Mutational likeliness and entropy help to identify driver mutations and their functional role in cancer

Giorgio Mattiuz[1,2]*, Salvatore Di Giorgio[1,3], Lorenzo Tofani[4], Antonio Frandi[5], Francesco Donati[1], Matteo Brilli[6,7]*, and Silvestro G Conticello[1,7]*

* Corresponding Authors

Affiliations

[1] Core Research Laboratory, ISPRO, Firenze, Italy.

[2] Department of Experimental and Clinical Medicine, University of Firenze, Italy.

[3] Department of Medical Biotechnologies, University of Siena, Italy

[4] Department of Neurosciences, Psichology, Drug Research and Child Health, University of Firenze, Italy.

[5] Department of Fundamental Microbiology, University of Lausanne, Switzerland.

[6] Department of Biosciences, Paediatric Clinical Research Center "Romeo ed Enrica Invernizzi", University of Milano, Italy.

[7] These authors contributed equally.

Abstract

Genomic alterations in cancer originate from mutational processes taking place throughout oncogenesis and cancer progression. Here we define two properties of somatic mutations crucial for cancer evolution, and we show that cancer-driver mutations do not conform to the mutational pattern characterizing the evolution of the cancer genome. Our analysis can also identify novel putative cancer driver genes and differentiate between gain of function and loss of function mutations.

Mutational likeliness and entropy help to identify driver mutations and their functional role in cancer

Cancer genomes display a complex landscape determined by the accumulation of mutations. In individual cancer types specific mutational signatures[1] have been identified, that can be ascribed to errors during DNA replication or repair, as well as to other effects of exposure to mutagens. However, in each tumour only a handful of mutations define the cancer phenotype and direct the cancer evolutionary process (*driver* mutations)[2–4]. The majority of somatic mutations found in a given tumour are mostly neutral with respect to cancer evolution (*passenger* mutations), as they hitchhike on fitness-increasing mutations. Passenger mutations greatly exceed driver mutations, hence they can be used to describe the *neutral* mutational landscape of cancer genomes.

Identifying driver mutations in the haystack of passenger mutations is a major outstanding problem in cancer research. Several approaches to discriminate between driver and passenger mutations have been developed, based on factors such as mutation frequency[5–7], gene expression[8], protein domain analysis[9,10], markers of positive selection[11], network enrichment analysis[12] and recurrently amino acid change analysis[13–16].

Since driver mutations are under positive selection[11], their mutational pattern should diverge from that observed in the much more numerous passenger mutations. In order to pinpoint these differences, we have used a Markov model trained on synonymous mutations in order to calculate the probability of non-synonymous mutations. We chose a zero-order model, rather than a higher order one, as we are dealing with coding sequences where higher order patterns are confounded by constraints related to the protein sequence (*via* the triplets of the genetic code). The parameters of the transition matrix of our model were worked out based on the synonymous mutations in a dataset of cancer mutations derived from the one generated by Chang et al.[16] using several cancer-genome resources. The full dataset comprises ~2 million single-nucleotides variants present in over 11,000 cancer exomes from patients who had one of 41 tumour types. We trained the model on synonymous mutations as they are mostly neutral, and therefore they reflect the outcomes of errors in the replicative/repair pathways and/or exposure to mutagens during cancer onset and progression (*Mutational Background Model*, **Fig.1a**). We then used the background model to calculate two scores for each group of non-synonymous mutations (GNSM; the set of all mutations hitting the same codon in the same transcript among all

patients): (1) the *mutational likeliness*, measures the probability for a given GNSM to be generated by the background model. It allows the identification of nucleotide changes which diverge from the mutational pattern of the tumour; (2) the *mutational entropy* score, calculated applying the Shannon entropy to the amino acid changes determined by mutations, measures the bias towards a specific amino acid encoded by a GNSM compared to the expectations from the background model (**Fig.1a).** While *mutational likeliness* quantifies the distortion caused by selective pressures on mutations, the entropy assesses if the resulting set of amino acid(s) is biased towards a few of them: entropy is zero when, of all possible outcomes, we observe the presence of a single amino acid in a given GNSM and reaches its maximum when all amino acids are equally represented.

