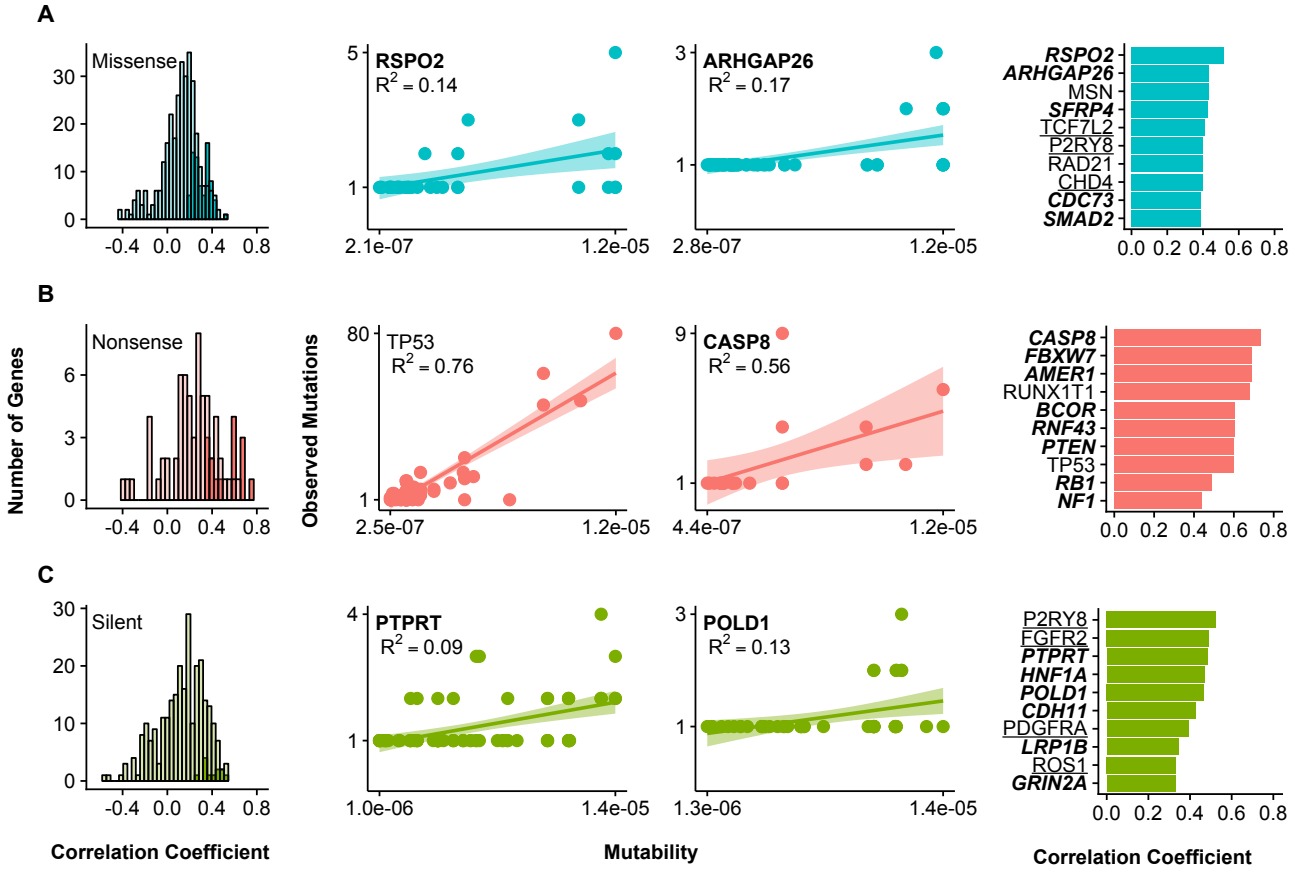


Figure 3. Relationship between codon mutability and frequency of mutations. Histograms show the Spearman rank correlation coefficients between observed mutations and mutability across cancer genes with at least 10 observed mutations of each type: **(A)** missense (blue), **(B)** nonsense (red) and **(C)** 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.

