Supplementary Information for

Criticality in Tumor Evolution and Clinical Outcome

Erez Persi*, Yuri I. Wolf, Mark D.M. Leiserson, Eugene V. Koonin*, Eytan Ruppin*

* Correspondence to: **E.P.** (erezpersi@gmail.com), **E.V.K.** (koonin@ncbi.nlm.nih.gov) or **E.R.** (eruppin@gmail.com)

This PDF file includes:

Figs. S1 to S21

Tables S1

Other supplementary materials for this manuscript include the following:

Datasets S1

www.pnas.org/cgi/doi/10.1073/pnas.1807256115

Supplementary Information

Figures

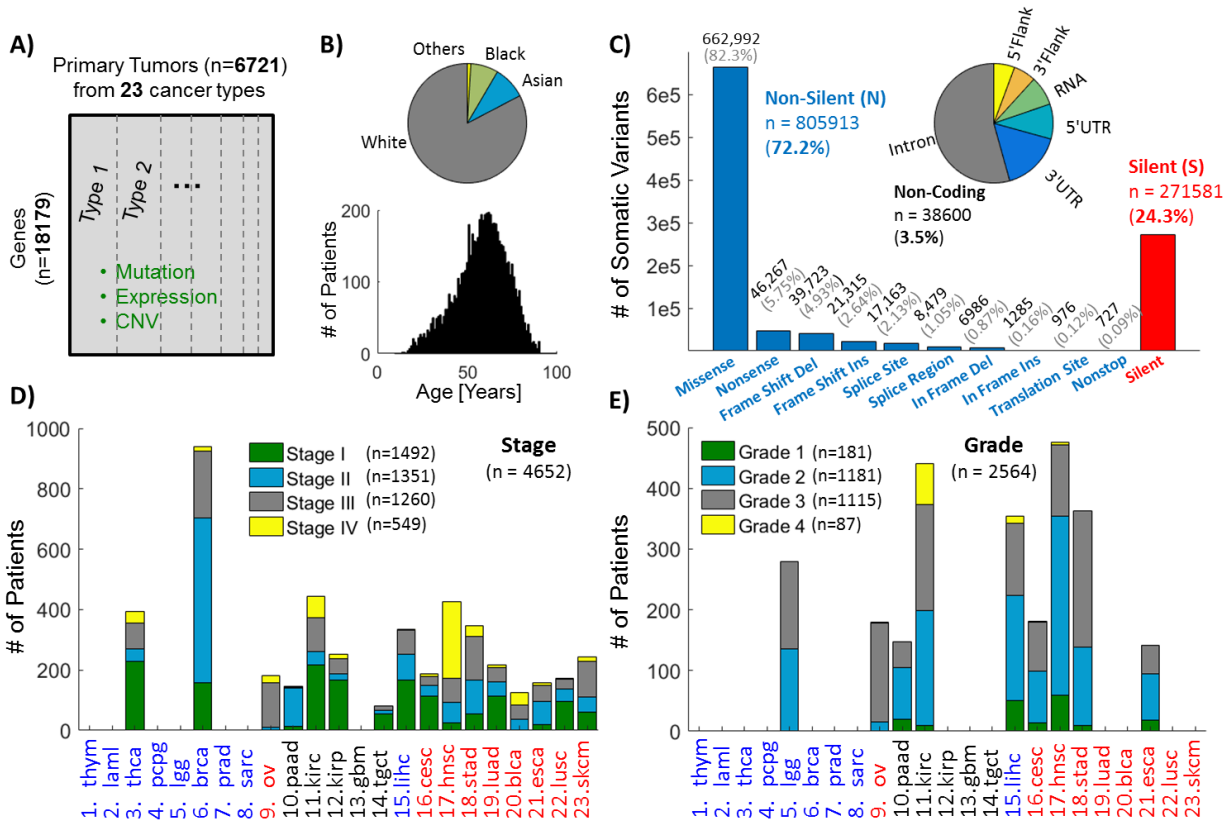

Figure S1: **Data Statistics**: Primary tumor data downloaded from cBioPortal of all TCGA cancer types containing at least 100 patients each, covering 6721 patients across 23 cancer types (including 287 Melanoma metastatic tumors). **A)** We analyzed all "3-way complete" samples, for which gene expression, CNV and somatic mutations data exist, and considered all protein coding genes that have both unique NCBI-Entrez and SwissProt IDs (n=18179). **B)** Distribution of race and age across patients. **C)** Distribution of the number of mutations in the proteome associated with each mutation class across patients. **D)** Distribution of stage across cancer types. **E)** Distribution of grade across cancer types.

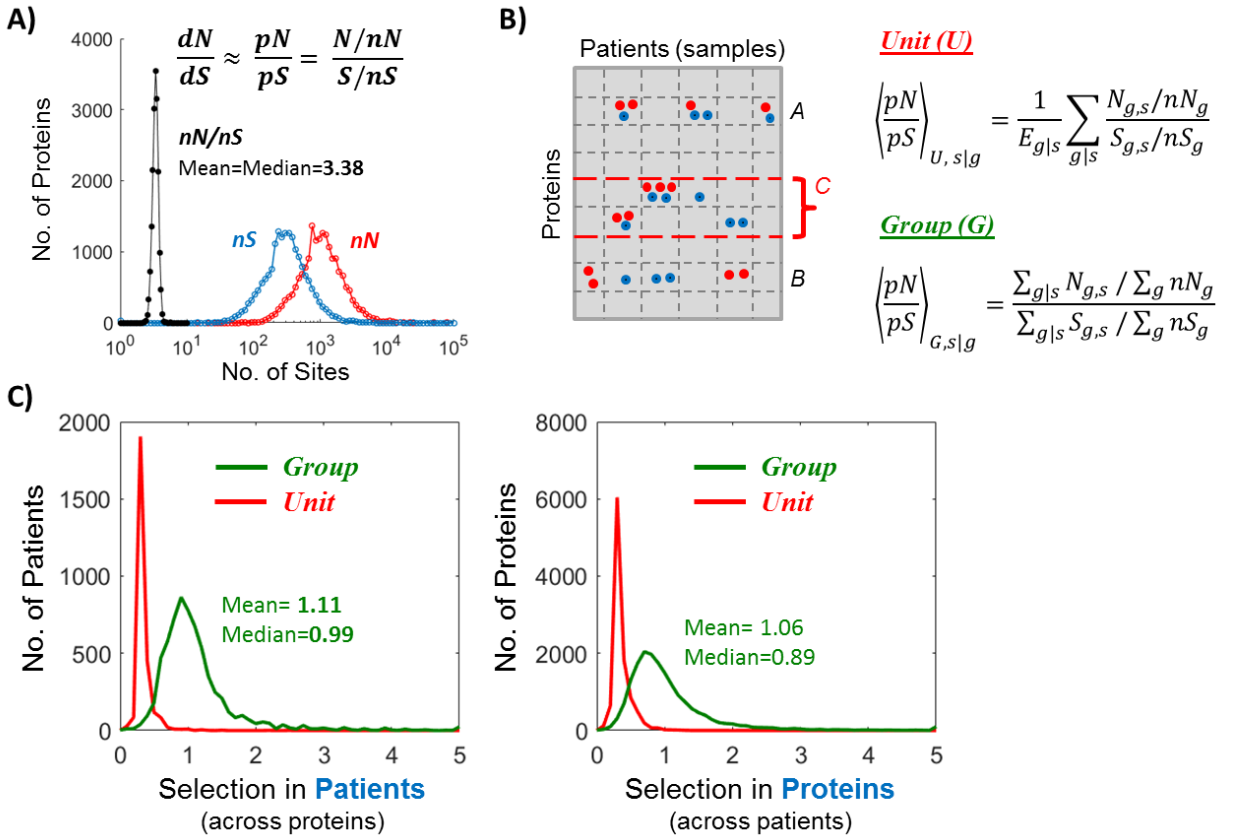

Figure S2: **Integrative Measures of selection (*dN/dS*). A)** Distribution of the number of non-synonymous (*nN*, red) and synonymous (*nS*, blue) sites, and their ratio (*nN/nS*, black) for all protein-coding genes (n=18179) as inferred by considering all alternative nucleotides in each position in each protein-coding sequence (**Methods**). Selection (*dN/dS*) in cancer is approximated by the ratio between the number of *N* mutations per *N* sites (*pN=N/nN*) and the number of *S* mutations per *S* sites (*pS=S/nS*). **B)** Illustration of the highly sparse mutation data, exemplified by (protein-sample pair) units that contain few mutations (e.g., protein 'A') and by the fact that in the vast majority of cases only one type of mutation (i.e., *N* or *S*) exists (e.g., protein 'B'). Precisely, out of the 6721x18179 units (n=122181059), there are 963,048 cells with either *N* or *S* mutations, but only 35278 contain both *N* and *S* mutations. **C)** Unit-based and group-based estimates of selection. In principle, *dN/dS* at the proteome level can be estimated either by taking the average of *dN/dS* across units (U), or by considering a group (G) of genes (sum over *g;* for example the set 'C' of cancer-genes) or a group of samples (sum over *s*), and estimating *dN/dS* based on the total number of mutations in the group. Given the highly sparse mutation data, Unit-based estimators are highly biased and inadequate for analysis, whereas Group-based estimators are adequate, as exemplified by their distribution around unity. For analysis and comparisons across patients, we measured the selection in each patient, exploiting the statistical power of the overall distribution of all mutations across the proteome (summing over genes), providing a measure of selection acting on the entire proteome for each patient. Although at the pan-cancer level both the integration over genes (*left*) and over sample (*right*) capture dominance of neutral evolution, the later is sensitive to the low statistical power, limited by the number of patients (**Figure S3**), whereas the former is not (cf. **Figure 2C**).