The significance of the two scores for each GNSM was then calculated by simulating 10,000 equally sized groups of mutations following the probability in the background model. We then transformed *mutational likeliness* and *entropy* for each GNSM into the corresponding Z-scores using the estimated average and standard deviation obtained from the simulations (**Supplementary Table S1**).

We used the Cancer Gene Census[17] to compile a list of non-synonymous mutations in *bona fide* cancer driver genes (*Driver*), and a list of non-synonymous mutations in other genes (*non-Driver*), presumably representing passenger mutations. Next, we compared the Z-score distributions of the *mutational likeliness* and *mutational entropy* for the *Driver* and the *non-Driver* sets of mutations (**Supplementary Table S1**).

The distribution of the *mutational likeliness* for *Driver* mutations is shifted towards lower values than that of *non-Driver* mutations ($p < 0.0001$, **Fig.1b**). This means that a substantial fraction of nucleotide changes in the driver genes, those with a very small probability, are not explained by the mutational background model: we suggest this to be due to positive selection acting on the cancer driver mutations. This effect is even stronger ($p < 0.0001$) when we only consider a set of validated driver codons (*Validated*, 293 positions on 75 genes, **Supplementary Table S1**), whose oncogenicity has been experimentally demonstrated **(Fig.1b**). This further reinforces the notion that the distribution of mutations that have been the subject of Darwinian selection during cancer development stand out as not representative of the overall outcomes of the mutational processes taking place in the cell during tumorigenesis or in the tumour.

With respect to *mutational entropy*, again the values we obtained are lower for *Drivers* than for *non-Drivers*, albeit with lower statistical significance ($p = 0.0281$, **Fig.1c**): this suggests that selection favours a reduced set of amino acid changes at these positions compared to the background model.

However, we did not observe a significant difference between *Validated* and *non-Driver* mutations with regards to *mutational entropy*.

We reasoned that the entropy of a GNSM might correlate tightly with protein function: only a small subset of the amino acid changes that can be derived from a given codon will positively alter the function of the protein; conversely, a wider set of amino acid changes can lead to loss of function. For instance, among TP53 mutations, A161T results in a gain of function (GOF)[18], whereas R181P results in loss of function (LOF)[19]: their entropy scores are -0.91 and 0.08 respectively, as mutations at A161 very often result in a Threonine, while mutations at R181 have more freedom concerning the resulting amino acid. When we further analysed *Validated* amino acid changes that are associated with GOF (n=100) or with LOF (n=172) (**Supplementary Table S1**), we found no difference between their *mutational likeliness* distributions (**Fig.1b**); on the other hand, the distribution of entropy values for GOF mutations when compared to that of LOF mutations was significantly shifted towards smaller values ($p = 0.0016$; **Fig.1c**). Thus, whereas the same evolutionary pressure applies to both GOF and LOF mutations (as they bear similar likeliness distributions), the difference in *mutational entropy* points towards the functional differences acquired through the amino acid changes. These scores thus highlight the dichotomy between gain- and loss-of-function mutations, which is fundamental in cancer biology, as in general GOF is characteristic of oncogenes and LOF of tumour suppressor genes.

Based on these results, we examined whether *mutational likeliness* and *mutational entropy* could provide a way to identify novel driver genes. We performed bibliographic searches on the 31 mutant genes in the *non-Driver* list for which *mutational likeliness* and *entropy* were below the 1st percentile of the *Driver* distribution (**Fig.1d**; **Supplementary Table S2**): we found that 18 of them (58%) are convincingly associated with oncogenesis and/or cancer phenotypes (on 6 there is no information: see **Supplementary Table S2**). The same was true for 68 out of 172 genes of *non-Driver* within the 5th percentiles with bibliographic information (**Supplementary Table S2**). Among the 31 genes below the 1st percentile, only one has been identified by other approaches[5,11,15–17] (11 below the 5th percentile) (**Supplementary Table S2**). Interestingly, among these mutations, one on RUNX2, a transcription factor associated to cancer and bone metastasis, is not currently present in any of the cancer driver gene lists (**Supplementary Table S2**). This suggests that our approach might be able to point towards a different set of cancer driving genes not usually discoverable.

