Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/7/1/eabc2100/DC1

Supplementary Materials for

Synthetic lethality across normal tissues is strongly associated with cancer risk, onset, and tumor suppressor specificity

Kuoyuan Cheng, Nishanth Ulhas Nair, Joo Sang Lee, Eytan Ruppin*

*Corresponding author. Email: eytan.ruppin@nih.gov

Published 1 January 2021, *Sci. Adv.* **7**, eabc2100 (2021)
DOI: 10.1126/sciadv.abc2100

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/7/1/eabc2100/DC1)

Contents

Supplementary Notes

Supplementary Figures

figure S4. Reproduced result of Tomasetti & Vogelstein[2] and Klutstein et al.[6]

figure S5. The cSL gene pairs used in this study are more specific to cancer and have much weaker synthetic lethal effect to the normal tissues.

figure S6. Spearman correlation between Tissue cSL load and cancer onset age computed using different thresholds.

figure S7. Randomized control analysis for the correlation between tissue cSL load and cancer onset age using pseudo-cSL load.

figure S8. Expression of the cSL partner genes of tumor suppressor genes (TSGs) by each TSG.

figure S9. Correlation between tissue cSL load computed with experimentally identified cSLs and lifetime cancer risk across tissues.

figure S10. Correlation between cSL load and lifetime cancer risk is independent from stem cell division.

figure S11. Correlation between cSL load and lifetime cancer risk or cancer onset age is not confounded by the number of samples available for each tissue.

figure S12. Correlation between cSL load and lifetime cancer risk is not confounded by the number of poised genes in each tissue.

figure S13. Correlation between cSL load and lifetime cancer risk is not confounded by the expression levels of oncogenes in each tissue

figure S14. Correlation between cSL load and lifetime cancer risk or cancer onset age is not confounded by the abundance of immune cells or fibroblasts in each tissue.

figure S15. The case of leukemia where tissue cSL load is correlated with cancer risk by age.

Supplementary Tables

table S1. ISLE-inferred and experimentally identified cSL gene pairs

table S2. Tissue cSL load in the older population and its correlation with lifetime cancer risk

table S3. Correlation between tissue cSL load in the younger population and cancer onset age

table S4. Highly specific vs lowly specific cSLs_the TCLs computed from them and correlation with cancer risk and onset age

table S5. Pathways specifically enriched by the highly specific cSLs in each tissue

table S6. Tissue-specific tumor suppressor genes and their cSL partner genes

Supplementary Notes

1. Details of the cSL gene pairs

The cSLs obtained from Lee *et al.*[22] were computationally identified with the ISLE (identification of clinically relevant synthetic lethality) pipeline, which mined data from 8749 cancer TCGA patient tumors across 28 cancer types and applied a 4 step procedure to identify the cSL network. The pipeline starts with an initial set of experimentally identified cSL interactions either by double knock-out/knock-down in a single cell line or by inference from the single RNAi/CRISPR screens in a large number of cancer cell lines. Then, it analyzes patient tumor data across the 28 cancer types to further filter those pairs that show the evidence of (i) negative selection of and (ii) better patient survival due to co-inactivation while controlling for the effect of individual genes, followed by (iii) a phylogenetic screen (described in detail in Lee *et al.*[22]). We used the cSL network identified with FDR < 0.2 and FDR < 0.1 as described in the Methods. Experimentally identified cSLs were obtained by pooling the experimentally identified cSL pairs from 17 double knock-down/knock-out screens compiled in Lee *et al.*[22] and combining it with the cSL pairs identified with CRISPRi experiments by Horlbeck *et al.*[34] as described in the Methods. The merit of using ISLE-inferred cSL pairs over the experimentally identified cSLs is that the selected cSL pairs may be more likely to be clinically relevant across many cancer types and less likely to be cancer-type specific.

2. Effect size measure of Wilcoxon tests and multiple testing

Throughout the text we measured the effect size of the Wilcoxon tests by the rank-biserial correlation, which ranges from -1 and 1. Strong effects are represented by rank-biserial correlations whose absolute values are close to 1, and weak effects represented by values close to 0. The exact formula for calculating the rank-biserial correlation can be found in Wendt *et al.*[39] We corrected for multiple testing using the Benjamini-Hochberg method throughout our study.

3. Robustness of the correlation between tissue cSL load and lifetime cancer risks, and cancer onset age

Alternative cSL load measures

We used various alternative approaches to compute cSL load and correlated them with lifetime cancer risk in order to show the robustness of our findings. These alternative approaches are described below:

a) We computed the correlation between lifetime cancer risk and cSL load (for older populations) using a more stringent FDR threshold (FDR < 0.1 instead of FDR < 0.2) for selecting the cSL network using ISLE (leading to 2326 pairs). We still get a significant correlation (Spearman's $\rho = -0.613$, $P = 0.000668$, fig. S1a).

b) We removed genes that have zero expression in more than 90% of all GTEx samples. We then computed cSL load (for older populations) with the remaining genes, and found a significant correlation with lifetime cancer risk (Spearman's $\rho = -0.605$, $P = 0.000838$, fig. S1b). Here the ISLE-inferred cSLs with FDR < 0.2 were used.

c) For our main results presented in Fig. 2 of the main text, tissue cSL load (TCL) was computed by taking the median (i.e. the 50 percentile) cSL load of the samples from a tissue. We additionally computed TCL by taking the 65 percentile or the 40 percentile of the cSL loads in the samples (age ≥ 50 years) of a tissue. TCLs obtained as such are also significantly correlated with lifetime cancer risk ($\rho = -0.794$, $P = 7.59e\text{-}7$, fig. S1c; and $\rho = -0.52$, $P = 0.0054$, fig. S1d, respectively).

d) We have defined a cSL gene pair to be inactivated if the expression levels of both genes in the cSL pair are below the 33 percentile within the population (i.e. each tissue type). We also used the definition of inactivated cSLs based on other percentiles, including 40 percentile, 20 percentile, and 10 percentile. The corresponding correlations with lifetime cancer risk are shown in fig. S1e-g, which remains significant. We also computed cSL load by considering only zero-expressed gene pairs, the correlation is further weaker and marginally significant (Spearman's $\rho = -0.313$, $P = 0.11$, fig. S1h). The issue with using very low gene expression cutoffs is that it will lead to very few cSL gene pairs where both genes defined as "inactivated" by that cutoff, and by doing this although we may reduce noise via keeping only the cSL pairs with strong effects, we also lose a lot of signal contributed by the additive effects of many more weaker cSL pairs.

Taken together, these results show that tissue cSL load is robustly (negatively) correlated with lifetime cancer risk across tissues.

Alternative mappings between normal and cancer tissue types

In the main result as presented in Fig. 2 of the main text, we used the mapping between 16 GTEx normal tissues to 27 cancer types as given in table S2a. However certain mappings were not exact due to the limitations of the data available, for example, salivary gland was the closest GTEx tissue type in terms of anatomical proximity to head and neck carcinoma (with or without HPV-16), but histologically this mapping may not be valid. Therefore we ensured that our observed correlation between tissue cSL load (TCL) with cancer lifetime risk is not affected by including or removing this mapping (Spearman's rho = -0.664 vs -0.662, P = 0.0002 vs 0.0003, fig. S1i). As another example, we additionally removed the mapping between "Brain - Cerebellum" in GTEx and Medulloblastoma, and still get a good correlation between TCL and risk (Spearman's rho = -0.623, P = 0.00067, fig. S1j). This was done because we had normalized (inverse normal transformation across samples and genes) each GTEx tissue individually and all the brain tissue samples were normalized together. However for the Medulloblastoma to normal tissue mapping only a portion of the brain samples corresponding to Brain - Cerebellum was used for computing tissue cSL load. So the mapping between "Brain - Cerebellum" in GTEx and Medulloblastoma was removed to show robustness.

Alternative cancer onset age definition

We computed the onset age for each cancer type from SEER incidence data[1]. We mapped the cancer types in SEER to the tissue type in GTEx, given in table S3. We then computed TCL as described above for the samples below 40 years old in each GTEx tissue, and correlated the TCL with cancer onset age across tissues (for the result in Fig. 3 of main text). TCLs were also computed for other upper cutoffs of ages (including 45 and 50 years old), and their correlations with onset age are shown in fig. S6, in order to show that the result is robust to the parameters used. We find that in general, TCL for younger populations correlates positively with cancer onset age in a robust way. This is consistent with the hypothesis that the lower the TCL for a tissue, the earlier the cancer onset time, thus showing that tissue cSL load can impede cancer development.