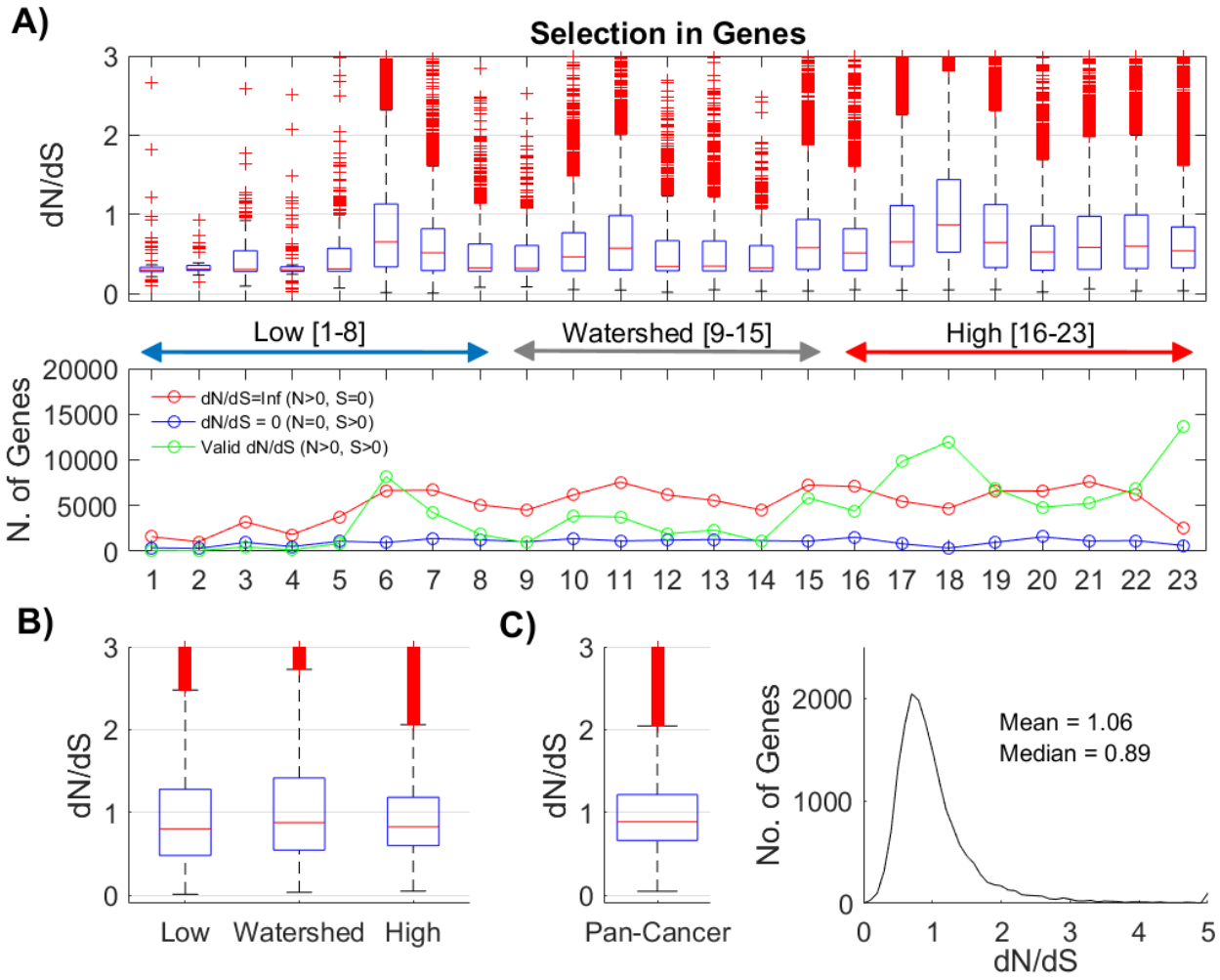

Figure S3: **Sensitivity of dN/dS measures to statistical power**. **A)** The *dN/dS* distribution of selection in genes, evaluated by integrating mutations across patients (typically, ~100-500), in each cancer type (*top*), and the number of genes with valid and invalid *dN/dS* values (*bottom*). Many gene have *dN/dS=Inf* (*N>0, S=0*; red) and fewer have *dN/dS=0* (*N=0, S>0*; blue), because the detection threshold of *N* is lower than that of *S*. These genes are discarded from analysis; hence many *N* mutations are discarded. Genes with valid *dN/dS* (*N>0, S>0*; green) are statistically biased to be those that harbor *S* mutations, hence the biased *dN/dS<1*. Cancers for which a substantial number of valid values (picks in the green curve), display *dN/dS* distributions closer to neutrality, due to increased statistical power. **B)** Increasing statistical power by integrating mutations in each gene across more patients (by grouping cancers to low, medium and high *ML*) leads higher *dN/dS* values, correcting for the bias in (A), with dN/dS values distributed closer to unity. **C)** Increasing more the statistical power, by integrating mutations across all patients leads to *dN/dS* distributed around unity. The *dN/dS* at the proteome level, used in our analysis, provides a larger statistical power by integrating mutations across genes (n=18179), such that *dN/dS* is insensitive to this statistical bias, as evident from **Figure 2C**, depicting dominance of neutral evolution across all cancers (except Melanoma, a particular case analyzed in detail; See main text and **Figures 4-5**). Only values of $dN/dS \leq 3$ are displayed, for clarity of the figures.

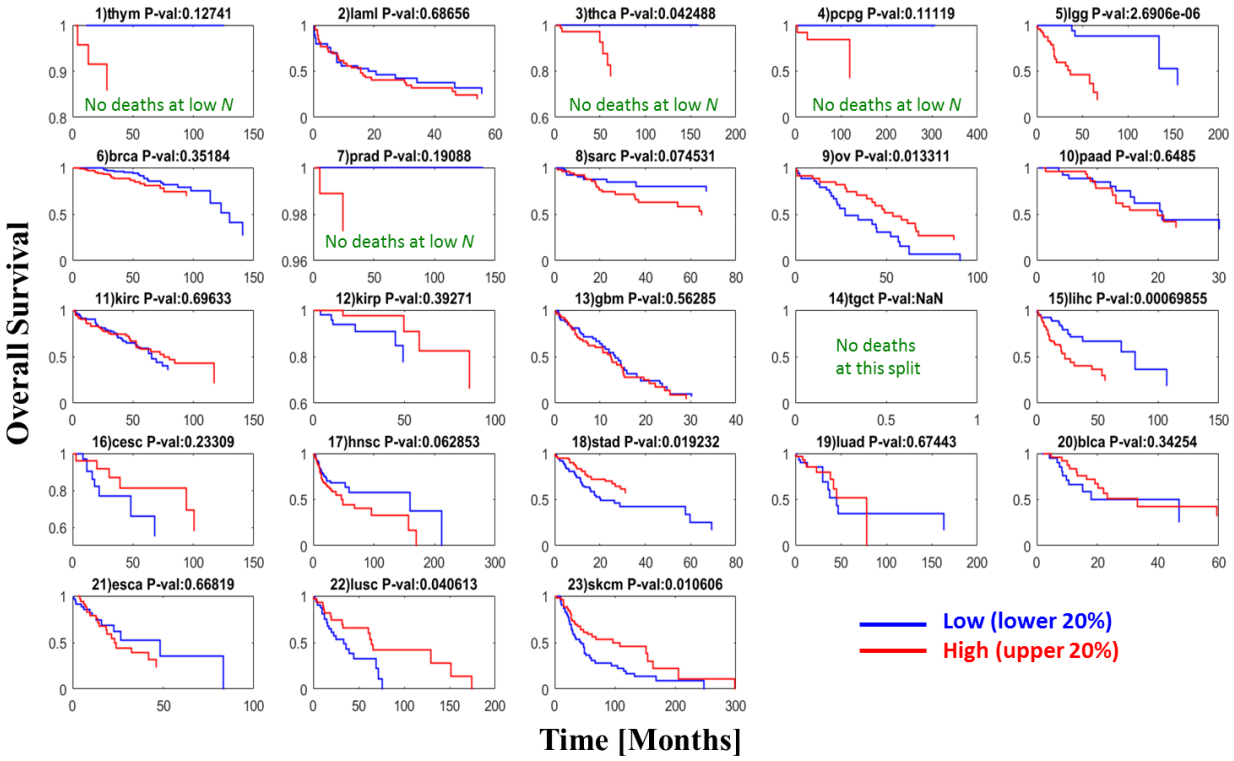

Figure S4: **KM of mutation load (*ML*) in each cancer type**. For each cancer type, survival rates were compared between patients with low (lower 20%) and high (upper 20%) number of non-silent (*N*) mutations (i.e., considering 40% of the data). Cancer types are ordered by *ML* as in **Figure 1** in the main text. At low *ML* (#1-8), low number of mutations is associated with better prognosis in most cases, whereas at high *ML* (#16-23), the opposite trend is observed (significant in Stad, Lusc, Skcm and to a lesser extent in Cesc and Blca). At the mutation watershed (#9-15), there is no obvious trend, and Ovarian (#9) and Liver (#15) cancers at the edges of the watershed behave oppositely, similar to the Cox regression results summarized in **Figure 1** of the main text. Imagining the *watershed* transition as a point, it is easy to imagine that Ovarian may belong to the high *ML* cancer type and that Liver may belong to the low *ML* cancer type (main text). See **Figure 1** for the Oncotree codes of cancer types.

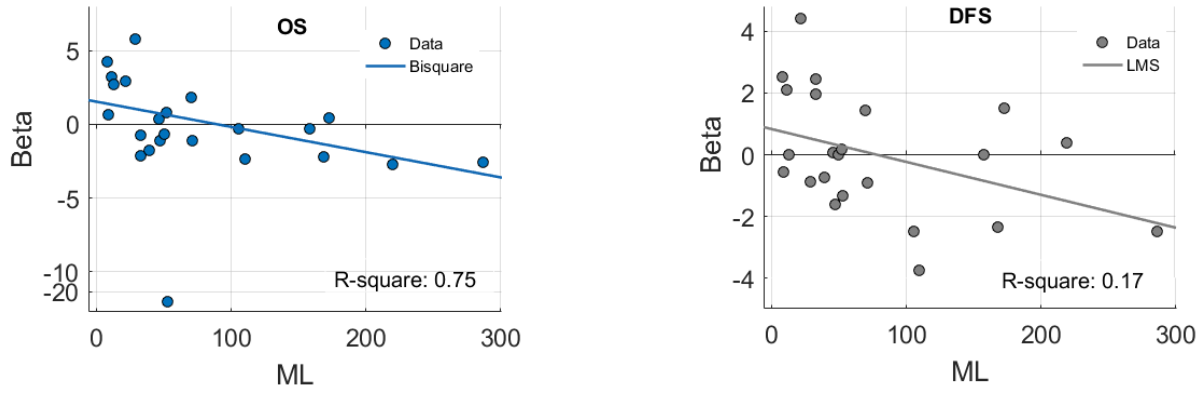

Figure S5: **Correlation between Beta and ML**. **A)** Betas (β) from the Cox analysis in **Figure 1** (i.e., applied to each cancer type separately) vs. the median mutation load of patients in each cancer, using OS data. Due to the outlier (with $\beta<-20$, in Tgct cancer) we used a Bisquare fit. **B)** Similar analysis using DFS data, where we used LMS fit (verifying that Bisquare gave similar results). Note that these correlations are likely underestimated, because the slope in low ML cancers is steeper than that in high ML cancers.