The cancer genome is riddled with somatic mutations that are in part spontaneous and in part result from exogenous mutagens. Only a few of the mutant genes are subject to selection, and this has made it

difficult to disentangle driver mutations from passenger mutations within the mutational landscape. We have found that *likeliness* and *entropy* of individual mutations can identify known driver mutations and predict some that have yet to be confirmed.

The different likeliness between driver and passenger mutations distributions indicates that the mutational signatures observed in cancer genomes mainly reflect passenger mutations. Even though known cancer driver mutations conform to the mutational background (e.g. PIK3CA mutations in HPV-related cancers[20]) the shift towards low *mutational likeliness* suggests that –overall– cancer driver mutations often escape the usual outcomes of the mutational processes acting on the cell.

This analysis may also enable for the first time to differentiate between gain- and loss-of-function mutations. Whereas mutations are stochastic phenomena, and a set of mutations may bear the 'signature' of a mutagenic agent, evolutionary pressure depends on GOF, or LOF, or any change of function entailed by any particular mutation, regardless of its original signature. Our methodology can identify cancer-driving mutant genes that are missed by other approaches and can guide the selection of potential cancer drivers for experimental validation. This is of special importance for precision medicine, since driver mutations are preferred potential targets of new therapies.

Methods

The cancer dataset

We started with a dataset of cancer mutations obtained from Chang et al.[16], comprising of 2 million single-nucleotides variants (SNVs) identified in 11,115 cancer exomes from 41 tumours types. From the starting dataset, we removed the SNPs and all that SNVs that did not match the correct position in the coding sequence with regards to grch37 coding sequences. The dataset we used thus consisted of 1,799,208 mutations.

Mutational background model

For all patients with more than 100 total mutations in coding regions, we selected the synonymous ones to calculate the transition matrix of a zero-order Markov model describing the background mutational process of each tumor type. All data from patients affected by the same tumor type contribute to the transition matrix for the given tumor type (Adrenocortical Carcinoma, Adenoid Cystic Carcinoma, Hypodiploid Acute Lymphoid Leukemia, Bladder Cancer, Breast Invasive Carcinoma, Cervical Squamous Cell Carcinoma And Endocervical, Chronic Lymphocytic Leukemia, Colorectal Carcinoma, Cutaneous Squamous Cell Carcinoma, Non Hodgkin Lymphoma, Esophageal Carcinoma, Gallbladder Carcinoma, Glioblastoma, High Grade Pontine Glioma, Head And Neck Squamous Cell Carcinoma, Kidney Chromophobe Cancer, Kidney Renal Clear Cell kirc Carcinoma, Kidney Renal Papillary Cell Carcinoma, Acute Myeloid Leukemia, Brain Lower Grade Glioma, Liver Hepatocellular Carcinoma, Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, Lung Small Cell Carcinoma, Medulloblastoma, Mantle Cell Lymphoma, Myelodysplasia, Multiple Myeloma, Rhabdoid Cancer, Neuroblastoma, Nasopharyngeal Carcinoma, Adenocarcinoma Ovarian Serous Cystadenocarcinoma, Pancreatic Adenocarcinoma, Pancreatic Neuroendocrine Carcinoma, Pilocytic Astrocytoma, Skin Cutaneous Melanoma, Stomach Adenocarcinoma, Thyroid Carcinoma, Uterine Corpus Endometrial Carcinoma, Uterine Carcinosarcoma, Prostate Adenocarcinoma). For cancer types for which the transition model could not be calculated as there were no tumours with at least 100 mutations we built an average model obtained by averaging all models from the other cancer types.