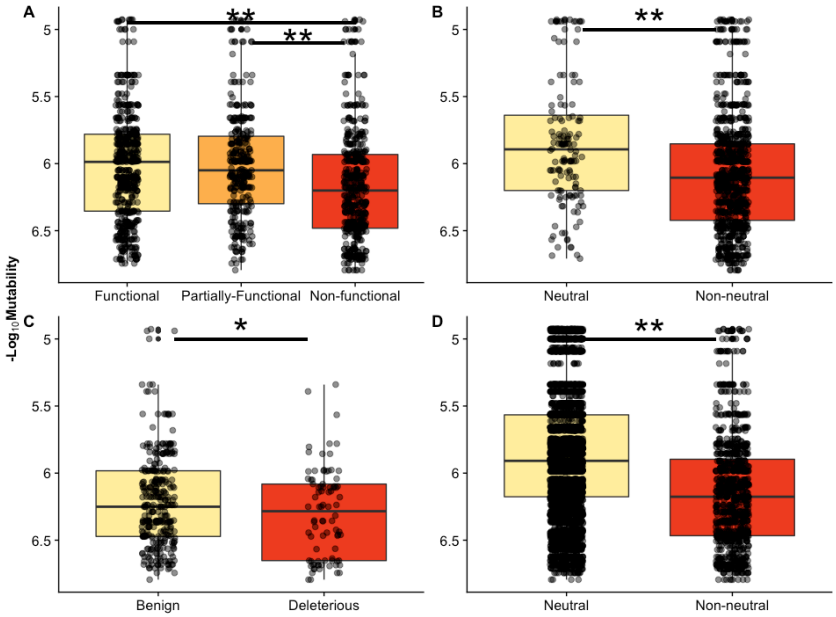


Figure 4. Codon mutability of missense mutations grouped by the effect on protein function. **(A)** Mutations in TP53 gene were characterized as functional, partially functional and non-functional. (B) Mutations from Martelotto *et al.* set were categorized as neutral and non-neutral. (C) Mutations in BRCA1 gene were annotated as benign and deleterious. (D) Mutations from the combined dataset were categorized as neutral and non-neutral. Significant differences with $p < 0.01$ are marked with a double asterisk and difference with $p < 0.05$ is marked with a single asterisk. Mutability was calculated with pan-cancer background model; see Supplementary Figure S4 for analysis with different background models.