4. Randomized control analysis

We did three types of random control analysis to test that the various observed effects in this study are specific to the cSL gene pairs and are not random. The three random controls were performed using: (i) random gene pairs; (ii) shuffled cSL gene pairs; and (iii) degree-preserving randomized cSL network (same size as the actual cSL network). Below we describe the detailed procedure testing for the cSL load correlation with cancer risk as an example.

(i) Random gene pairs: we randomly sampled 20171 gene pairs from all the genes in the GTEx data and computed a tissue "pseudo-cSL load" with these random gene pairs, for each tissue for the older population (≥ 50 years), and then computed its correlation with lifetime cancer risk. This procedure was repeated for 1000 iterations and a histogram showing correlation distributions are shown in fig. S3a. We see that the correlations are not significant (median Spearman's $\rho = -0.21$, $P = 0.29$). An empirical p-value was computed using the correlation obtained from the actual cSL network and the correlation distribution of the tissue pseudo-cSL loads ($P < 0.001$, fig. S3a), indicating that the signal from the true cSL network is always much stronger than random signals.

(ii) Shuffled cSL gene pairs: starting from the 20171 cSL gene pairs listed in a table (20171-by-2, each row is a gene pair), we randomly shuffled the columns of the table to obtain a list of shuffled gene pairs, and computed a tissue pseudo-cSL load for each GTEx tissue for older populations (age ≥ 50 years). This was correlated with cancer lifetime risk. This procedure was repeated for 1000 iterations and a histogram showing correlation distributions are shown in fig. S3b. We see that many of these random correlations are significant (median Spearman's $\rho = -0.537$, $P = 0.0039$). This can be due to an effect on the single gene level, which we described in the main text, and showed that the genetic interaction effect actually is the dominant one (vs the single-gene effect, fig. S3d-g). For more details on the single-gene effect analysis, see the section "*Details on the single-gene effect analysis*" below. Additionally, the empirical p-value for the randomization test computed as above is $P < 0.001$ (fig. S3b), indicating that the signal from the true cSL network is still the strongest.

(iii) Degree-preserving randomized cSL network: starting from the actual cSL network consisting of 20171 cSL gene pairs (edges), we performed a degree-preserving

randomization, rewiring the network but retaining the original degree distribution. Tissue pseudo-cSL load was computed as above and correlated with cancer lifetime risk. This procedure was repeated for 1000 iterations and a histogram showing correlation distributions are shown in fig. S3c. The random correlations have a median value of Spearman's $\rho = -0.537$, $P = 0.0039$, which as elaborated above and further explored in the main text, can be due to the single-gene effect (fig. S3d-g). Nevertheless, again the empirical p-value is $P < 0.001$ (fig. S3c), indicating a strong and robust cSL-specific effect.

Similarly, such control tests were performed for the correlation with cancer onset age (fig. S7) and experimentally identified cSLs (fig. S9). We also performed control tests for the analysis of the tissue type-specificity of tumor suppressor genes (TSGs). As shown in Fig. 4 of the main text, the true cSL partner genes of a TSG tend to have higher expression in the tissue type(s) where the TSG is a known driver and the rest of the tissue types where the TSG is not an established cancer driver, supporting our hypothesis that the tissue type-specificity of TSGs can be (partly) explained by cSL. Specifically, out of the 23 cases of TSGs, 17 of them show such a trend, and only 6 cases show the opposite trend but with much weaker effect sizes and P values. We simply uses the difference in the number of cases on both sides as a test statistics (with the true cSL pairs, the value is 17-6=11), with a larger value of the statistics indicating a stronger evidence supporting our hypothesis. For each TSG, given its true cSL partner genes identified by the ISLE method, we randomly sampled the same number of genes from all the genes as its "pseudo-partner genes", and we analyzed the differential expression (DE) of these pseudo-partner genes between the tissue type(s) where the TSG is a known driver and the rest of the tissue types where the TSG is not an established cancer driver using linear model, as described in Methods. Summarizing the results of all TSGs, the test statistics was calculated. After repeating the random sampling, we computed the empirical P value based on the fraction of times that the statistics obtained using pseudo-partner genes is greater than that obtained using the true cSL partners. We also performed another control test by randomly shuffling the mapping between the TSG and the tissue types where they have established driver functions, and the empirical P value was computed in a similar fashion. Both control tests yielded empirical $P < 0.05$. Additionally, we also performed the control tests as above for each of the TSG, with the test statistics being the linear model

coefficient, which represents the mean difference in the expression level of the cSL partner genes between the tissue(s) of the TSG and the tissues where the TSG is not a known driver, and the empirical P values obtained for each TSG is given in fig. S8.

5. Details on the single-gene effect analysis

As we described above and in the main text, some of the random control tests yielded significant signals in terms of, e.g. correlation with cancer risk or onset age, although the signals from the actual cSL gene pairs are always much stronger. Notably, it is the shuffled cSL gene pairs or the degree-preserving randomization that produced significant signals, but not the purely random gene pairs. We therefore reasoned that the genes within the cSL network (*cSL genes*) may have an effect on the single gene level, and although the cSL load is meant to capture the pairwise cSL genetic interaction effect, it actually also captures some of the single-gene effect due to the way it is computed. Accordingly we computed the *tissue cSL single gene load (SGL)* as described in the main text, and confirmed the single-gene effect wrt cancer risk (fig. S3d) and onset age (fig. S7e). Interestingly in both cases, the single-gene effect is only exhibited by the cSL genes, and no significant signal can be found using random sets of genes or all genes (fig. S3e-f, S7f), which could be due to additional weak genetic interaction effects among the cSL genes not captured by the current cSL network. Further, as described in the main text, we examined the partial correlation between tissue cSL load (TCL) and risk/onset age after controlling for the SGL, and showed that it is the cSL genetic interaction effect that dominates the observed correlations (fig. S3g, S7g). Specifically, we regressed out the SGL component from both TCL and cancer risk/onset age with linear regression, then correlated the residues. This is the standard procedure of partial correlation and is equivalent to a multiple regression model for cancer risk/onset (as the dependent variable) with both SGL and TCL as the independent variables (covariates). We adopted the partial correlation procedure here (instead of the multiple regression) since it gives the correlation value, which is easier to compare to the rest of our results.

6. Reproducing the result of Tomasetti & Vogelstein[2] and Klutstein *et al.*[6]

For our analysis, only a subset of the cancer types in Tomasetti & Vogelstein 2015[2] have matched normal tissue types available from the GTEx database (table S2a), as

explained in the Methods. To facilitate comparison, we reproduced the result of Tomasetti & Vogelstein 2015[2] that there is a strong association between the number of tissue stem cell divisions with cancer lifetime risk only on this subset of tissue types (Spearman's ρ = 0.717 , P = 2.57e-5, fig. S4). As a result of the drop-out of cancer types, this result is numerically different from that reported in their original manuscript[2] (Spearman's ρ = 0.81; P = 3.5e-8). Similar situation exists for reproducing the result of Klutstein *et al.*[6] (fig. S4).

7. Computing the level of negative selection of cSL gene pairs

For each gene, we define that it's lowly expressed in a sample if its expression level is below the 33 percentile of those in all samples of the same tissue type, which is consistent with the approach in Lee *et al.*[22] and as elaborated in the main text Methods. For each cSL gene pair, we computed the fraction of samples where both genes are lowly expressed for each of the normal and cancer tissue types, which were compared to those computed with random gene pairs with Wilcoxon rank-sum tests. Rank-biserial correlation, the effect size metric of Wilcoxon test was used to measure the level of negative selection, since effectively this metric will take zero values on average for random gene pairs, and negative values for gene pairs whose co-downregulation in a tissue is selected against. Also, a smaller negative value (i.e. larger absolute value) indicates a stronger negative selection.