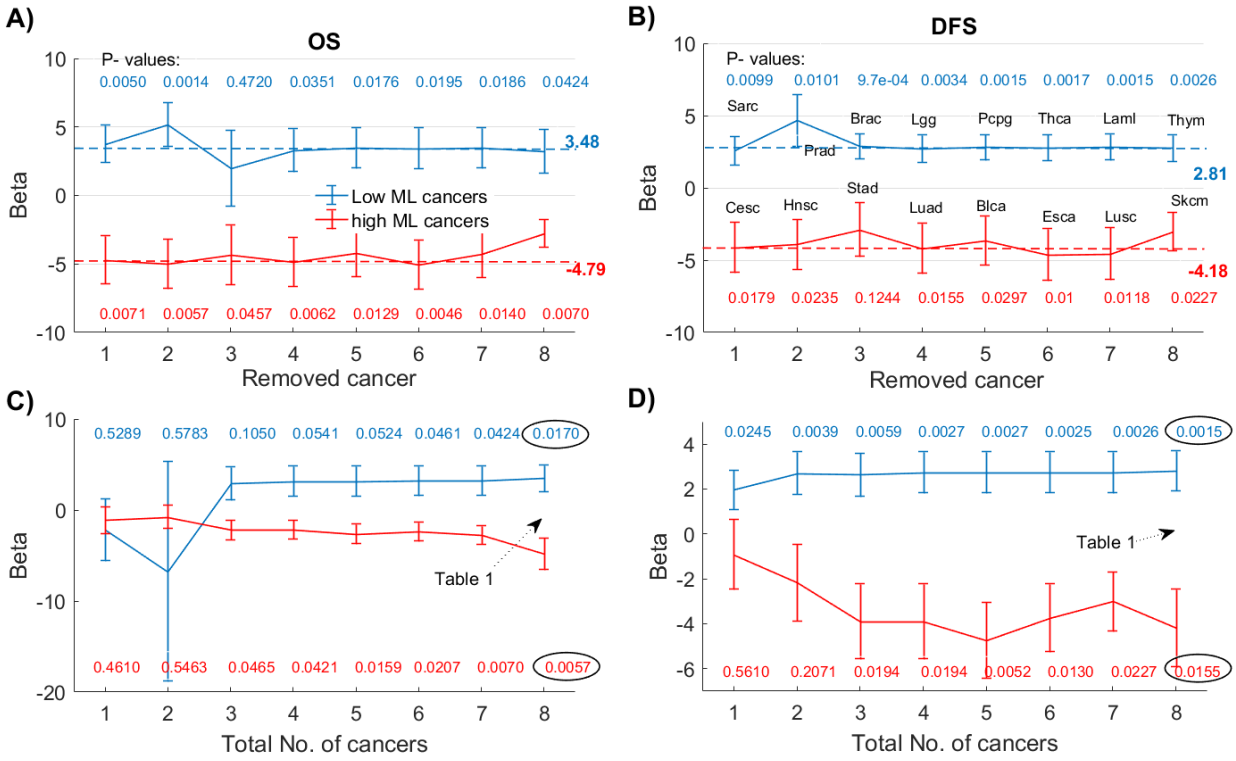

Figure S6: **Stability and significance of beta in the low (n=8) and high (n=8) ML A)** Beta (β) of stratified Cox analyses, using OS data, applied to the low ML group (blue) and high ML group (red), when each time a cancer is removed from the analysis (one from each group). Values obtained in **Table 1** are shown as dashed lined, for comparison. **B)** Similar analysis using DFS data. Cancers are removed from the tips of the watershed away (i.e., towards the lowest and highest ML in each group respectively), as indicated in B (Oncotree code as **Figure 1**). **C)** In the same order, cancers are now accumulated and a Cox analysis was applied each time, using OS data; indicating how rapidly the values of betas increase and become more significant as more cancers are considered in each group. The last values correspond to the values in **Table 1. D)** Similar analysis as in C, using DFS data.

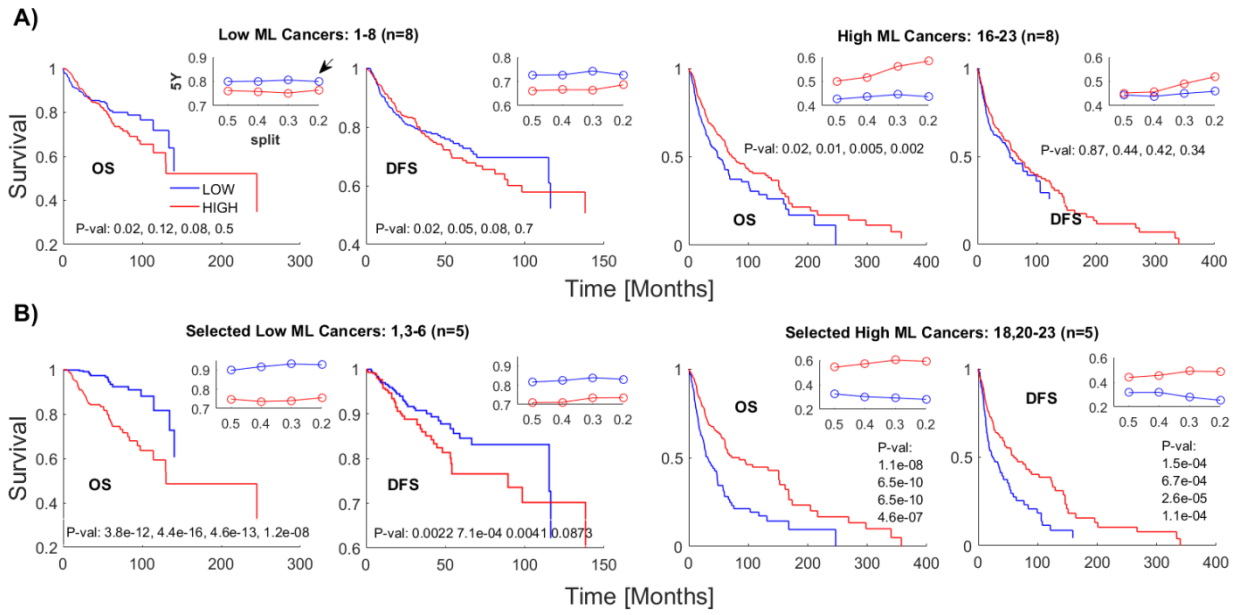

Figure S7: **Results of KM analysis applied for groups of cancers. A)** Analysis performed on the aggregated patients data belonging to the low ML cancers (n=8) and to the high ML cancers (n=8), showing a weak signal of opposing regimes, at the border of significance. Inset shows 5-years survival of patient with low number of mutation (blue) and those with high number of mutations (red), splitting the data by 50,40,30,20 upper and lower percentiles (x-axis values 0.5-0.2). Survival curves of the lower and upper 20% (arrow) are shown in the larger panels. **B)** Removing the 2 cancers closet to the watershed (i.e., Prad and Sarc from the low ML cancers, and Cesc and Hnsc from the high ML cancers), as well as the cancer with the lowest association with survival (i.e., $\beta \sim 0$), from each side/group (i.e., Laml from the low ML cancers and Luad from the high ML cancers); thus analyzing 5 cancers in each group, recapitulate the signal of opposing regimes (**Figure 1** and **Table 1**).

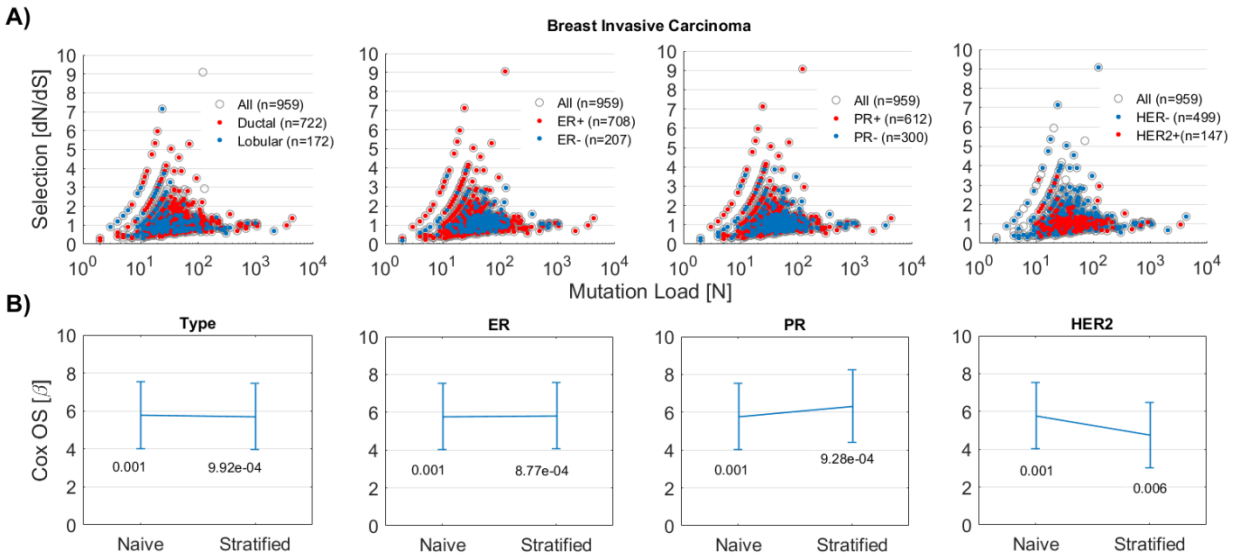

Figure S8: **Analysis of breast cancer subtypes**. **A)** Selection (*dN/dS*) vs. the mutation load (*ML*) is shown for all breast cancer patients (n=959), with different classifications of tumors to subtypes superimposed (color): ductal/lobular and ER, PR, HER2 status, from left to right. Subtypes are distributed comparably in the *dN/dS-ML* phase-plane. **B)** Beta (β) of the naïve Cox model (obtained in **Figure 1**) and when tumors are stratified by the corresponding classification. The values of β are highly robust and significant.