As expected, considering the directionality of coding sequences, complementary mutations (e.g. A → C and T → G) are not symmetric and we treated them separately. Each model is thus composed by a vector of 12 probabilities, one for each possible nucleotide change:

i.e. $M = [p(A \rightarrow C), p(A \rightarrow G), p(A \rightarrow T), ..., p(T \rightarrow C), p(T \rightarrow G)]$.

As we analyzed mutations in their codon context, we define as a group of non-synonymous mutations (GNSM) the set of all mutations hitting the same codon in the same transcript in different patients. We considered for analysis only those codons for which at least three mutations existed in the dataset.

Mutational likeliness

The mutational background model is used to extrapolate the background probability distribution of amino acids resulting from non-synonymous mutations hitting a certain codon (**Fig.1a)**.

This distribution is then compared with the set of amino acids observed at a given codon mutated. For instance, let's suppose that several patients have a given codon CAC (Histidine) mutated. From the background model we know the values of p(C→A, G, T) and p(A→C, G, T) and therefore we can calculate the probability of going from one codon to another by means of a single point mutation, e.g. p(CAC → AAC, GAC,...). From the probabilities towards each possible codon, we calculate the corresponding expected distribution of amino acids by merging the probability of codons coding for the same amino acid. In the case of the CAC codon we get all the possible resulting codons, which lead to N, D, Y, P, R, L, Q, Q. Thus, each amino acid change has its own probability to happen in this codon context:

p(H→N) = p(CAC→AAC) and p(H→Q) = p(CAC→CAA) + p(CAC→CAG).

Since we only consider non-synonymous mutations, while some of the amino acids reachable from a certain codon are synonymous, the probabilities for non-synonymous changes are then rescaled to 1. We define the mutational likeliness score as:

$$L = \sum_{i=1}^{n} log_{10} p_i^{bkg}$$

with *n* the number of observed mutations and p_i^{bkg} the probability of a certain mutation given by the background model; this formula therefore allows to calculate the probability of a certain set of mutations at a certain codon for a specific tumour background model. Mutations in line with the background model will have large *mutational likeliness*, as they will tend to have large p_i^{bkg}, while those that do not conform to the background model show a bias towards smaller scores. We then perform random sampling to assess the significance of the score observed for each GNSM: we use the background model to generate 10,000 equally sized sets of mutations starting from the wild type codon

and we calculate the average and standard deviation of the score. This procedure allows to calculate the Z-score and therefore the significance for each observed GNSM. This measure allows the identification of the codons where the set of observed mutations diverges the most with respect to the background model.

Mutational entropy

In order to analyze whether (a subset of) driver mutations tend to prefer certain amino acid changes, we consider all amino acids changes found in each patient genome for a given GNSM, and we calculate its entropy using the Shannon formula[21]:

$$H = -\sum_{i=x,y,z..} f_i log_2 f_i$$

Where the sum runs over the different amino acids encoded by the GNSM (x, y, z...) with frequencies f_i. As in the previous case, the entropy of each set of mutations is transformed in a Z-score by using the background model to produce 10,000 equally sized group of amino acids from which we calculate the expected average entropy and its standard deviation. Therefore, while the *mutational likeliness* identifies mutations with a pattern of nucleotide changes differing significantly from the expected (given the background model), the *mutational entropy* identifies those mutations that might be in line with the background probabilities but not with the expected amino acid distribution.

Cancer Gene Census List

We exploited the Cancer Gene Census[17] to define a list of known cancer driver genes. In this work, we excluded all those genes that are not present in our dataset and whose oncogenicity derives from copy number alterations, gene fusions and truncations, insertions, or deletions. Our *Driver* dataset comprises of 2666 mutations on 399 driver genes (**Supplementary Table S1**). The complementary dataset of mutations outside these driver genes (*non-Driver*) includes 31037 mutations on 9846 *non-Driver* genes (**Supplementary Table S1**).