8. Details on the analysis of highly specific vs lowly specific cSLs (hcSLs and lcSLs)

Note that for each tissue type, a distinct set of hcSLs and lcSLs were identified. The sets of hcSLs of the tissues have pairwise Jaccard indices ranging from 0.136 to 0.251 with a mean value of 0.163, which represents significantly more overlap compared to random sets of the same size ($P < 0.001$ in a randomization test for the mean Jaccard index). For the set of hcSLs and lcSLs for each tissue, we obtained the unique genes within the cSL pairs and identified the pathways they are enriched in as described in Methods. We focused on the pathways that are specifically enriched for the hcSL genes rather than the lcSL genes, since as we have shown the hcSLs are the ones most relevant in cancer risk.

9. Contribution of cSL to the tissue-specificity of tumor suppressor genes (TSGs)

A recent study reported that the expression levels of the genetic interaction (including cSL) partners of cancer driver genes differ across cancer types, in a way that is dependent on the tissue-specific functional status of the driver gene[32]. Their analysis was performed in cancer tissues, and thus does not directly support the role of cSL in cancer development. Our approach is different, however, in that we investigated the expression levels of the cSL partners of the TSGs in normal, non-cancerous tissues, and we sought to establish the link between the co-inactivation status of cSL gene pairs in normal tissues and the tissue type-specificity of carcinogenesis.

10. The correlation between tissue cSL load and lifetime cancer risk is not confounded by the number (or rate) of normal tissue stem cell division

Klutstein *et al.*[6] have suggested a link between abnormal DNA methylation and DNA replication in stem cells, consistent with the strong correlation they observed between LADM and NSCD. On the contrary, we showed in the main text that tissue cSL load (TCL) significantly adds to either NSCD or LADM in terms of the correlation with lifetime cancer risk and is likely not a corollary of tissue stem cell divisions. We further performed partial correlation of TCL and lifetime cancer risk, conditioned on either the rate of tissue stem cell division or the number of stem cells residing in each normal tissue (both obtained from Tomasetti *et al.*[2]) and found that strong correlations remain (fig. S10a,b). We also computed the proliferation index (PI) for each of the samples in the GTEx normal tissue from gene expression using a proliferation signature named meta-PCNA[40]. PI was computed as the median expression values of the set of meta-PCNA genes defined in Venet *et al.* 2011[40], and is regarded as a proxy for the rate of cell proliferation within a sample. We then took the median value of the PI's for all samples of a tissue type, and correlated the median PI of a tissue with its lifetime cancer risk (across tissues), showing that there's no significant correlation (fig. S10c). Also, we note that stem cell proliferation *rates* (rather than the total number of stem cell divisions), as estimated by Tomasetti *et al.*[2], correlates only weakly with lifetime cancer risk (Spearman's ρ = 0.386, P = 0.046).

11. Association of lifetime cancer risk with tissue cSL load controlled for the number of poised genes

To investigate whether the number of poised genes in the tissues may underlie the observed correlation between TCL and lifetime cancer risk, we obtained the "core 15-state model" chromatin state data from the ROADMAP epigenomics project[42]. We focused on human tissue samples, where the tissue type is also contained in the GTEx data. These tissue samples include liver, brain angular gyrus, brain anterior caudate, brain cingulate gyrus, brain germinal matrix, brain hippocampus middle, brain inferior temporal lobe, brain dorsolateral prefrontal cortex, brain substantia nigra, colonic mucosa, sigmoid colon, esophagus, lung, ovary, pancreas, small intestine, and stomach mucosa. For each of these tissue samples, we identified genes that are associated with the bivalent/poised transcription starting site (i.e. type 10_TssBiv) using the *matchGenes* function from the R package *bumphunter*[43], and used these as tissue-specific poised genes. For the normal tissues where we have such poised gene information, we re-computed the correlation of TCL (age ≥ 50) and lifetime cancer risk. As expected, we find a significant negative correlation as before (Spearman's ρ = -0.707, P = 0.0032, fig. S12a). We then control for with the number of poised genes in each tissue, by regressing out the number of poised genes from TCL and correlating the residuals with lifetime cancer risk. We see a similar association between lifetime cancer risk and TCL controlling for the number of poised genes (Spearman's ρ = -0.67, P = 0.0061, fig. S12b). This suggests that the observed correlation between TCL and lifetime cancer risk across human tissues is not dominant by the effect due to tissue-specific poised genes.

12. Association of lifetime cancer risk with tissue cSL load after removing oncogenes

To test whether TCL is predictive of lifetime cancer risk after removing oncogenes, we removed all cSL pairs containing oncogenes (the list of Tier 1 and Tier 2 oncogenes is taken from COSMIC[11]), and recomputed cSL load. We still see a strong negative correlation of cSL load and cancer risk in older populations (Spearman's ρ = -0.64, P = 0.000347, fig. S13a). We also see a strong positive correlation between cSL load and

cancer onset age in younger populations (Spearman's ρ = 0.47, P = 0.0171, fig. S13b). These correlations are quite similar to what we obtained earlier (i.e. without removing oncogenes). We note that these results support our currently proposed interpretation that the association between tissue cSL load and lifetime cancer risk is due to the overall additive synthetic lethal/sickness effect of all the cSL gene pairs, rather than being driven by a particular small subset of genes.

13. Association of lifetime cancer risk with tissue cSL load after controlling for immune and fibroblast/stromal cell abundance

We estimated the immune cell and fibroblast abundances for each of the GTEx samples from gene expression data using the ESTIMATE algorithm[41] then obtained the median immune cell or fibroblast/stroma abundance levels for each tissue type. We then control for the immune cell or fibroblast abundance in a linear model while checking the association between TCL and lifetime cancer risk or cancer onset age. We still see significant negative associations between lifetime cancer risk and TCL controlling for predicted immune cell abundance (Spearman's ρ = -0.66, P = 0.0002, fig. S14a); between lifetime cancer risk and TCL controlling for predicted fibroblast/stroma abundance (Spearman's ρ = -0.57, P = 0.002, fig. S14b); and between lifetime cancer risk and TCL controlling for both predicted immune and fibroblast abundance (Spearman's ρ = -0.65, P = 0.00025, fig. S14c). We also see significant positive associations between cancer onset age and TCL controlling for predicted immune cell abundance (Spearman's ρ = 0.48, P = 0.014, fig. S14d); between cancer onset age and TCL controlling for predicted fibroblast abundance (Spearman's ρ = 0.52, P = 0.0081, fig. S14e); and between cancer onset age and TCL controlling for predicted immune and fibroblast abundance (Spearman's ρ = 0.45, P = 0.0245, fig. S14f).

14. Tissue cSL load variation within tissue in blood and its association with age-specific cancer risk in leukemia

To investigate the possibility of the association TCL with age-specific cancer risk within a tissue (instead of across tissues), we focused on leukemia (SEER tissue site recode name "Leukemia", GTEx tissue type name "Blood"). We note that although

leukemia can happen both at very young and very old ages, as shown in the plot below the older age group has much higher risk than the younger age group (fig. S15a).

In the GTEx blood samples, we identified 1144 genes within the SL network whose expression changes with age (Wilcoxon rank-sum test adjusted $P<0.01$); comparing GTEx blood samples to the TCGA LAML leukemia samples, we identified 2068 differentially expressed genes independent of age (controlling for age with linear model, adjusted $P<0.01$). These two sets of genes do not have significant overlap (Fisher's test $P=0.47$). We computed the cSL load using the GTEx blood samples for every individual. GTEx data has normal non-cancerous individuals between 20-80 years and their age information is provided as a range of 10 years, therefore we computed age-specific TCL in bins of 10 years (20 to 29, 30 to 39 etc.). Correlating these age-specific TCLs with the corresponding age-specific leukemia risk from the SEER data, we find a strong negative correlation (Spearman's $\rho = -0.943$, $P = 0.0167$, fig. S15b). These results seem to provide evidence that cSL load is not only associated with lifetime cancer risk across tissues, but also may partly account for the variation in cancer risk by age in leukemia. However a deeper study is required before ascertaining whether cSL load can indeed be predictive age-specific cancer risk within leukemia, and since it is beyond the scope of this current study, it can be looked at in a future analysis.