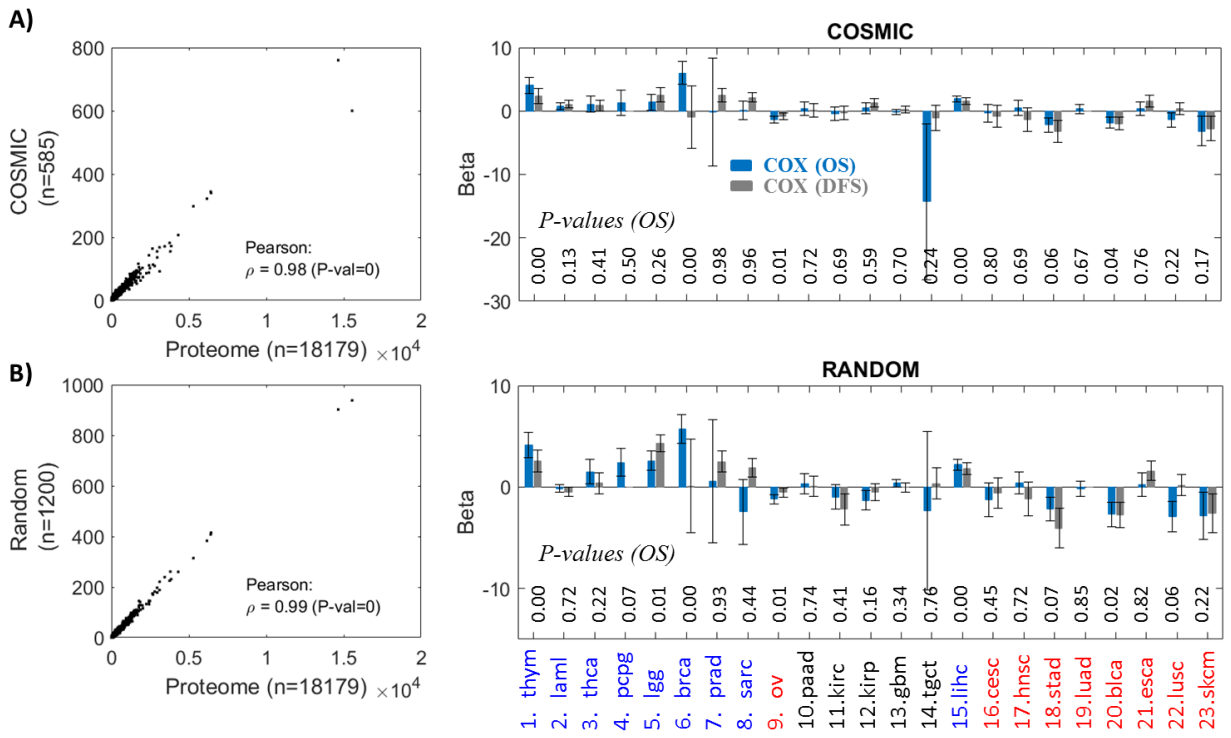

Figure S9: **Robustness of *ML* to the gene cohort used for estimation**. **A)** The correlation between *ML* estimated for the entire proteome and for cancer-genes (COSMIC) is extremely high and significant (left). Thus, it is not surprising that the pattern observed in **Figure 1** of the main text is robust to the choice of genes that are used to estimate *ML*. **B)** Similar analysis, performed for random sets of genes. A larger set of random genes had to be used to achieve a number of mutations comparable to that in the COSMIC genes which, by definition, harbor more mutations. Nonetheless, for a sufficiently large set of random genes, *ML* is highly correlated with *ML* evaluated across the entire proteome, and thus captures the transition in the clinical outcome.

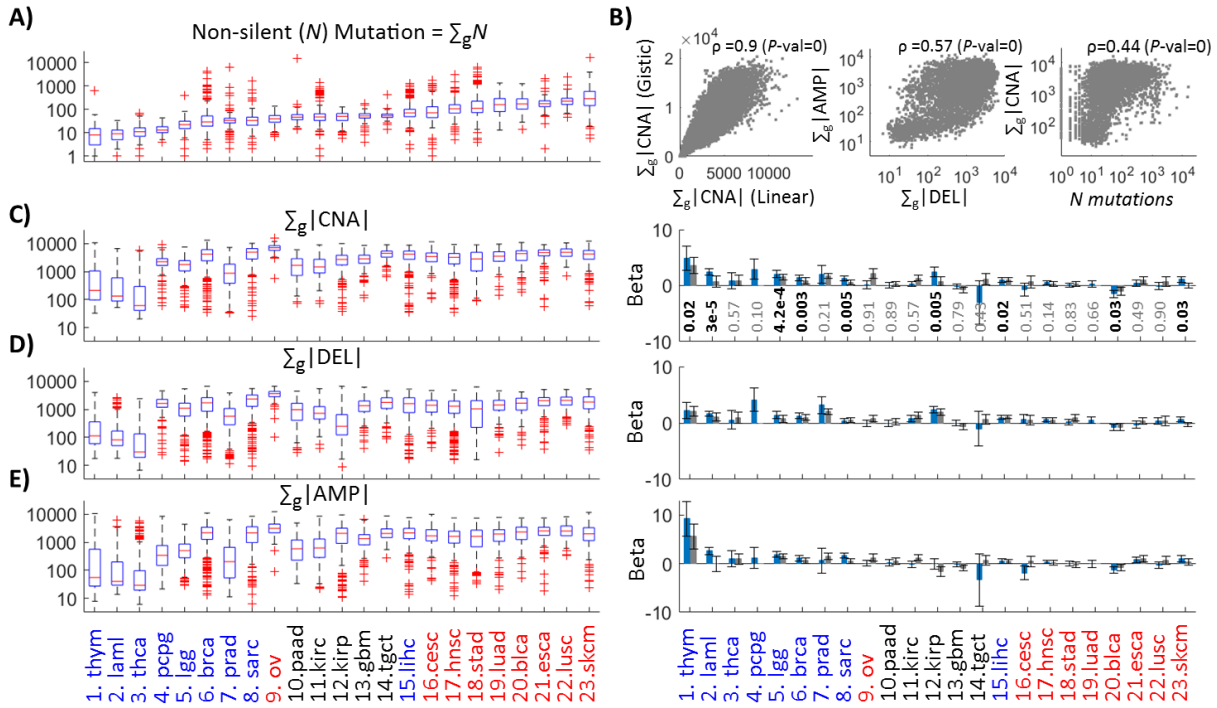

Figure S10: **Copy number alteration (CNA) analysis and comparison to *ML*. A)** The distribution of non-silent (*N*) mutations across cancer types as in **Figure 1** is shown, for the ease of visual comparison. **B)** Similarly to the mutation load, we estimate the overall CNA as the sum of the absolute values, using two standard measures: 'Linear' (a continuous variable) and 'Gistic' (a rounded integer variable) (**Methods**). As these measures are highly correlated they provide comparable association with survival. Continuing with the 'Linear' estimator, we observe that the overall CNA deletions (|DEL|) and the overall CNA amplifications (AMP) are also correlated. Also, to a lesser extent, the overall CNA (|CNA|) is correlated with the *ML*, i.e. the total number of non-silent (*N*) in each tumor proteome (Spearman correlation, ρ=0.44). This correlation suggests that CNA would capture a similar prognostic signal to that of *ML*. **C-E)** However, although CNA predicts pooper survival in patients harboring large CNA at low *ML*, it does not capture any transition in the clinical outcome of the type observed for *ML* in **Figure 1** of the main text. This is the case for the overall CNA (C) as well as deletions (D) and amplifications (E) each when tested separately. Complementary stratified Cox regression analysis of cancer in low and high *ML* verify these results, and show that also the copy-number DNA burden, measured as the fraction of altered genes (gain/loss) (**Methods**) behaves similarly to the overall CNA (**Table 1**).

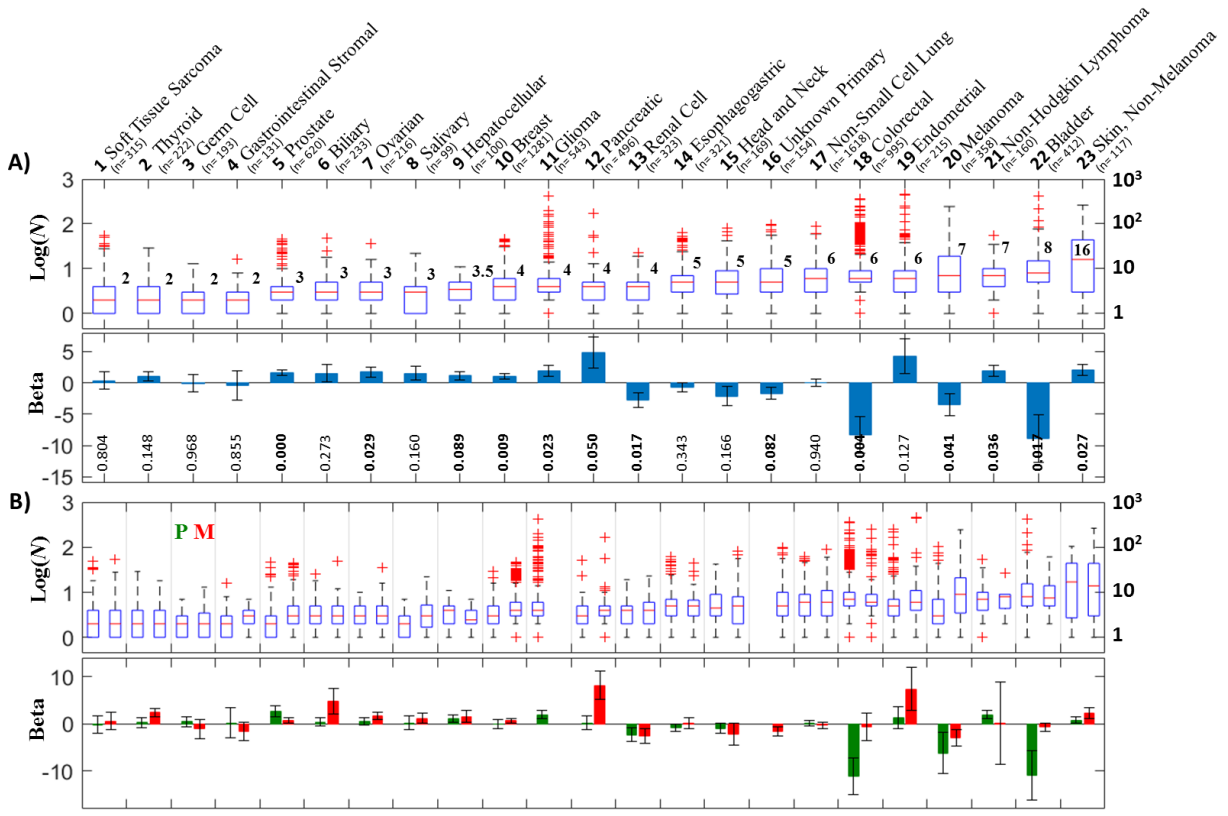

Figure S11: **Validation by analysis of the MSK-impact-2017 cohort**. A recent cohort of >10,000, which contains 43% samples from metastatic sites across >30 cancer types, was analyzed (**Methods**). Only 414 genes were sequenced in this study, for which only calls of non-silent (*N*) mutations are available. Given that our measures are apparently insensitive to the exact set of genes used to estimate *ML* (**Figure S9**), using this cohort we sought to validate our main result in **Figure 1**, that is, the existence of a transition in clinical outcome at high mutation loads (*ML*). The survival times in this cohort are provided in days intervals corresponding to the time from procedure to last follow-up. **A)** As in the main text, this analysis centered on cancer types that included at least 100 patients (one type with 99 patients was included). Discarding the sample site (i.e., including both primary and metastatic sites), we find a comparable pattern of both the mutation distribution (top) and the transition in clinical outcome, using Cox regression (bottom). **B)** The analysis was then repeated for metastatic (M, red) and primary (P, green) sites separately. Under clonal evolution, a metastatic site, by definition, should contains at least all the mutations of the corresponding primary site when taken from the same individual. The upper panel shows that, in most cases, metastatic sites indeed contain more mutations however this is not always the case because samples are taken from different individuals (top). Nonetheless, Cox regression indicates a transition in both cases although with lower significance compared to (A), presumably, because of the reduced number of patients.