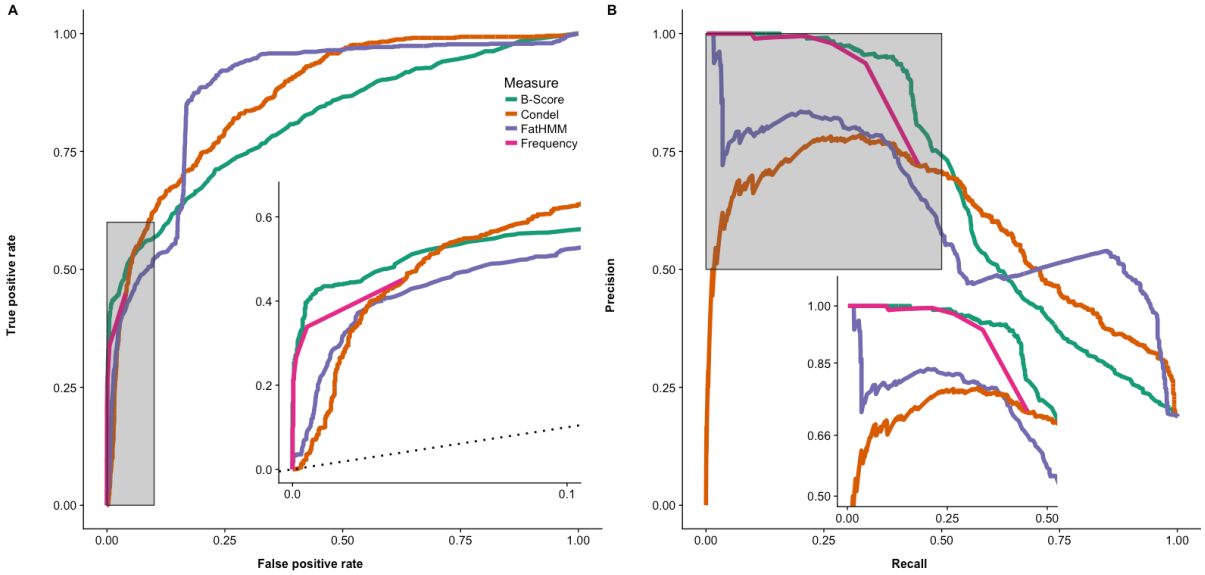
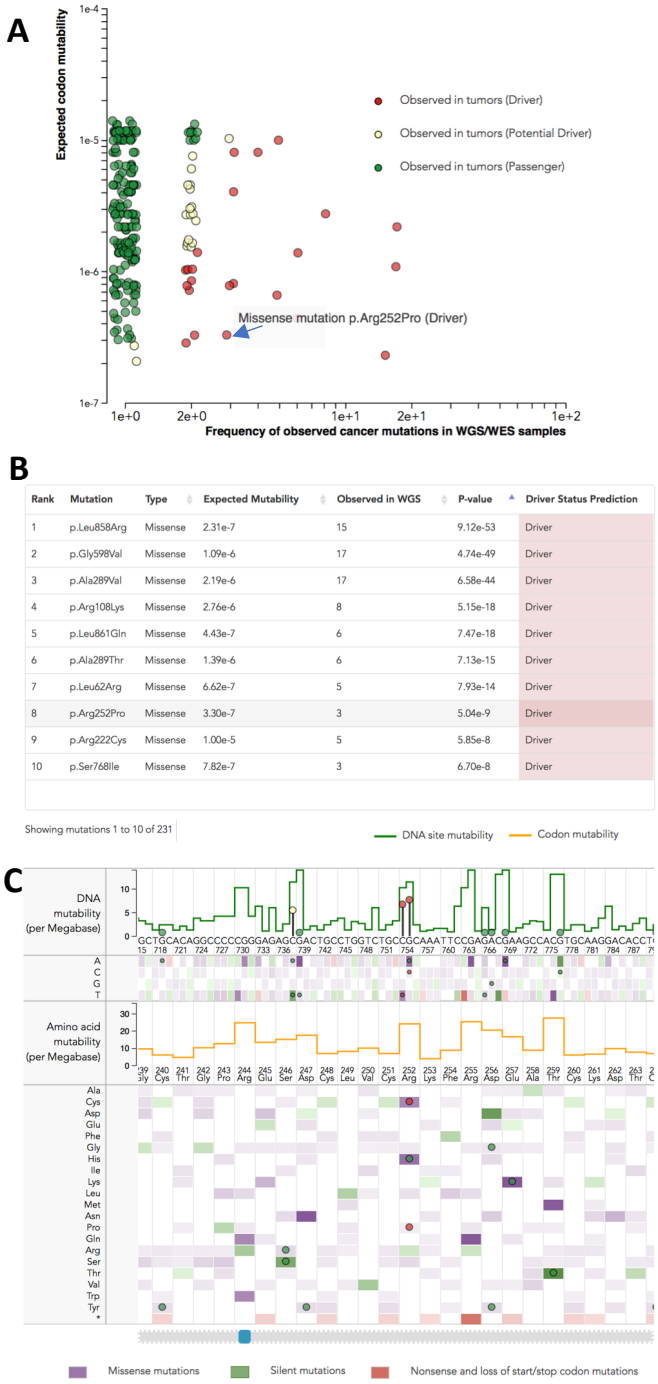


Figure 5. Assessment of classification performance between neutral and non-neutral mutations in a combined dataset. **(A)** ROC curves for B-Score, Condel, FatHMM, and observed mutational frequency. Inset shows the performance of highlighted area corresponding to up to 10% FPR. **(B)** Precision-recall curves for the same benchmark set. Inset shows the performance of highlighted area corresponding to precision over 50%. The ROC for observed frequency cannot be calculated for all mutations because some experimentally validated mutations were not observed in tumor samples.



Rank	Mutation	Type	Expected Mutability	Observed in WGS	P-value	Driver Status Prediction
1	p.Leu858Arg	Missense	2.31e-7	15	9.12e-53	Driver
2	p.Gly598Val	Missense	1.09e-6	17	4.74e-49	Driver
3	p.Ala289Val	Missense	2.19e-6	17	6.58e-44	Driver
4	p.Arg108Lys	Missense	2.76e-6	8	5.15e-18	Driver
5	p.Leu861Gln	Missense	4.43e-7	6	7.47e-18	Driver
6	p.Ala289Thr	Missense	1.39e-6	6	7.13e-15	Driver
7	p.Leu62Arg	Missense	6.62e-7	5	7.93e-14	Driver
8	p.Arg252Pro	Missense	3.30e-7	3	5.04e-9	Driver
9	p.Arg222Cys	Missense	1.00e-5	5	5.85e-8	Driver
10	p.Ser768Ile	Missense	7.82e-7	3	6.70e-8	Driver

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.