However, we note that we further checked and found that the above correlation between age-specific cSL load and age-specific cancer risk is not consistently significant for all tissue/cancer types beyond blood/leukemia. This suggests that the potential role of cSL load in cancer risk variation by age within each tissue can be independent from its role in the variation of lifetime cancer risk across tissues, as different factors can contribute to these two kinds of variations.

Supplementary Figures

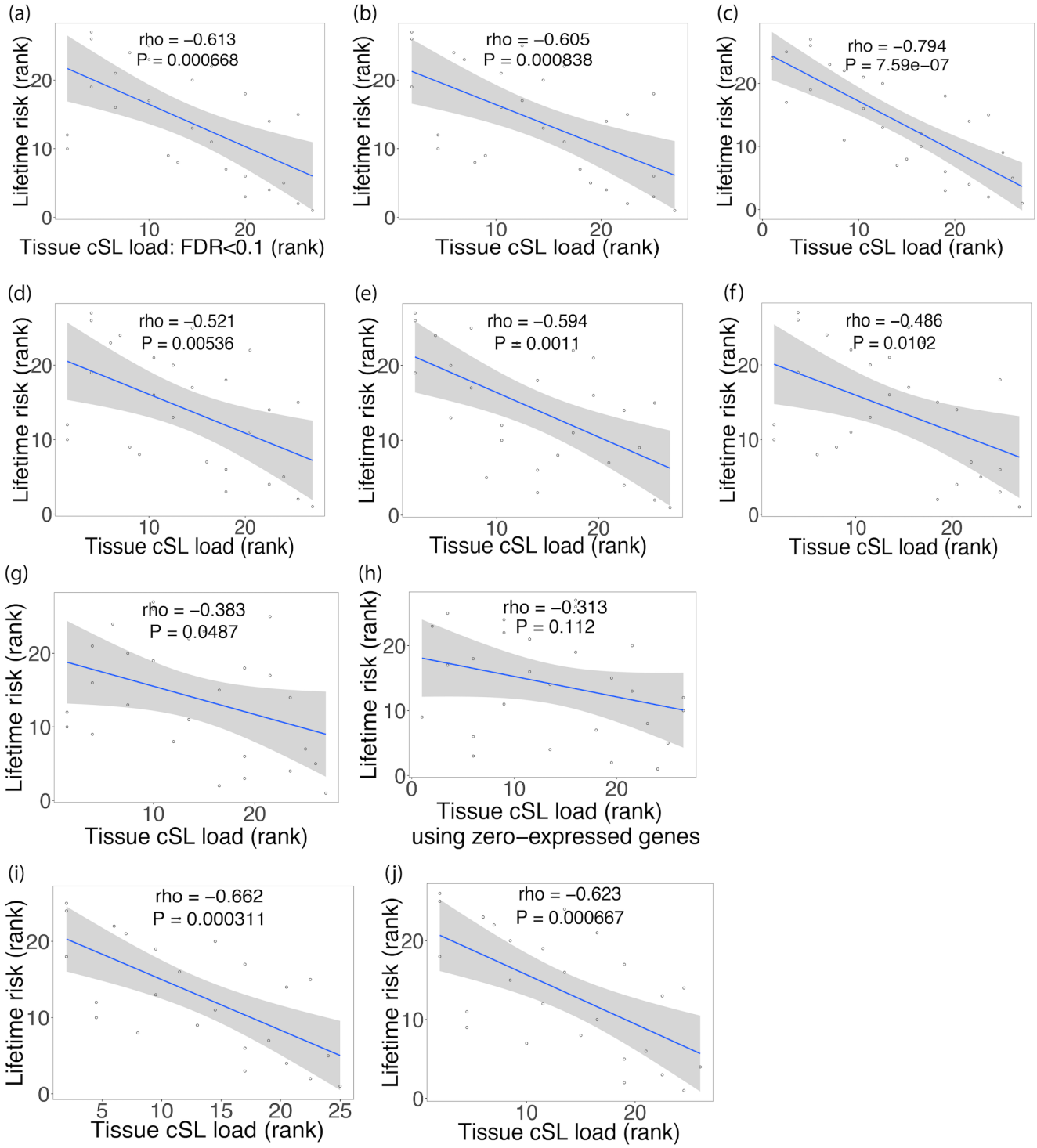


figure S1. Correlation between tissue cSL load and lifetime cancer risk across tissues is robust to variations to the computational method used. *Scatter plots*

showing Spearman correlations (rho) between cancer lifetime risk and: (a) cSL load (computed using FDR<0.1); (b) cSL load (computed by removing genes which have zero expression in more than 90% of the samples); (c) cSL load (by taking 65 percentile cSL load value across samples for each tissue, instead of median value); (d) cSL load (by taking 40 percentile cSL load value across samples for each tissue, instead of median value); (e) cSL load (using low expression threshold as less than 40 percentile); (f) cSL load (using low expression threshold as less than 20 percentile); (g) cSL load (using low expression threshold as less than 10th percentile); (h) cSL load (considering only genes with zero expression); (i) cSL load (removed 2 mappings between Head & neck squamous cell carcinoma and salivary gland); (j) cSL load (removed data point which mapped Medulloblastoma to Brain-Cerebellum). cSL analysis was done on older populations (age ≥ 50 years) for all sub-figures.

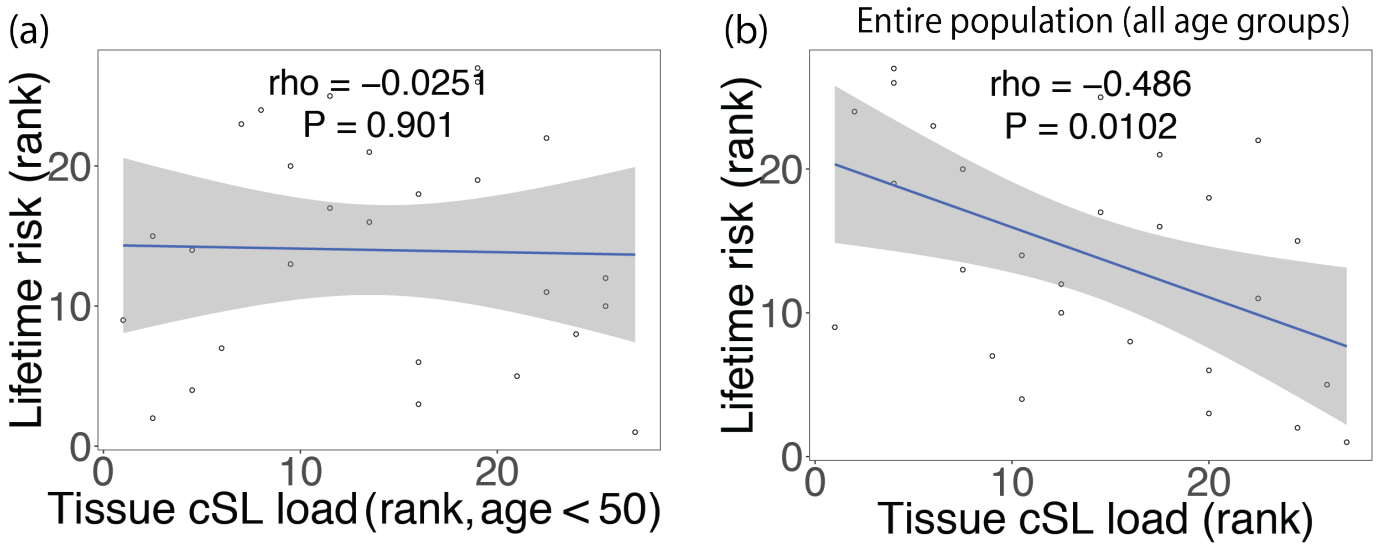


figure S2. Correlation between tissue cSL load computed with different age ranges and lifetime cancer risk. *Scatter plots showing Spearman correlations (rho) between cancer lifetime risk and: (a) Tissue cSL load (younger population, age < 50 years); (b) Tissue cSL load (on entire population which includes all age groups).*