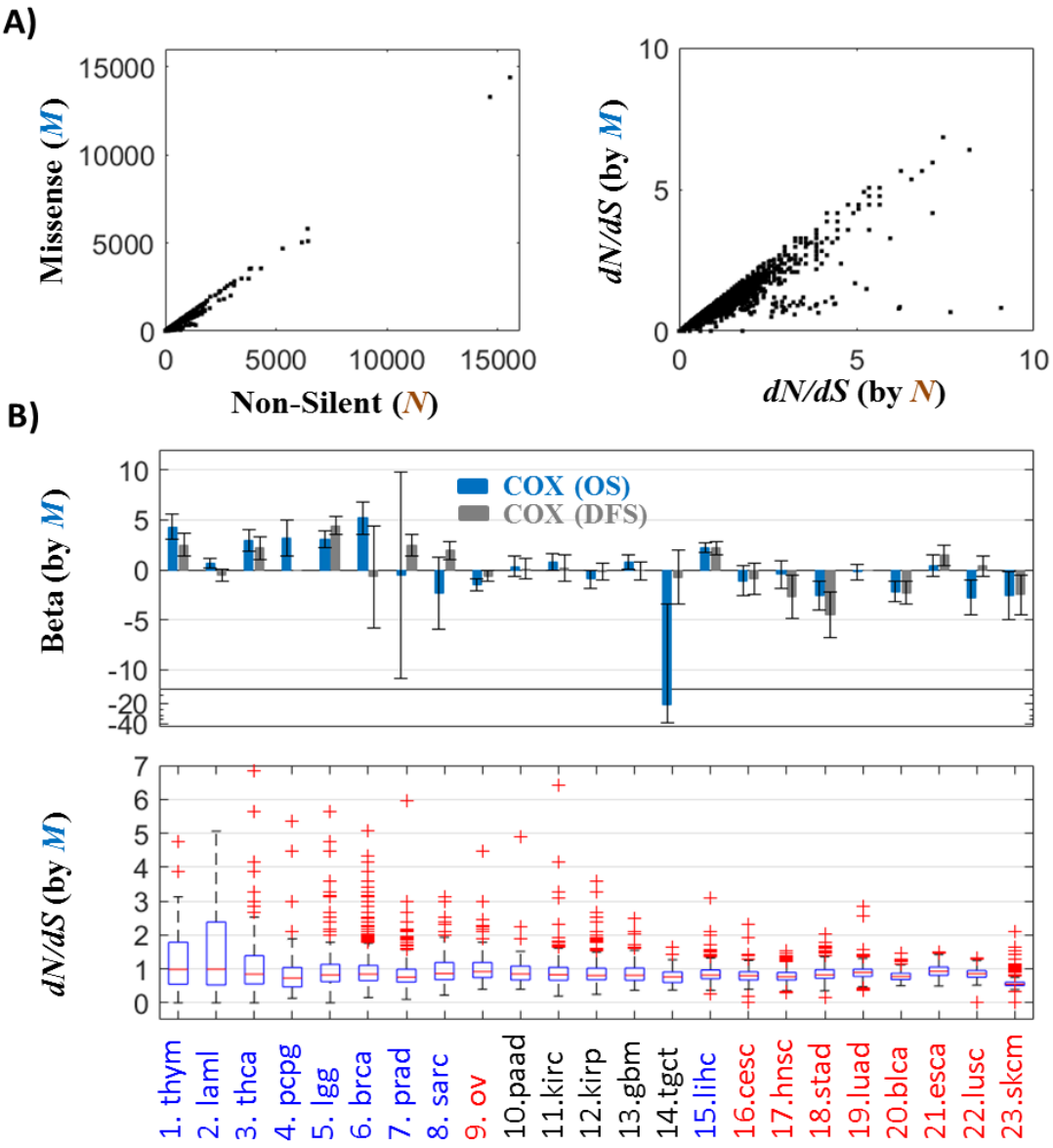

Figure S12: **Robustness of *ML* and *dN/dS* to mutation class**. The analysis presented in **Figures 1-2** of the main text was repeated for missense mutations alone. **A)** *ML* and *dN/dS* each is highly correlated for all non-silent (*N*) and missense only (*M*) mutations. **B)** The transition in clinical outcome around the mutation *watershed*, where the distribution of *ML* is flat across cancer types (**Figure 1**) is recapitulated (Top). The distributions of *dN/dS* for missense mutations only slightly shifted toward lower *dN/dS* ratios (i.e., over-estimating purifying selection), compared with the neutral evolution depicted in **Figure 1**. The heavier tails of positive selection at low *ML* and the lack thereof at high *ML*, are evident (Bottom).

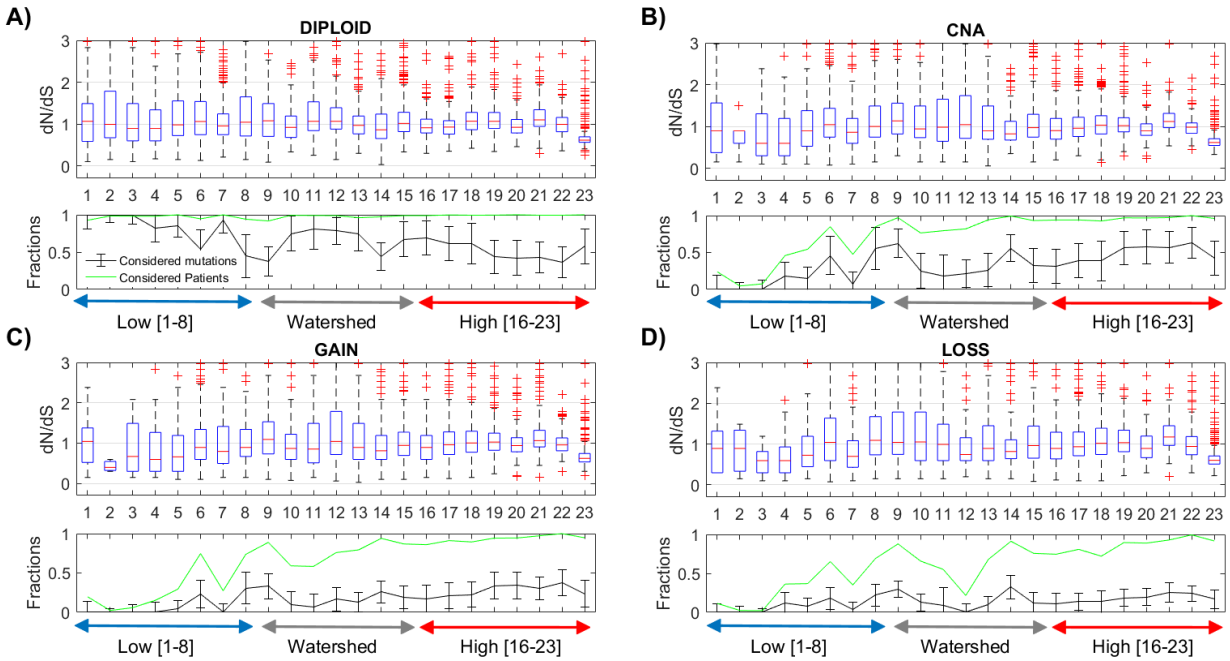

Figure S13: **dN/dS of patients in diploid and regions affected by copy-number alterations (CNA)**. We used the CNA gistic metric (Methods) to separate between mutations occurring in different regions, and re-estimated the selection *dN/dS* acting on the different regions, in each tumor. **A)** Distributions of *dN/dS* in samples based on mutations occurring in diploid regions (gistic=0). All cancers evolve near neutrality (except Melanoma). *dN/dS* can be quantified in almost all patients despite the lower number of mutations considered for evaluation (lower panel). **B)** Distributions of *dN/dS* based on mutations occurring regions affected by CNA (gistic≠0). All cancers evolve near neutrality (except Melanoma). Deviations from neutrality in low *ML* cancers are due to lose of statistical power (patients with *dN/dS=Inf* are discarded, leading to loss of *N* mutations a lowered *dN/dS*), as indicated in the lower panel. **C)** Distributions of *dN/dS* based on mutations occurring in gained regions (gistic>0). **D)** Distributions of *dN/dS* based on mutations occurring in gained regions (gistic<0). These regions also evolve similarly. Cancers are ordered as in **Figure 1**.