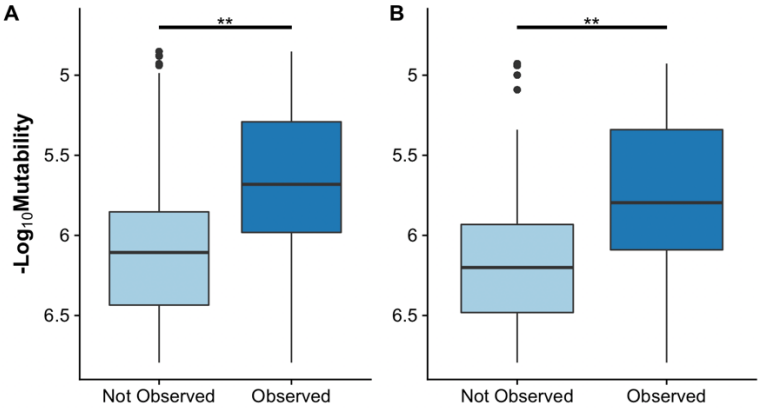
726 **Table 1.** Comparison of different measures to distinguish neutral from non-neutral mutations using
727 combined experimental dataset. Scores developed in this study (B-Score and LR) are underlined. The best
728 performance is shown in bold font. AUC-ROC and AUC-PR values for observed frequency count were
729 extrapolated since some experimentally validated mutations were not observed in tumor samples.
730
731

Measure	**AUC-ROC**	**AUC-PR**	**Matthews correlation**	**Sensitivity at 10% FPR**
LR	0.80	0.61	0.47	0.50
B-Score	0.82	**0.68**	**0.58**	0.57
Condel	0.87	0.59	0.53	**0.62**
FatHMM	**0.88**	0.63	**0.58**	0.52
Mutability	0.67	0.33	0.21	0.24
Observed frequency	0.71	0.58	0.52	0.48

732
733

734 **Supplementary Figures and Tables**

735



736

737

738 **Figure S1.** Mutability values for observed mutations in COSMIC v85 subset cohort and all theoretically
739 possible mutations that were not observed: **(A)** Codon mutability; **(B)** Nucleotide mutability. Differences
740 on Mann-Whitney-Wilcoxon test significant at $p < 0.01$.

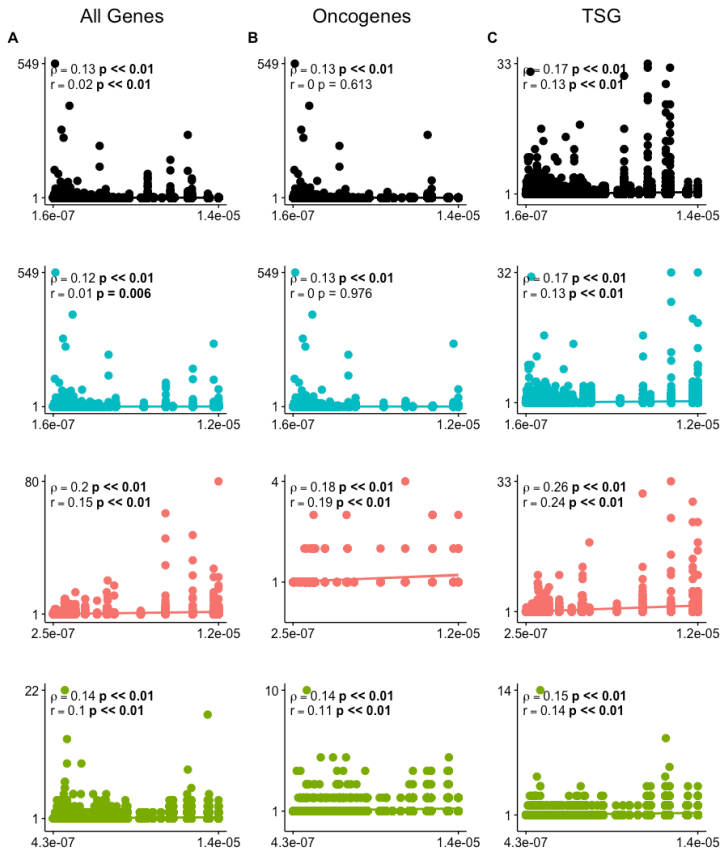


Figure S2. Relationship between codon mutability and observed mutation frequency by mutation type and role in cancer. **(A)** Scatterplot showing the observed mutation frequency for all genes and mutation types. **(B)** Scatterplots for oncogenes (n = 202) and **(C)** TSG (n = 166) for all mutation types. 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.