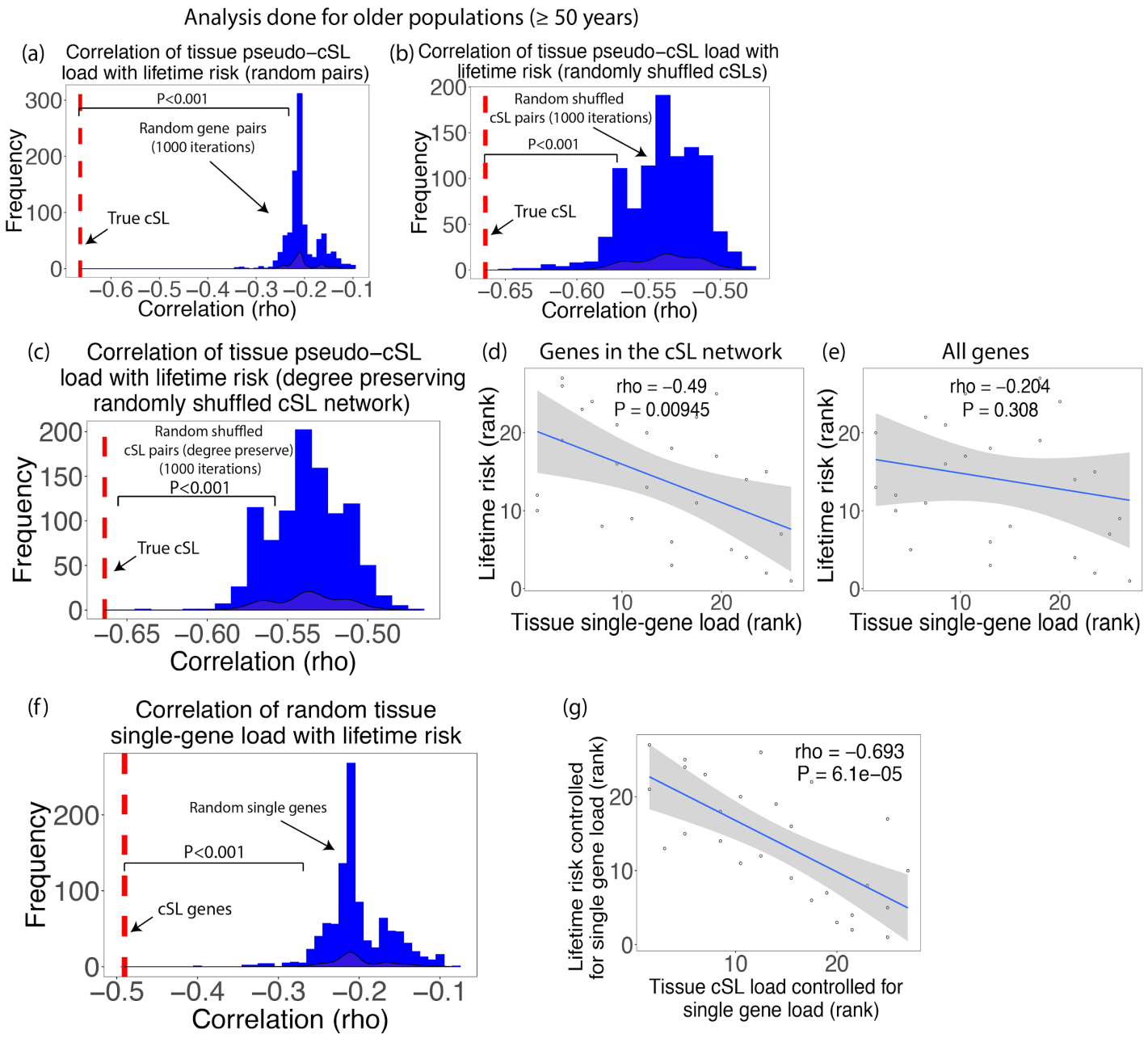


figure S3. Randomized control analysis for the correlation between tissue cSL load and lifetime cancer risk using pseudo-cSL load. *(a) Histogram showing Spearman correlations of pseudo-cSL load (computed from random gene pairs) with cancer lifetime risk (no. of iterations = 1000). Correlation for tissue cSL load (True ISLE-inferred cSL) and cancer risk is shown in red. Randomization test shows that the correlation obtained from ISLE-inferred tissue cSL network is always more significant than those obtained from*

tissue pseudo-cSL loads (P < 0.001). (b) Histogram showing Spearman correlations of pseudo-cSL load (computed from shuffled cSL pairs) with cancer lifetime risk (no. of iterations = 1000). Correlation for tissue cSL load (True ISLE-inferred cSL) and cancer risk is shown in red. Randomization test shows that the correlation obtained from ISLE-inferred cSL network is much more significant than those obtained from tissue pseudo-cSL loads (P < 0.001). (c) Histogram showing Spearman correlations of pseudo-cSL load (computed from degree-preserving randomized cSL network) with cancer lifetime risk (no. of iterations = 1000). Correlation for tissue cSL load (True ISLE-inferred cSL) and cancer risk is shown in red. Randomization test shows that the correlation obtained from ISLE-inferred cSL network is much more significant than those obtained from tissue pseudo-cSL loads (P < 0.001). Scatter plots showing Spearman correlations (rho) between cancer lifetime risk and: (d) tissue single-gene load (computed from only the genes in the ISLE-inferred cSL network); (e) tissue single-gene load (computed from all genes). (f) Histogram showing Spearman correlations of tissue single-gene load (computed from random single genes) with cancer lifetime risk (no. of iterations = 1000). Correlation for tissue single-gene load (inferred from ISLE-inferred cSL genes) and cancer risk is shown in red. Randomization test shows that the correlation obtained from cSL genes is much more significant than those obtained from random single genes (P < 0.001). (g) Scatter plots showing Spearman correlations (rho) between cancer lifetime risk and tissue cSL load controlled for tissue single-gene load (single-gene load is computed from only the genes in the ISLE-inferred cSL network).

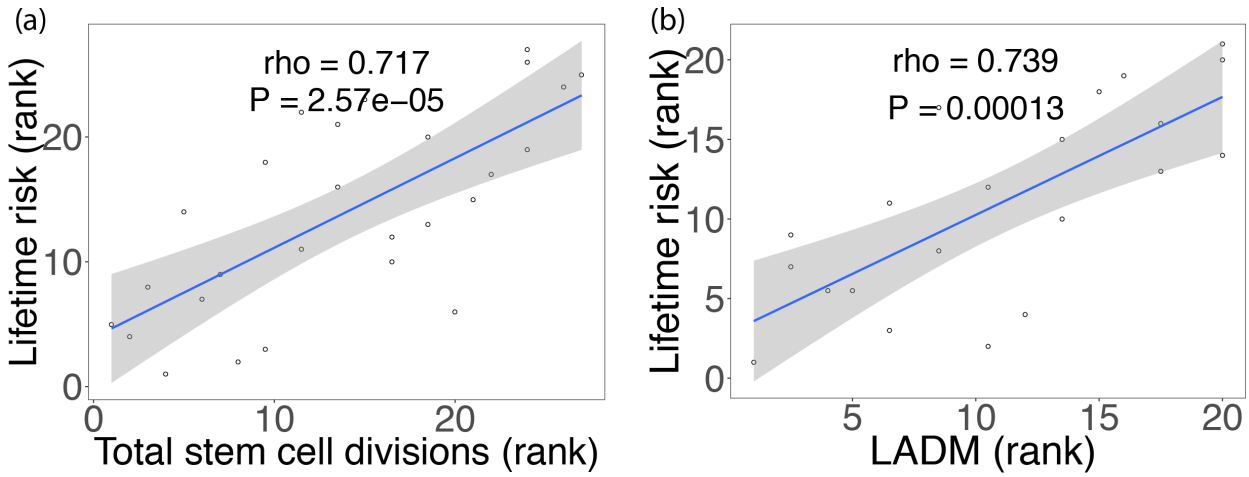


figure S4. Reproduced result of Tomasetti & Vogelstein[2] and Klutstein et al.[6] *Scatter plots showing Spearman correlations (rho) between lifetime cancer risk and: (a) total no. of stem cell divisions (NSCD); (b) LADM (levels of abnormal DNA methylation), across tissues.*