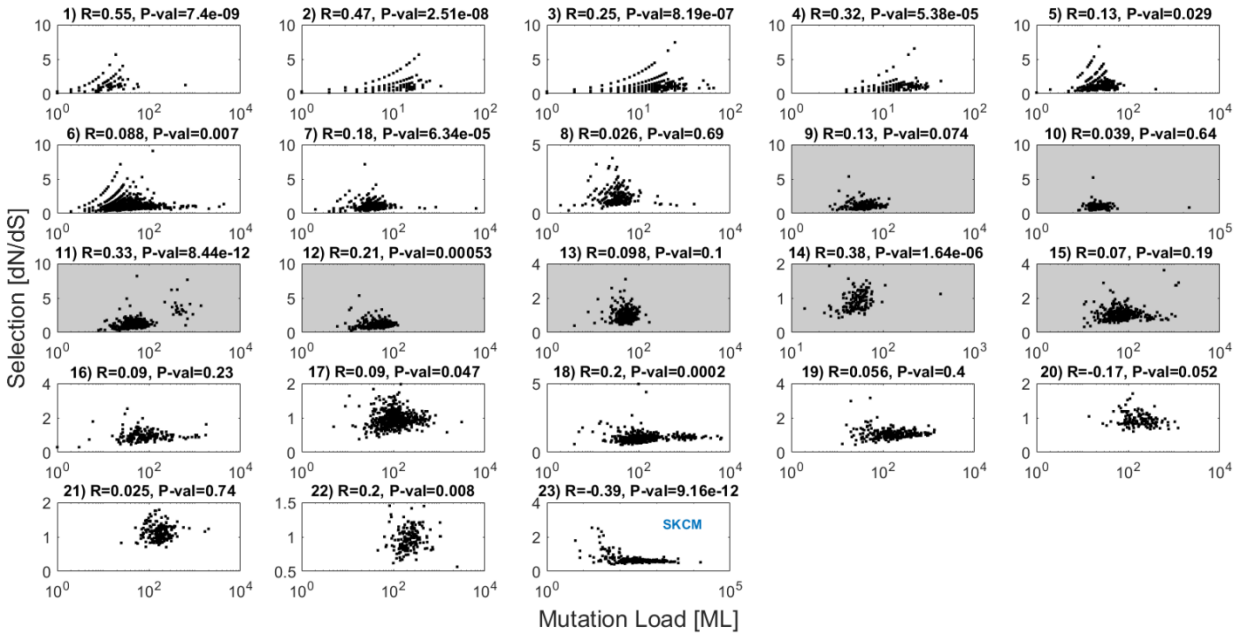

Figure S14: **Selection (dN/dS) in patients vs. the mutation load (ML).** Cancers show a diversity of relationships between *dN/dS* and *ML*. Melanoma (#23) is the only case that exhibits significant negative correlations. High *dN/dS* are not a consequence of low number of mutations, and are more frequently observed among patients that harbors large number of mutations. Cancers are ordered as in **Figure 1**.

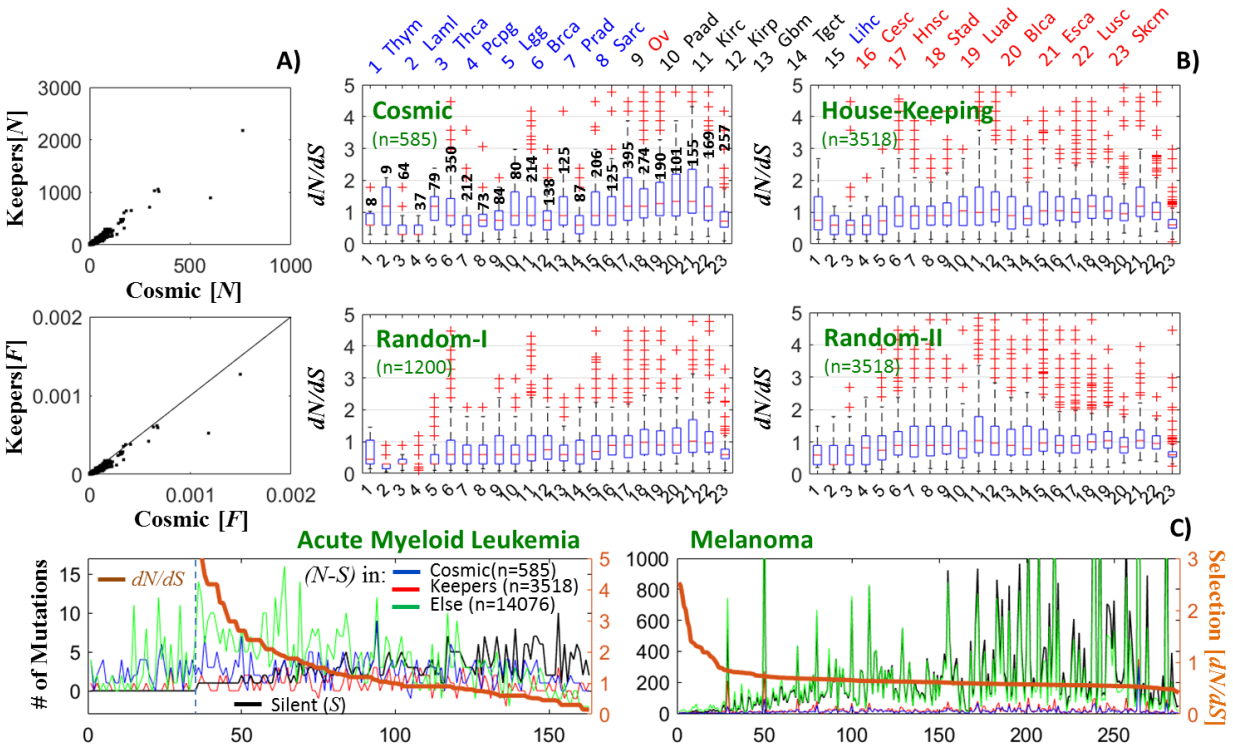

Figure S15: Distributions of mutations and selection (*dN/dS*) across different groups of genes. We evaluated *dN/dS* across the cohorts of patients based on the mutations in different groups of genes: Cosmic (i.e cancer-related) genes (n=585), house-keeping genes (n=3518; 145 genes that overlapped the Cosmic set were removed from the original set of 3663 house-keeping genes) and the rest of the proteome (n=14076). For each set of genes, we generated a corresponding randomized set (for the Cosmic genes, the same set as in **Figure S9** was used). **A)** Number of non-silent (*N*) mutations in Cosmic versus House-keeping genes across patients, showing the expected high correlation (*top*). Cosmic genes have higher mutation rate per unit length ($F = N/L$ where L is the length of the concatenated coding sequence of the set of genes) (*bottom*). **B)** *dN/dS* distributions across patients evaluated for cancer genes and house-keeping genes (*top panels*) and respective randomized sets of genes (*bottom panels*), shown for each cancer type (ordered as in **Figure 1**). Known cancer genes display a much higher *dN/dS* then the respective random set, across all cancer types. In low *ML* cancers, *dN/dS* in cancer genes could be evaluated only for a small number of patients (numbers shown in figure), while in the majority of patients cancer genes harbor only *S* mutations, such that *dN/dS=Inf* and is discarded from analysis. Hence, **cancer genes manifest signatures of positive selection across all cancers**. In contrast, *dN/dS* ratios in house-keeping genes are distributed around unity in most cancer types, similarly to the respective randomized set. **Thus, overall, selection acts differently across different parts of the proteome, with the sum of effects leading to neutrality in most cases except in Melanoma (Figure 2).** In Melanoma, the cancer with the highest *ML*, purifying selection prevails in the entire proteome, and acts on each of the examined group of genes (except for cancer genes). **C)** For convenience, the distribution of mutations shown in **Figure 3** for AML and Melanoma are displayed again.

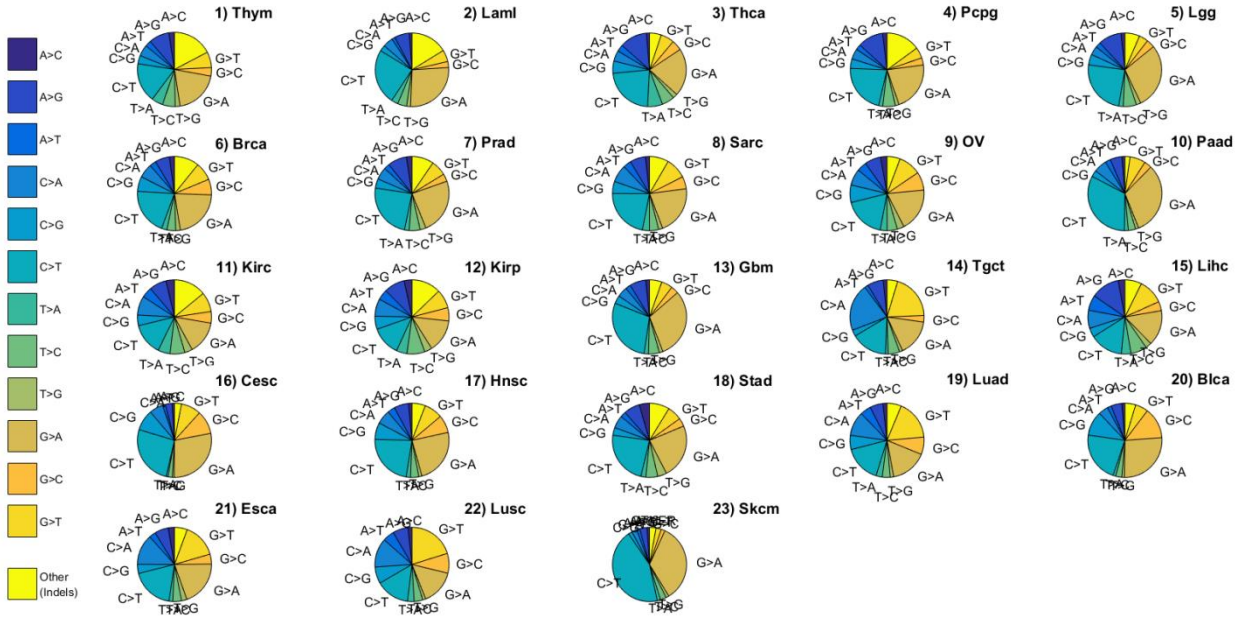

Figure S16: **Composition of mutation types in the 12 category.** The frequencies of the 12 contexts (A>x, C>x, T>x, G>x) of all mutations (*N* and *S*) in each tumor proteome were evaluated. Pie charts depict the average frequency of each context across patients, in each cancer type. The fraction of C>T/G>A mutations in Melanoma (#23) is substantial, and higher than in any other cancer type. Cancers are ordered as in **Figure 1**.