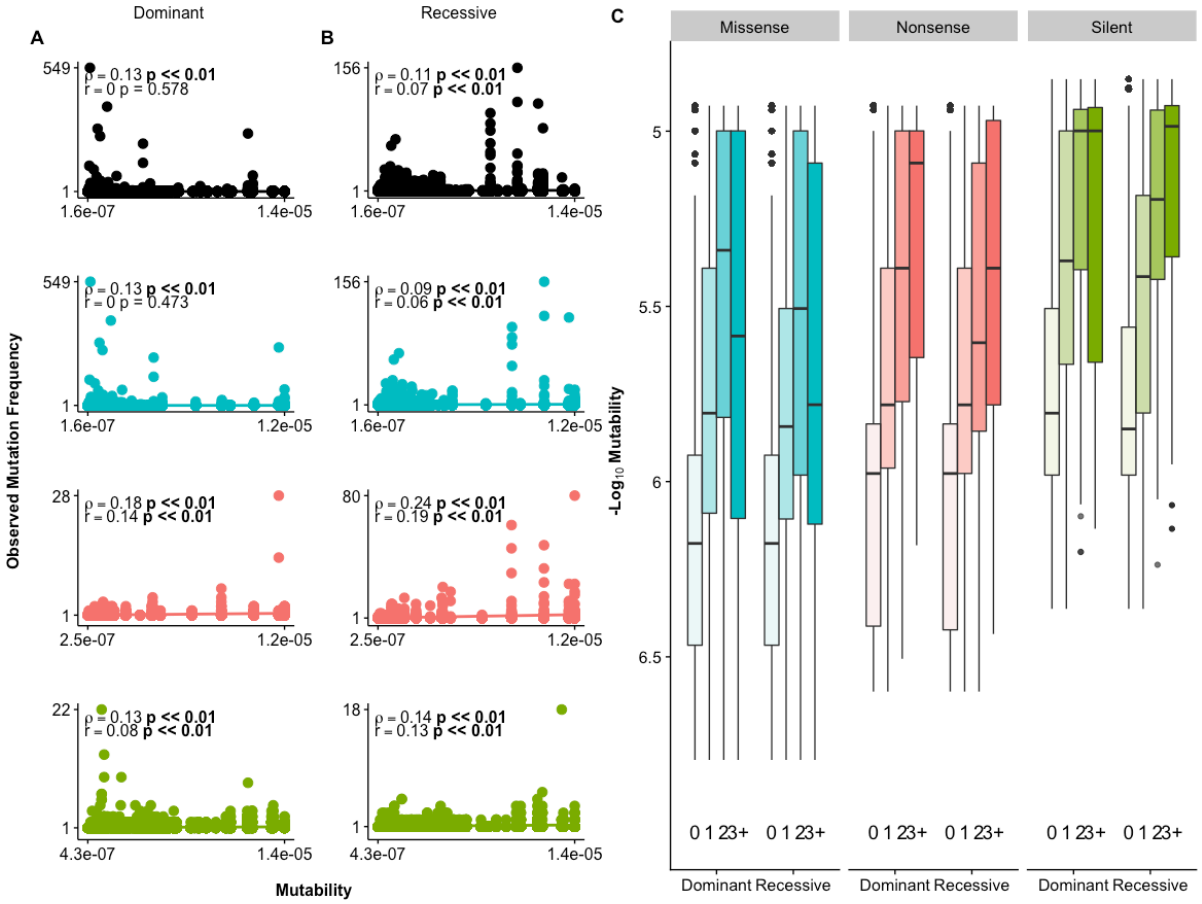


Figure S3. Relationship between codon mutability and observed mutation frequency by mutation type and molecular genetics. **(A)** Genes with only dominant mutations, **(E)** Genes with only recessive mutations. Different colors show scatterplots broken down by mutation type: missense (blue), nonsense (red) and silent (green). As in Figure 1E, 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.