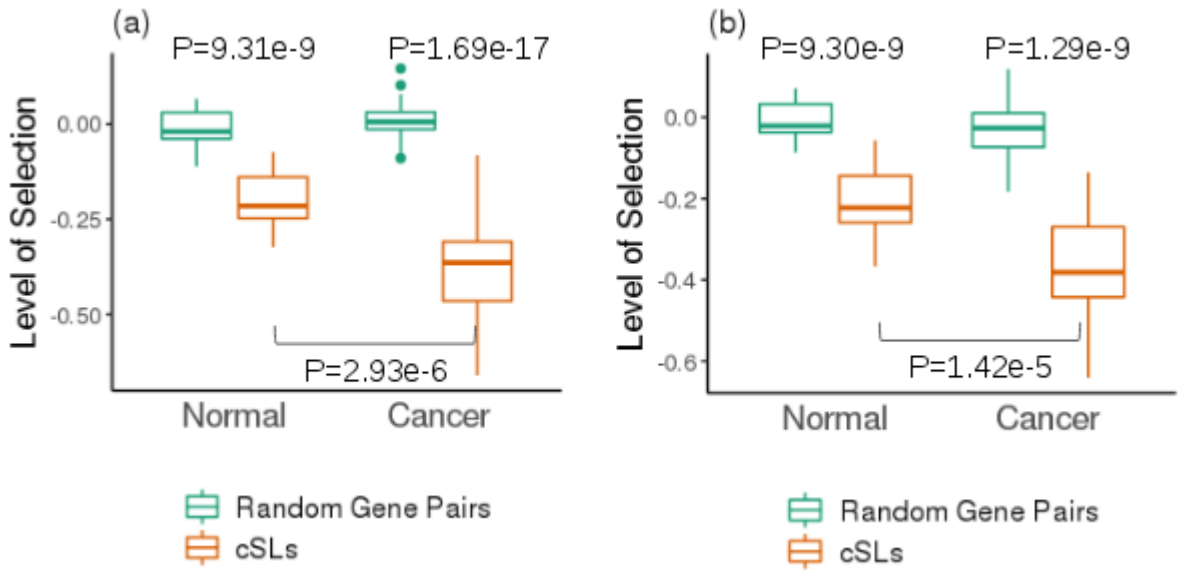


figure S5. The cSL gene pairs used in this study are more specific to cancer and have much weaker synthetic lethal effect to the normal tissues. *(a) The level of negative selection against the inactivation of both genes in cancer-derived cSL gene pairs, compared to that of random gene pairs in both normal tissues from GTEx and cancers from TCGA. Co-inactivation of the genes in cSL gene pairs are much weaker selected against in GTEx normal tissues than in TCGA cancer samples. (b) Similar to (a), but with the analysis performed using cross-validation. ISLE was applied to a random subset of TCGA cancer samples to identify the cSLs, and the negative selection level in cancer was computed for the held-out samples not used by ISLE. P values for one-sided Wilcoxon rank-sum tests are shown.*

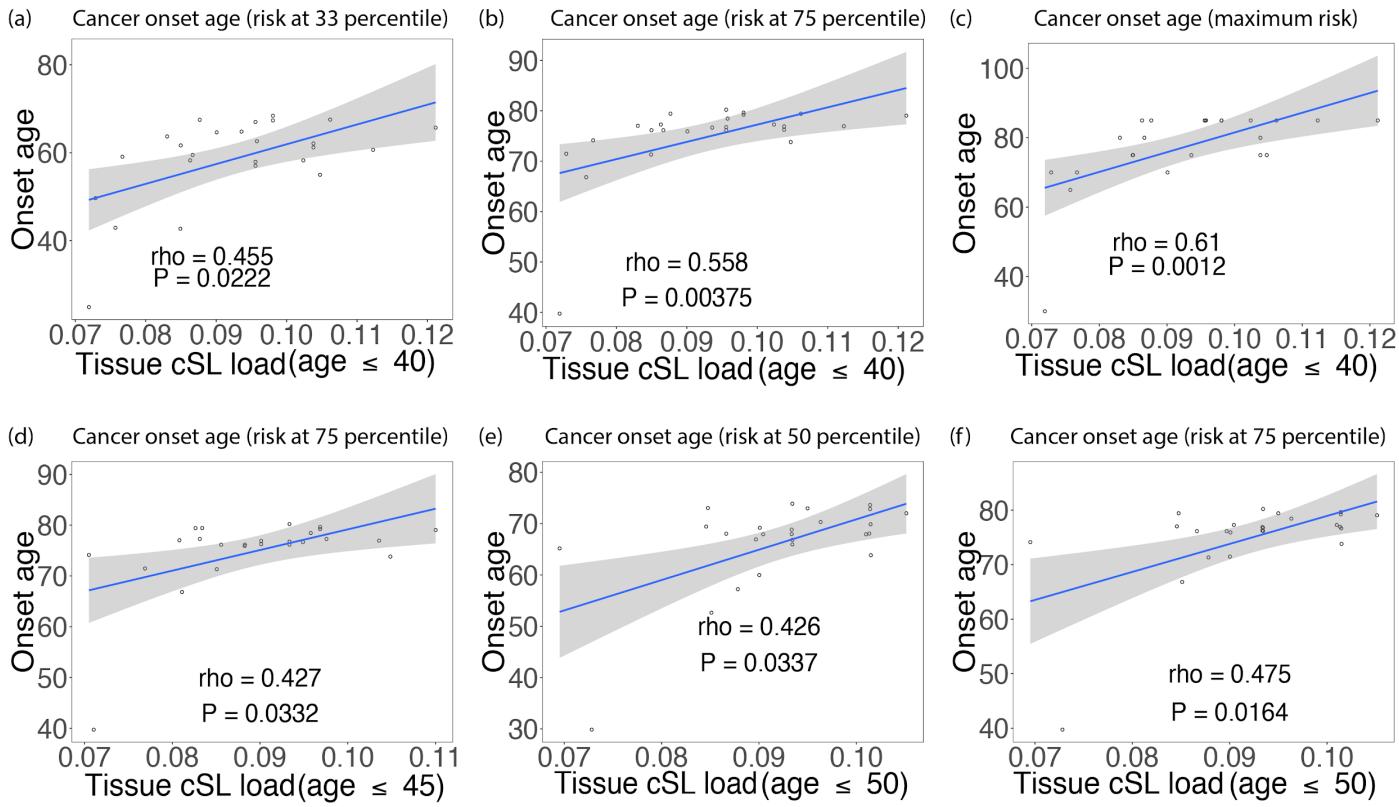


figure S6. Spearman correlation between Tissue cSL load and cancer onset age computed using different thresholds. *Scatter plots showing correlation between: (a) Tissue cSL load (age ≤ 40) and cancer onset time (computed at 33 percentile of maximum risk); (b) Tissue cSL load (age ≤ 40) and cancer onset time (computed at 75 percentile of maximum risk); (c) Tissue cSL load (age ≤ 40) and cancer onset time (computed at maximum risk); (d) Tissue cSL load (age ≤ 45) and cancer onset time (computed at 75 percentile of maximum risk); (e) Tissue cSL load (age ≤ 50) and cancer onset time (computed at 50 percentile of maximum risk); (f) Tissue cSL load (age ≤ 50) and cancer onset time (computed at 75 percentile of maximum risk). All of them have FDR < 0.1.*