Experimentally Validated Driver Codons (*Validated*)

Starting from our list of 2666 driver mutations, we manually selected 293 codons for which the effects of the specific mutation on the gene have been reported in the literature. Many of the validated

positions were taken from the JAX Clinical Knowledgebase (The Jackson Laboratory; https://ckb.jax.org) and confirmed through further bibliographic search. We divided the *Validated* amino acid changes in two subgroups containing mutations inducing either gain-of-function (GOF = 100) or loss-of-function (LOF = 172) depending on the effect of the amino acid change on protein function. To create GOF and LOF lists, we usually had to consider only the starting amino acid position (e.g. BRAF V600); in rare cases a single codon could be classified both as GOF or LOF, depending on the resulting amino acid change; in our analysis we considered each mutation separately, in both subgroups (e.g. the mutation of Y646 in the EZH2 gene leads to LOF for Y646C or to GOF for Y646F) (**Supplementary Table S1**). It is noteworthy that only 10% of mutations in the *Drivers* dataset has been tested and catalogued as GOF or LOF. This highlights the need for experimental approaches to validate cancer-associated mutations and to allow a proper identification and functional characterization.

Driver mutations in the *non-Driver* dataset

We selected mutations in the *non-Driver* dataset whose *mutational likeliness* and *entropy* scores were below the 1st and the 5th percentiles of the *Driver* dataset distributions. For each gene, we performed a bibliographic analysis (using all aliases for the gene name) and selected those whose function had been linked to cancer processes and phenotypes. (**Supplementary Table S2**).

Statistical analysis

The distributions of the mutated codon groups were compared using the Mann-Whitney test with Bonferroni multiple comparison correction. Continuous variables were expressed as mean, standard deviation, median, 25th and 75th percentiles. The significance level was set to 5%. The statistical analysis was performed using SAS 9.3.

Table 1

The p-values for each comparison are shown. Significant differences are reported in **Fig.1b** and **c**.

Mutational Likeliness			
			Bonferroni (p-value)
Driver	vs	*non-Driver*	< .0001
Validated	vs	*non-Driver*	< .0001
non-Driver	vs	*GOF*	< .0001
non-Driver	vs	*LOF*	< .0001
GOF	vs	*LOF*	1.000

Mutational Entropy			
			Bonferroni (p-value)
Driver	vs	*non-Driver*	0.0281
Validated	vs	*non-Driver*	0.0696
non-Driver	vs	*GOF*	< .0001
non-Driver	vs	*LOF*	< .0001
GOF	vs	*LOF*	0.0016

Authors Contribution:

GM, AF, MB, and SGC conceived the study; GM, AF, MB, and SGC wrote the manuscript; MB developed the mathematical model; GM, SDG, and MB performed the bioinformatics analysis; LF performed the statistical analysis; GM, FD performed the bibliographic analysis.

References

1. Alexandrov, L. B., Nik-zainal, S., Wedge, D. C. & Aparicio, S. A. J. R. Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
2. Nordling, C. O. A new theory on cancer-inducing mechanism. *Br. J. Cancer* **7,** 68–72 (1953).
3. Luzzatto, L. & Pandolfi, P. P. Causality and Chance in the Development of Cancer. *N. Engl. J. Med.* **373,** 84–88 (2015).
4. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 118–23 (2015).
5. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–8 (2013).
6. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
7. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 1–20 (2016).
8. Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **5,** 3231 (2014).
9. Miller, M. L. *et al.* Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Syst.* **1,** 197–209 (2015).
10. Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48,** 827–37 (2016).
11. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* (2017). doi:10.1016/j.cell.2017.09.042
12. Merid, S. K., Goranskaya, D. & Alexeyenko, A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. *BMC Bioinformatics* **15,** 1–21 (2014).
13. Szpiech, Z. A. *et al.* Prominent features of the amino acid mutation landscape in cancer. *PLoS One* **12,** 1–12 (2017).
14. Anoosha, P., Sakthivel, R. & Gromiha, M. M. Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets. *Biochim. Biophys. Acta-*

Biomembranes **1862,** 155–165 (2016).

15. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173,** 371–385.e18 (2018).
16. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34,** 1–11 (2015).
17. Futreal, P. a. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).
18. Okada, M., Tessier, A., Bai, L. & Merchant, J. L. p53 mutants suppress ZBP-89 function. *Anticancer Res.* **26,** 2023–2028 (2006).
19. Kato, I. *et al.* Oxidized DJ-1 Inhibits p53 by Sequestering p53 from Promoters in a DNA-Binding Affinity-Dependent Manner. *Mol. Cell. Biol.* **33,** 340–359 (2013).
20. Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-Mediated Cytosine Deamination Links PIK3CA Helical Domain Mutations to Human Papillomavirus-Driven Tumor Development. *Cell Rep.* **7,** 1833–1841 (2014).
21. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).

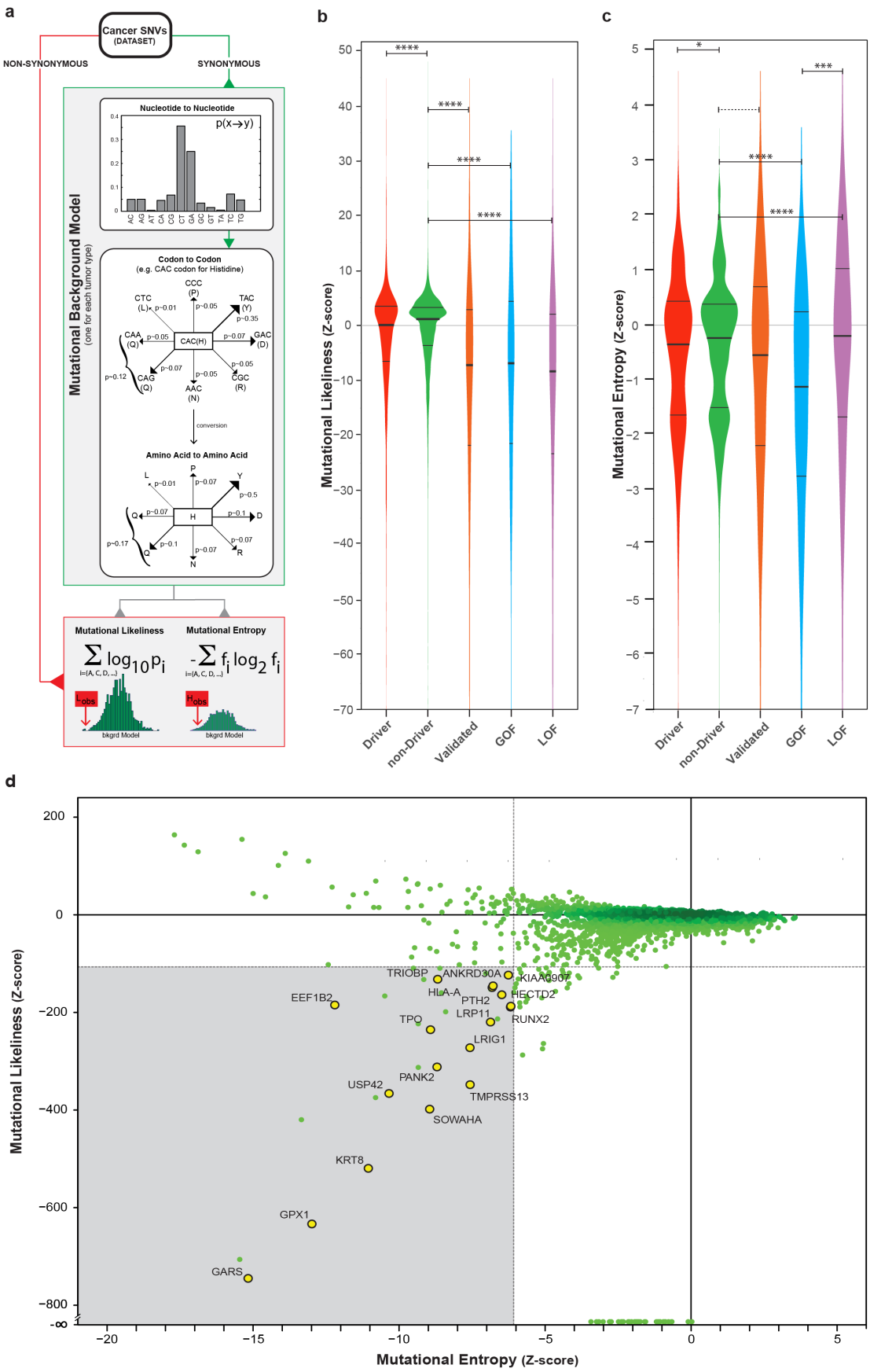
a
Cancer SNVs
(DATASET)
NON-SYNONYMOUS
SYNONYMOUS
Mutational Background Model
(one for each tumor type)
Nucleotide to Nucleotide
p(x→y)
Codon to Codon
(e.g. CAC codon for Histidine)
conversion
Amino Acid to Amino Acid
Mutational Likeliness
Mutational Entropy
bkgrd Model
b
Mutational Likeliness (Z-score)
Driver
non-Driver
Validated
GOF
LOF
c
Mutational Entropy (Z-score)
d
Mutational Likeliness (Z-score)
Mutational Entropy (Z-score)
TRIOBP
ANKRD30A
KIAA0907
HLA-A
PTH2
HECTD2
EEF1B2
LRP11
RUNX2
TPO
LRIG1
PANK2
USP42
TMPRSS13
SOWAHA
KRT8
GPX1
GARS