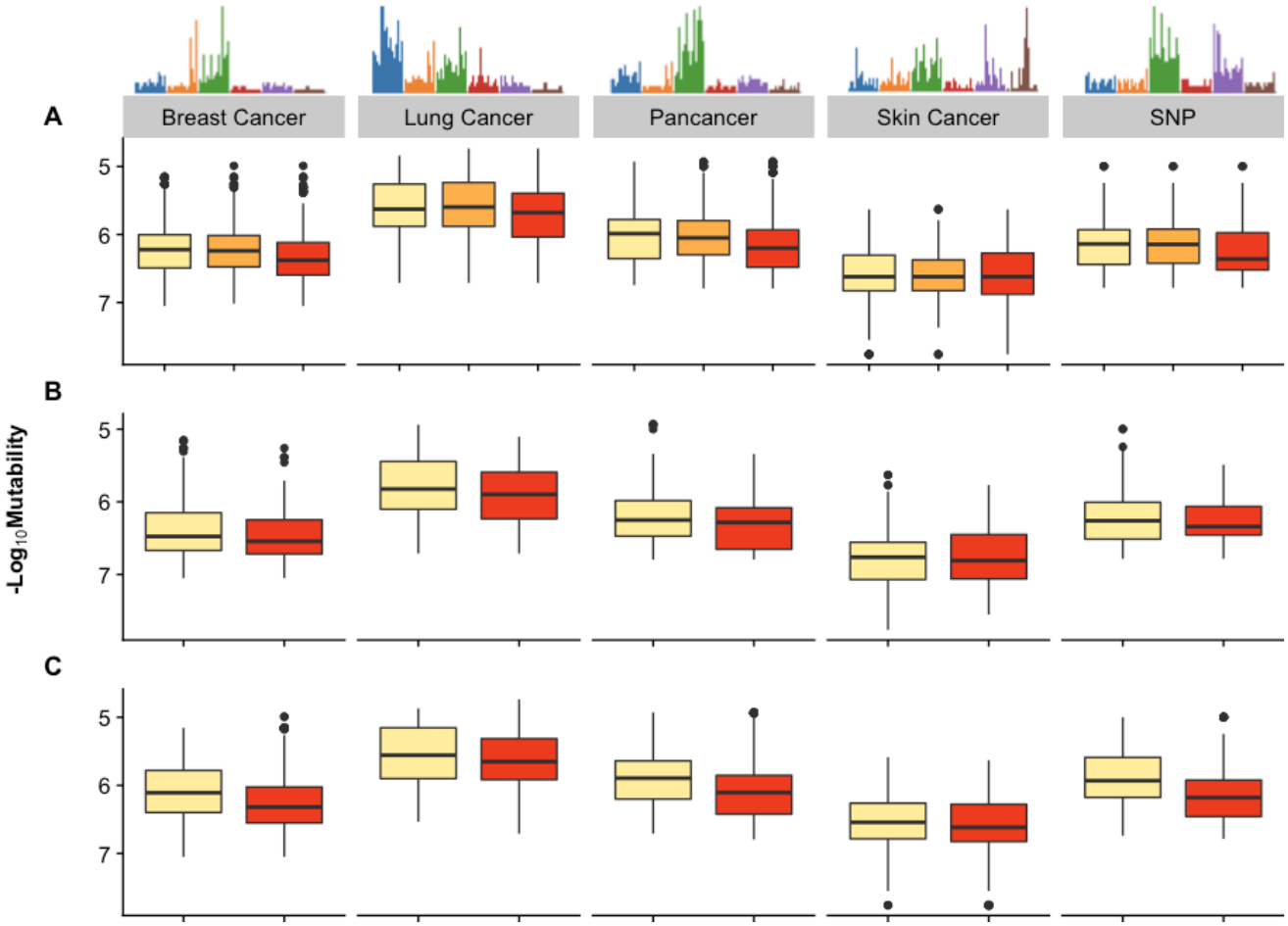
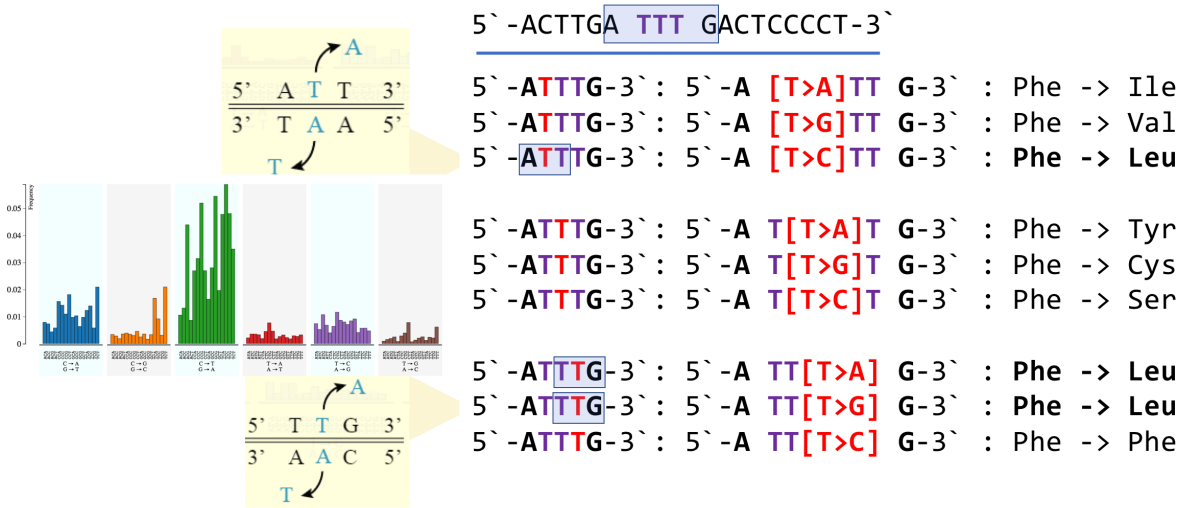