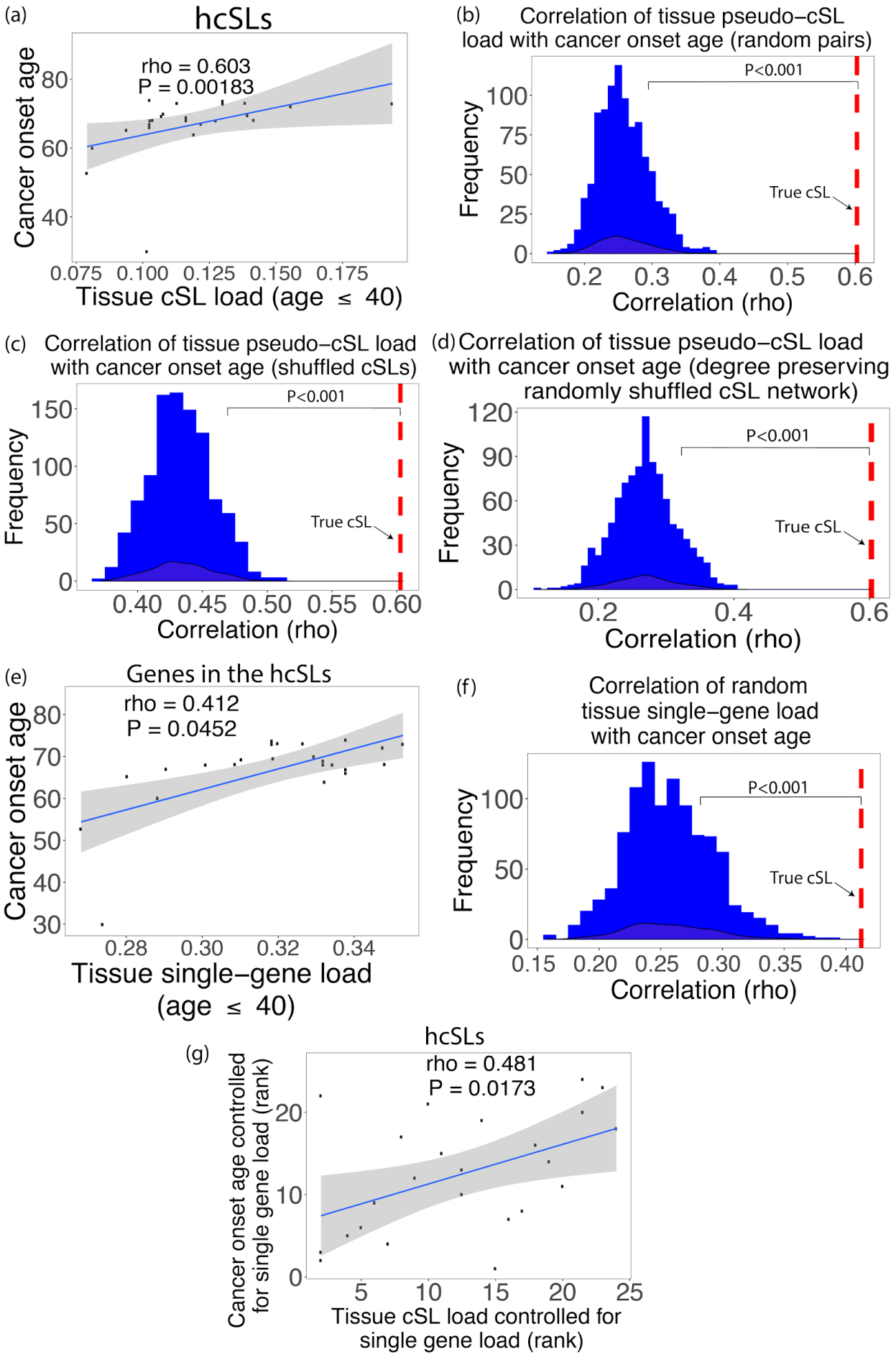

(a)
hcSLs
rho = 0.603
P = 0.00183
Cancer onset age
Tissue cSL load (age ≤ 40)
(b)
Correlation of tissue pseudo−cSL load with cancer onset age (random pairs)
P<0.001
True cSL
Frequency
Correlation (rho)
(c)
Correlation of tissue pseudo−cSL load with cancer onset age (shuffled cSLs)
P<0.001
True cSL
Frequency
Correlation (rho)
(d)
Correlation of tissue pseudo−cSL load with cancer onset age (degree preserving randomly shuffled cSL network)
P<0.001
True cSL
Frequency
Correlation (rho)
(e)
Genes in the hcSLs
rho = 0.412
P = 0.0452
Cancer onset age
Tissue single−gene load (age ≤ 40)
(f)
Correlation of random tissue single−gene load with cancer onset age
P<0.001
True cSL
Frequency
Correlation (rho)
(g)
hcSLs
rho = 0.481
P = 0.0173
Cancer onset age controlled for single gene load (rank)
Tissue cSL load controlled for single gene load (rank)

figure S7. Randomized control analysis for the correlation between tissue cSL load and cancer onset age using pseudo-cSL load. *(a) Scatter plot showing Spearman correlations (rho) between cancer onset age and tissue cSL load (computed from ISLE-inferred hcSL gene pairs). (b) Histogram showing Spearman correlations of pseudo-cSL load (computed from random gene pairs) with cancer onset age (no. of iterations = 1000). Correlation for tissue cSL load (True ISLE-inferred lcSL network) and cancer onset age is shown in red. Randomization test shows that the correlation obtained from ISLE-inferred hcSL gene pairs is always more significant than those obtained from pseudo-cSLs (P < 0.001). (c) Histogram showing Spearman correlations of pseudo-cSL load (computed from shuffled hcSL gene pairs) with cancer onset age (no. of iterations = 1000). Correlation for tissue cSL load (True ISLE-inferred hcSLs) and cancer onset age is shown in red. Randomization test shows that the correlation obtained from ISLE-inferred hcSL gene pairs is much more significant than those obtained from tissue pseudo-cSL loads (P < 0.001). (d) Histogram showing Spearman correlations of pseudo-cSL load (computed from degree-preserving randomized hcSL network) with cancer onset age (no. of iterations = 1000). Correlation for tissue cSL load (True ISLE-inferred lcSL) and cancer onset age is shown in red. Randomization test shows that the correlation obtained from ISLE-inferred hcSL network is much more significant than those obtained from tissue pseudo-cSL loads (P < 0.001). (e) Scatter plot showing Spearman correlations (rho) between cancer onset age and tissue single-gene load (computed from only the genes in the ISLE-inferred hcSL network). (f) Histogram showing Spearman correlations of tissue single-gene load (computed from random single genes) with cancer onset age (no. of iterations = 1000). Correlation for tissue single-gene load (inferred from ISLE-inferred hcSL genes) and cancer onset age is shown in red. Randomization test shows that the correlation obtained from hcSL genes is much more significant than those obtained from random single genes (P < 0.001). (g) Scatter plots showing Spearman correlations (rho) between cancer lifetime onset age and tissue cSL load controlled for tissue single-gene load (for hcSL network). Only age groups less than or equal to 40 years are considered for this analysis. Normal tissue GTEx gene expression data is used to compute the tissue cSL load.*