Fig.1. Mutational likeliness and mutational entropy distribution highlight differences between cancer driver and passenger mutations. (**a**) Synonymous mutations from the cancer dataset are used to derive the probability for each single nucleotide change in each tumour type (*Nucleotide to Nucleotide mode*l). The values obtained are used to build a model representing the probabilities for each non-synonymous change to occur given a tumour-specific mutational background (*Codon to Codon model*). From this, the probability for all the codons that can be produced by single nucleotide substitutions from any other codon was calculated (summing probabilities as needed; e.g. the case of Q) (*Amino Acid to Amino Acid model*). This model is then used to estimate the *mutational likeliness* and *entropy* scores for each non-synonymous mutation using the equations shown. p_i in the *mutational likeliness* formula represents the probability of a given amino acid change, depending on the wild type codon and the mutational background model. The *mutational entropy* is calculated considering the frequency (f_R) of amino acid changes observed at the mutated codon. The *mutational likeliness* and *mutational entropy* were converted to Z-scores and the relative distributions were plotted. (**b, c**) The box-plots show the Z-score distributions of codons arising from cancer driver mutations (*Driver*, n=2666 positions on 399 genes), non-driver mutations (*non-Driver*, n=31037 positions on 9846 genes), experimentally validated ones (*Validated*, n=293), gain-of-function (*GOF*, n=100) and loss-of-function (*LOF*, n=172). and the horizontal black lines indicate the median for each group. The statistical significance is indicated *-p≤0.05; **-p≤0.01; ***-p≤0.001; ****-p≤0.0001; full details in the **Table 1**); comparisons not reaching statistical significance are indicated by a dotted horizontal line. (**d**) Scatterplot of mutational likeliness and entropy of the *non-Driver* dataset (**Supplementary Table S2**). Areas below the 1st percentile of the *Driver* dataset (*mutational likeliness* = -113.68; *mutational entropy* = -6.16) are shaded in grey. Codon mutations on genes experimentally associated to cancer related processes are indicated in yellow.