Figure S4. Mutability of missense mutations calculated with different background models. **(A)** TP53 **(B)** BRCA1 **(C)** Martelotto et al. datasets shown in Figure 4. As in Figure 4, colors refer to functional annotation, yellow – Function/Neutral/Benign, orange – partially-functional, red – Non-functional/Non-neutral/Deleterious. Mutability values were calculated with background models of breast cancer, lung adenocarcinoma cancer, skin adenocarcinoma, 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.



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

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

Figure S5. 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.

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

		Count group			
		0 (not observed)	1	2	3+
		Codon mutations			
Mutation Type	Missense	2625306	31510	1898	844
	Nonsense	140706	3264	322	193
	Silent	425059	11009	770	136
		Nucleotide mutations			
	Missense	2923312	32510	1942	844
	Nonsense	174350	3418	327	197
	Silent	906077	11665	677	124
		Codon mutations - Oncogene			
	Missense	772789	10355	670	333
	Nonsense	40310	664	36	7
	Silent	124054	3890	327	47
		Codon mutations - TSG			
	Missense	1023631	13063	763	237
	Nonsense	55041	1907	224	140
	Silent	165909	4287	290	59
		Codon mutations - Dominant			
	Missense	1678008	20256	1201	440
	Nonsense	90646	1462	94	28
	Silent	270997	7414	554	83
		Codon mutations - Recessive			
	Missense	689246	9154	559	371
	Nonsense	36894	1605	212	163
	Silent	112121	2808	167	44
		Codon mutations - Dom/Rec			
	Missense	57126	727	44	10
	Nonsense	2818	95	8	1
	Silent	9190	278	9	1
		Codon mutations - Rec/X			
	Missense	5437	60	4	0
	Nonsense	283	19	2	0
	Silent	879	25	0	0
		Codon mutations - No Molecular Genetics Information Provided			
	Missense	195489	1313	90	23
	Nonsense	10065	83	6	1
	Silent	31872	484	40	8

771 **Table S2.** Correlation between mutability and recurrence of mutations in cancer-associated genes
772 https://www.ncbi.nlm.nih.gov/research/mutagene/static/data/TableS2.xlsx
773

774 **Table S3.** Combined dataset with experimentally annotated neutral and non-neutral mutations in 17
775 genes
776 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	2.17×10^{-9}	functional - non-functional
		0.36	functional - partially-functional
		1.41×10^{-8}	non-functional - partially-functional
	Breast cancer	1.84×10^{-9}	functional - non-functional
		0.47	functional - partially-functional
		6.11×10^{-8}	non-functional - partially-functional
	Lung adenocarcinoma	9.82×10^{-5}	functional - non-functional
		0.23	functional - partially-functional
		3.13×10^{-5}	non-functional - partially-functional
	Skin adenocarcinoma	0.33	functional - non-functional
		0.41	functional - partially-functional
		0.43	non-functional - partially-functional
	SNP	4.90×10^{-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	6.24×10^{-8}	
	Breast cancer	9.06×10^{-7}	
	Lung adenocarcinoma	0.02	
	Skin adenocarcinoma	0.28	
	SNP	8.90×10^{-12}	
BRCA1 - DMS	Pan-cancer	0.04	
	Breast cancer	0.07	
	Lung adenocarcinoma	0.12	
	Skin adenocarcinoma	0.29	
	SNP	0.49	