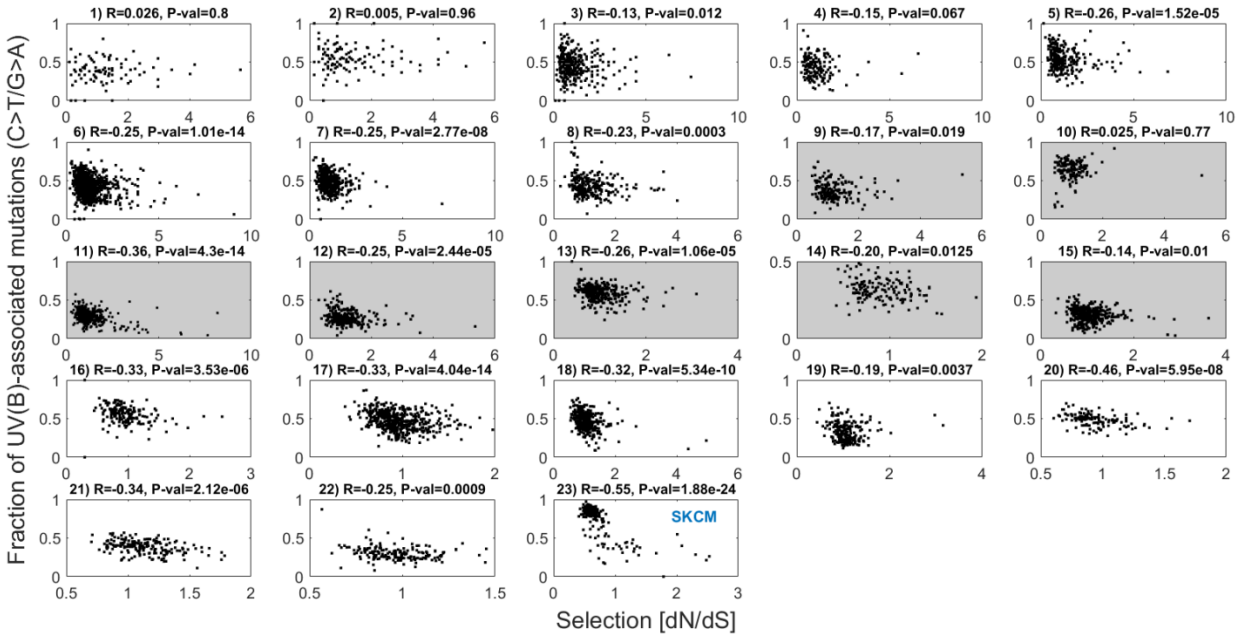

Figure S17: **Fraction of UVB associated mutations vs. selection (dN/dS) in patients.** The fraction of C>T/G>A mutations was evaluated in each patient. The relationship between these factions and the *dN/dS* values is weak in most low (and watershed) ML cancers. It becomes significant in high ML cancers. Despite this correlation all high ML have *dN/dS~1* (**Figure 2C**), except Melanoma (*dN/dS<1*) which also exhibits the highest and most significant negative correlation. Hence in Melanoma (#23), these mutations seem to affect *dN/dS* and may lead to an overestimation of the strength of negative selection. Cancers are ordered as in **Figure 1**.

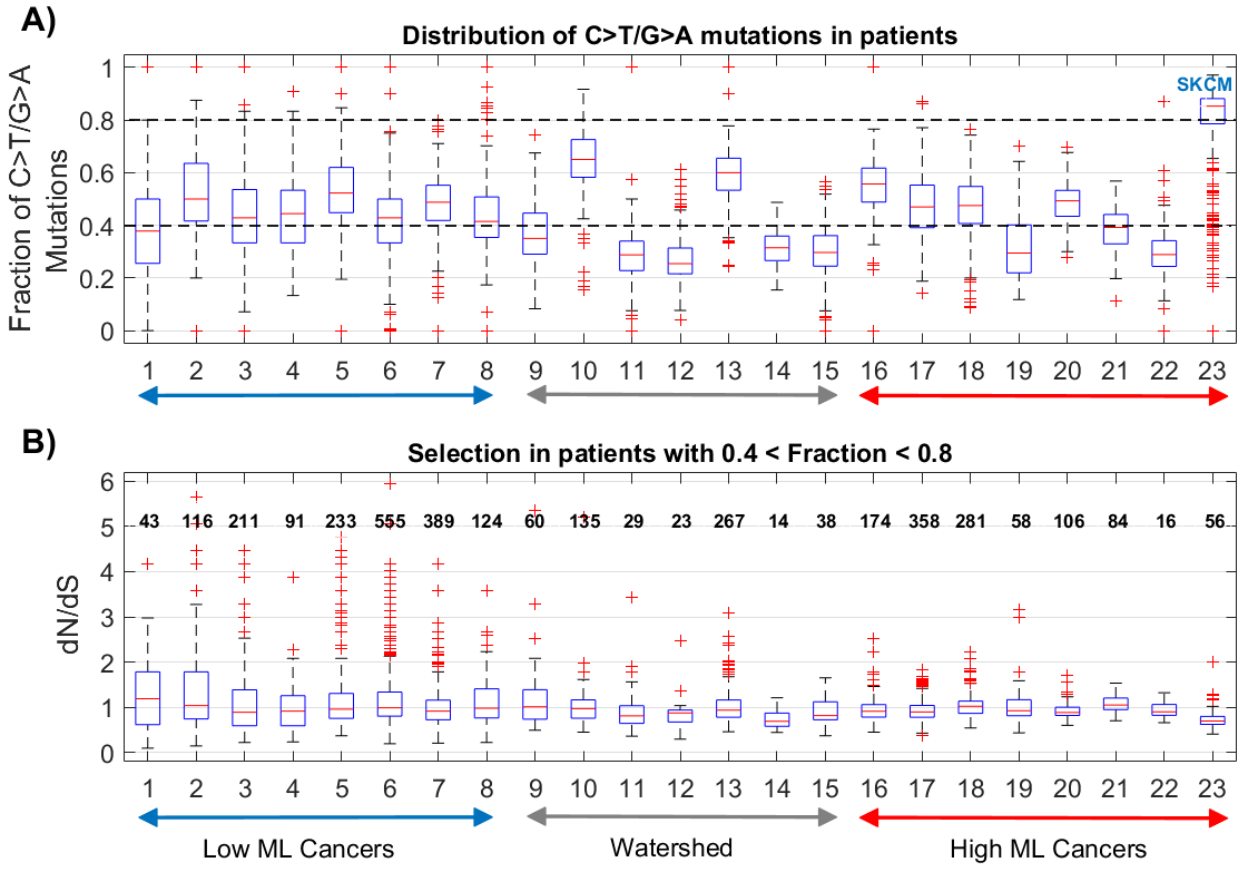

Figure S18: **Distributions of the fraction UV-associated mutations (C>T/G>A) in the proteomes of patients**. **A)** The distributions of the calculated fractions in patients' tumor proteomes are shown for each cancer type. Melanoma patients display the largest fractions compared with other cancer types (see also **Figure S16**). **B)** Distributions of *dN/dS* of patients with medium range of C>T/G>A fraction (40%-80%). While tumors of patients at this range (representing a large diversity) evolve near neutrality in all cancer types, Melanoma patients in this range (n=56) have *dN/dS<1*. Because UV-associated mutations may still be the primary driving force of Melanoma tumors (but not of other cancers); we further substantiated the existence of negative selection in Melanoma through extreme test (**Figure S19**). Cancers are ordered as in **Figure 1**.

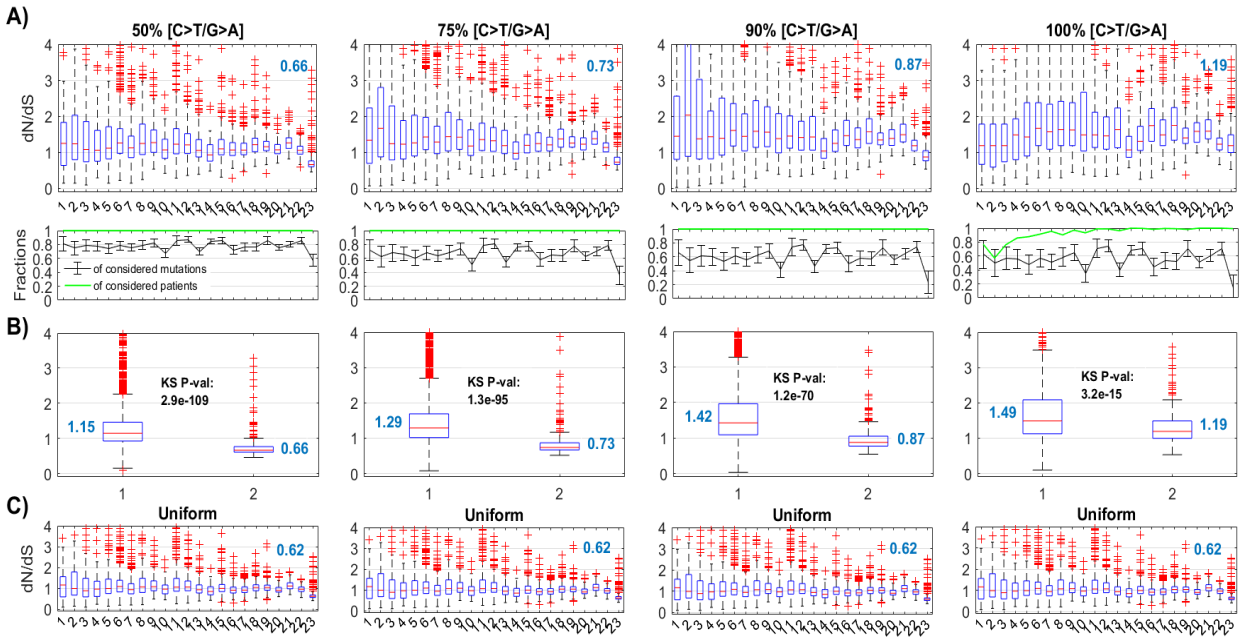

Figure S19: **Extreme test of the effect of removal of C>T/G>A mutations on the dN/dS of patients. A)** The distributions of dN/dS in patients after removal of 50%-100% (from left to right) of C>T/G>A mutations. Removing mutations biases *dN/dS* toward higher values (in all cancers). This because now we start with all mutations, and gradually remove them. *S* mutations are rarer than *N* mutations, such that their weight is underestimated following removal of mutations (i.e., at the edge of detection), leading to the increased *dN/dS* values (opposite from the effect in **Figure S3**). **Nonetheless, Melanoma patients consistently exhibit much lower *dN/dS*, indicative of negative selection.** Note that in the case of 100% removal of mutations the difference between *dN/dS* in Melanoma and other cancers appears to be reduced, because in some low ML cancers *dN/dS* cannot be quantified in many patients, as indicated in the lower panel (green). In these cases *S=0*, *dN/dS=inf*, such that many *N* mutations are discarded (similarly to the case of selection in genes; explained in **Figure S3**), leading to a lower attainable *dN/dS*. **B)** Distribution of *dN/dS* following mutation removal in the all cancers (PAN) vs. Melanoma. *dN/dS* in Melanoma is substantially lower than in pan-cancer data. The difference ($\Delta dN/dS \sim 0.5$) is stable (even in the extreme case of 100% removal of mutations, given the biases described above). **C)** Removing the exact number of mutations uniformly across the 12 contexts (A>x, T>x, G>x, C>x) Does not affect *dN/dS*. Cancers are ordered as in **Figure 1**.