780 **Table S5.** Performance of different measures on each individual dataset from the combined data set.
781

	AUC-ROC	AUC-PR	Matthews's correlation	Sensitivity at 10% FPR	Dataset
LR	0.57	0.35	0.25	0.29	BRCA1-DMS
B-score	0.58	0.32	0.23	0.29	
Condel	**0.78**	**0.56**	0.50	**0.53**	
FatHMM	0.68	0.52	**0.52**	0.35	
Mutability	0.57	0.33	0.24	0.29	
Frequency	0.51	0.23	0.03	0.11	
LR	0.77	0.95	0.32	0.45	Martelotto et. al
B-score	0.79	0.95	0.32	0.52	
Condel	0.85	**0.97**	0.45	**0.55**	
FatHMM	**0.88**	0.96	**0.69**	**0.56**	
Mutability	0.64	0.91	0.18	0.24	
Frequency	0.7	0.94	0.28	0.47	
LR	0.61	0.06	0.17	0.26	POLE/POLD1 Variants
B-score	0.65	**0.13**	**0.30**	0.32	
Condel	**0.90**	0.07	0.22	**0.65**	
FatHMM	0.36	0.01	0.04	0.47	
Mutability	0.57	0.02	0.05	0.24	
Frequency	0.57	0.09	0.21	0.23	
LR	0.82	0.77	0.54	0.58	TP53 Transactivation
B-score	**0.84**	**0.80**	**0.62**	**0.67**	
Condel	**0.84**	0.74	0.56	0.6	
FatHMM	0.80	0.65	0.51	0.46	
Mutability	0.61	0.47	0.19	0.18	
Frequency	0.79	0.76	0.58	0.67	

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

Gene Weight	AUC-ROC	AUC-PR	MCC	Sensitivity at 10% error	Dataset
SNP – 10,000 bp	0.84	0.68	0.58	0.57	**Combined**
SNP – 20,000 bp	0.84	0.68	0.58	0.57	
SNP – 100,000 bp	0.84	0.69	0.58	0.58	
SNP – 200,000 bp	0.84	0.69	0.58	0.58	
No-outlier based weight	0.67	0.58	0.52	0.50	
Mutation based weight	0.71	0.60	0.55	0.50	
No Weight	0.82	0.68	0.58	0.57	
Silent mutation based weight	0.82	0.67	0.58	0.55	
SNP – 10,000 bp	0.75	0.68	0.44	0.47	**Validation**
SNP – 20,000 bp	0.75	0.68	0.43	0.47	
SNP – 100,000 bp	0.75	0.68	0.43	0.47	
SNP – 200,000 bp	0.75	0.68	0.43	0.47	
No-outlier based weight	0.74	0.67	0.44	0.45	
Mutation based weight	0.74	0.67	0.44	0.45	
No Weight	0.76	0.69	0.44	0.49	
Silent mutation based weight	0.72	0.55	0.37	0.40	

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 in area

784

Table S7. Number of annotated mutations in combined dataset by gene and the gene's role in cancer.

	Neutral Mutations	Non-Neutral Mutations	Role in Cancer
BRAF	0	23	Oncogene
BRCA1	369	95	TSG
BRCA2	50	12	TSG
DICER1	0	11	TSG
EGFR	0	33	Oncogene
ERBB2	5	33	Oncogene
ESR1	0	7	Oncogene, TSG, Fusion
IDH1	0	1	Oncogene
IDH2	0	3	Oncogene
KIT	1	24	Oncogene
KRAS	0	25	Oncogene
MYOD1	0	1	Oncogene
PIK3CA	1	31	Oncogene
POLD1	945	5	Oncogene
POLE	1806	29	Oncogene
SF3B1	0	7	Oncogene
TP53	610	540	Oncogene, TSG, Fusion