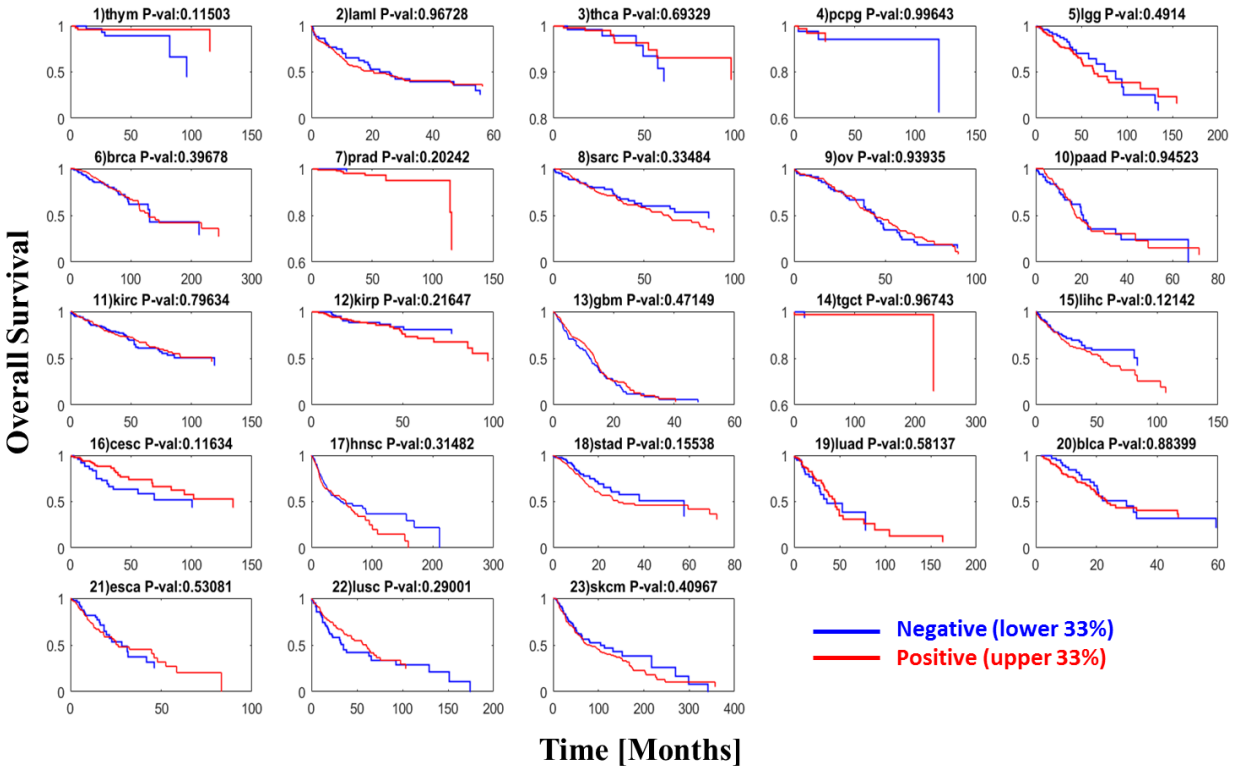

Figure S20: **KM analysis of positive versus purifying selection in each cancer type.** For each cancer type, overall survival was compared between patients with high *dN/dS* (upper 33%) indicating positive selection and patients with low *dN/dS* (lower 33%) corresponding to purifying selection. Cancer types are ordered by the mutation load (*ML*) as in **Figure 1** of the main text. No significant differences are observed, consistent with the Cox regression results shown in **Table 1** of the main text.

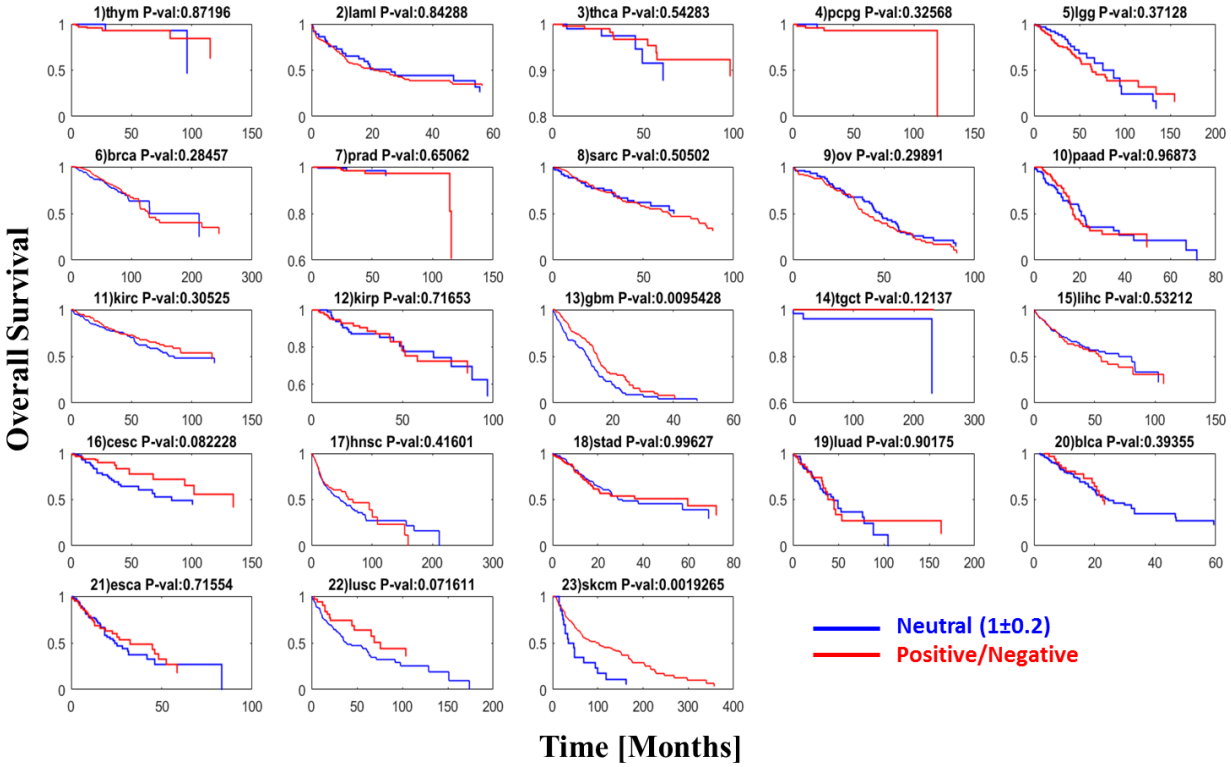

Figure S21: **KM analysis of neutral evolution versus positive/purifying selection for each cancer type.** KM analysis for each cancer type, comparing the overall survival (OS) between neutral evolution ($0.8<dN/dS<1.2$) and positive or purifying selection (i.e., $dN/dS>1.2$ and $dN/dS< 0.8$). Cancer types are ordered by the mutation load (*ML*) as in **Figures 1** of the main text. Significant differences are observed only in few a cases where neutral evolution leads to poorer prognosis (Gbm, Cesc, Lusc, Skcm). In Skcm, few patients follow neutral evolution (**Figure 2** of main text), yet their survival is much pooper than that for the rest of patients with this cancer type.

Tables

		OS		*DFS*	
Model_Variables	**Cancer Set**	**Beta (SE)**	***P*-value**	**Beta (SE)**	***P*-value**
U_ML	**L**	**3.48** (1.46)	**0.017**	**2.81 (0.89)**	**0.0015**
M_ML	L	**7.42** (2.34)	**0.002**	-8.73 (13.34)	0.51
M_Age	L	**3.73** (0.6)	**5.35e-10**	-0.03 (0.65)	0.96
M_Stage_II	L	**0.47** (0.28)	0.09	0.14 (0.27)	0.60
M_Stage_III	L	**1.30** (0.29)	**5.6e-06**	**1.07** (0.26)	**3.28e-05**
M_Stage_IV	L	**2.18** (0.35)	**6.3e-10**	**1.83** (0.36)	**4.28e-07**
M_CNV	L	**1.32** (0.51)	**0.001**	**0.97** (0.57)	**0.086**
Chi-square (6 dof) ***P-value*:**		**0**		**0**	
U_ML	**H**	**-4.79** (1.73)	**0.0057**	**-4.18** (1.73)	**0.0155**
M_ML	H	**-8.87** (3.26)	**0.0066**	**-8.62** (3.84)	**0.025**
M_Age	H	**2.01** (0.44)	**4.55e-06**	0.77 (0.5)	0.12
M_Stage_II	H	0.08 (0.21)	0.7	0.1 (0.21)	0.65
M_Stage_III	H	**0.6** (0.19)	**0.0021**	0.27 (0.21)	0.2
M_Stage_IV	H	**1.06** (0.21)	**4.50e-07**	**0.86** (0.24)	**0.00025**
M_Grade_II	H	0.35 (0.22)	0.10	**0.45** (0.25)	0.073
M_Grade_III	H	**0.41** (0.22)	**0.066**	0.55 (0.25)	**0.032**
M_Grade_IV	H	-11.16 (401)	0.98	-11.12 (428)	0.98
M_CNV	H	-0.22 (0.34)	0.52	0.18 (0.38)	0.64
Chi-square (9 dof) ***P-value*:**		**0**		**0**	

Table S1: Univariate versus multivariate stratified Cox regression analysis of the *ML* and confounding factors. For each tested variable the estimated scaling coefficient β (i.e., $HR = e^{\beta}$), its standard error (SE) and the corresponding *P*-value of the stratified (by cancer type) Cox regression model are shown for overall survival (OS) and disease free survival (DFS). Univariate (U) results of **Table 1** of main text are bolded and colored (β>0 blue; β<0 red) and are compared to the results of a multivariate (M) Cox model, considering Age, Stage, Grade and overall CNA as confounding factors. Analysis is done for each of the 2 sets of cancers, corresponding to low (L) and (H) mutational load (*ML*) cancer types (see **Figure 1**). In each test, variables are normalized to 0-1 within each group of cancers (**Methods**). Results of chi-square statistics, inferred by the difference in the log likelihoods of the Univariate versus Multivariate models, are shown below each test. Cancer grade is not available for the low *ML* cancers type (**Figure S1**); hence, it was used as a confounding factor only in set H. The results verify *ML* to be the only variable which captures the transition in clinical outcome.

Additional Data Table S1 (separate file)

Supplemental data, provided as Microsoft Excel file, contains the samples, genes, number of mutations and survival times, allowing fully reproducibility of the results this study.