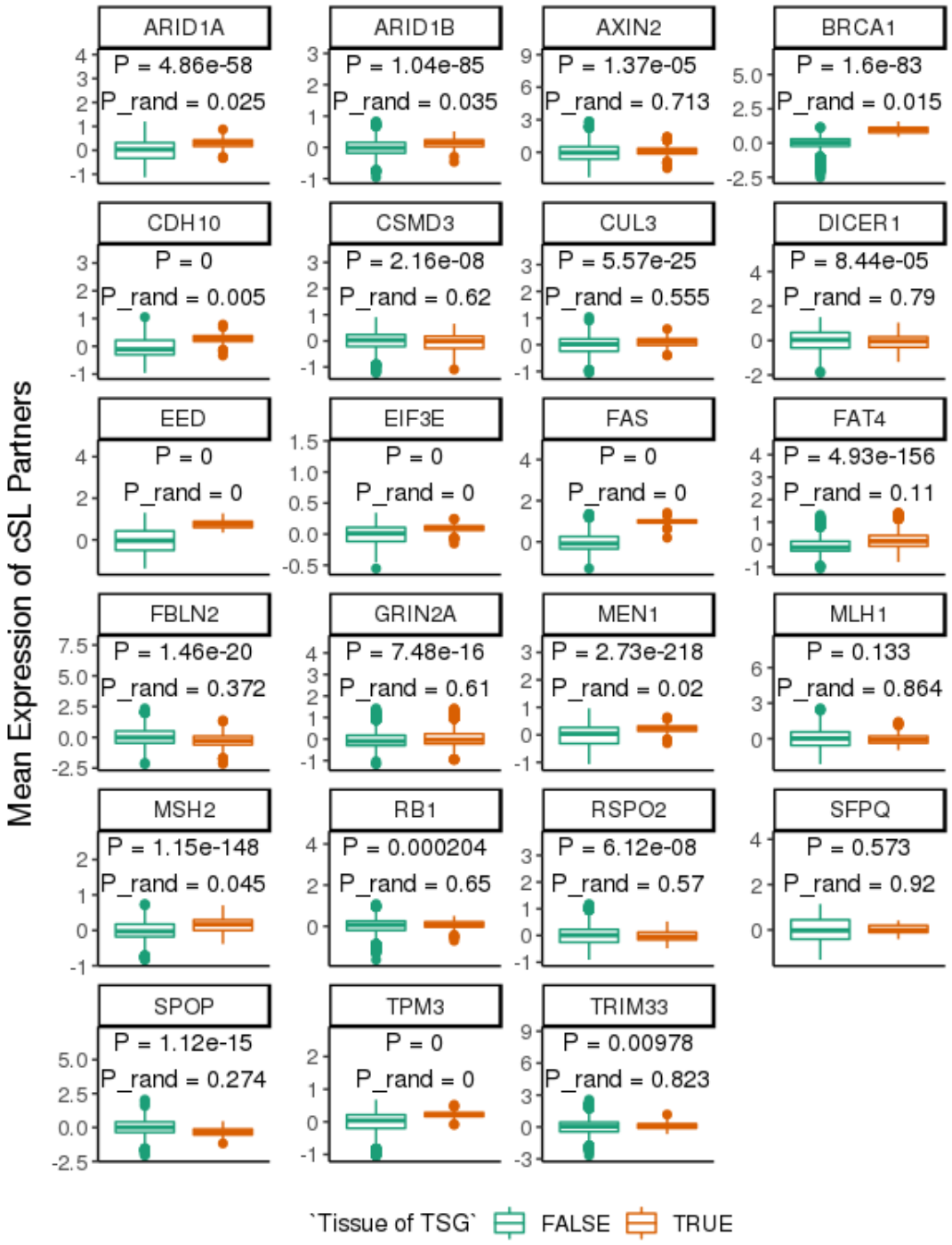


figure S8. Expression of the cSL partner genes of tumor suppressor genes (TSGs) by each TSG. *The difference between the mean expression levels of the cSL partner genes of each tumor suppressor gene (TSG) in the tissue type(s) where the TSG is a known driver ("Tissue of TSG" is TRUE) and in the rest of the tissue types where the TSG is not an established driver ("Tissue of TSG" is FALSE). The expression levels are those*

from the GTEx normal tissue samples. Each panel is a tissue-specific TSG, and the tissue types where they are established cancer drivers can be found in table S6 (Note that no cSL partner genes were identified for most of the TSGs in table S6, and we are analyzing only the TSGs with more than one cSL partners identified). The P values (top row) shown are from the linear models associated with the "Tissue of TSG" term as described in the main Methods, and the "P_rand" values (bottom row) are the empirical P values obtained from the random control tests described in the Supp. Notes. Where the P values are numerically extremely small, they are displayed as zero by the software.

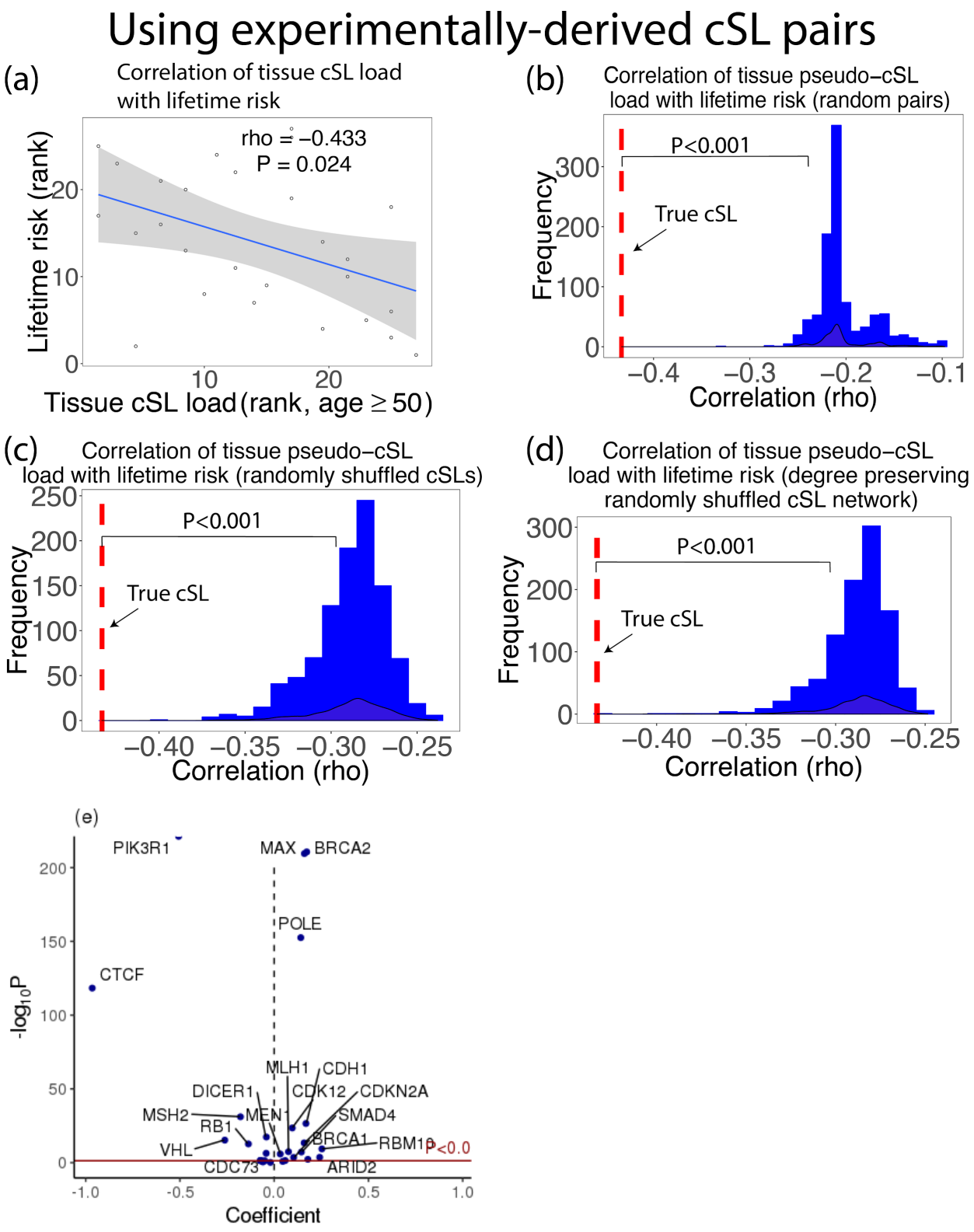


figure S9. Correlation between tissue cSL load computed with experimentally identified cSLs and lifetime cancer risk across tissues. *(a) Spearman correlation (rho)*

between tissue cSL loads computed from experimentally identified cSL pairs with lifetime cancer risk. (b) Histogram showing Spearman correlations of pseudo-cSL load (computed from random gene pairs) with cancer lifetime risk (no. of iterations = 1000). (c) Histogram showing Spearman correlations of pseudo-cSL load (computed from randomly shuffled experimentally identified cSL pairs) with cancer lifetime risk (no. of iterations = 1000). (d) Histogram showing Spearman correlations of pseudo-cSL load (computed from randomly shuffled experimentally identified cSL pairs) with cancer lifetime risk (no. of iterations = 1000). (d) Histogram showing Spearman correlations of pseudo-cSL load (computed from degree-preserving randomized experimentally identified cSL network) with cancer lifetime risk (no. of iterations = 1000). In subfigures (b,c,d) Correlation for cSL load (from experimentally identified cSLs -- true cSL) and cancer risk is shown in red. Randomization test shows that the correlation obtained from true cSL network is much more significant than those obtained from pseudo-cSL loads (P < 0.001). (e) For each tissue-specific tumor suppressor (TSG) gene Gi, the expression levels of its experimentally identified cSL partner genes in the tissue type(s) where gene Gi is a TSG were compared to those where gene Gi is not an established TSG, using GTEx normal tissue expression data. The volcano plot summarizes the result of comparison with linear models. Positive linear model coefficients (X-axis) mean that the expression levels of the cSL partner genes are on average higher in the tissue(s) where gene Gi is a TSG. Many cases have near-zero P values and are represented by points (half-dots) on the top border line of the plot. All TSGs with FDR < 0.05 are labeled.

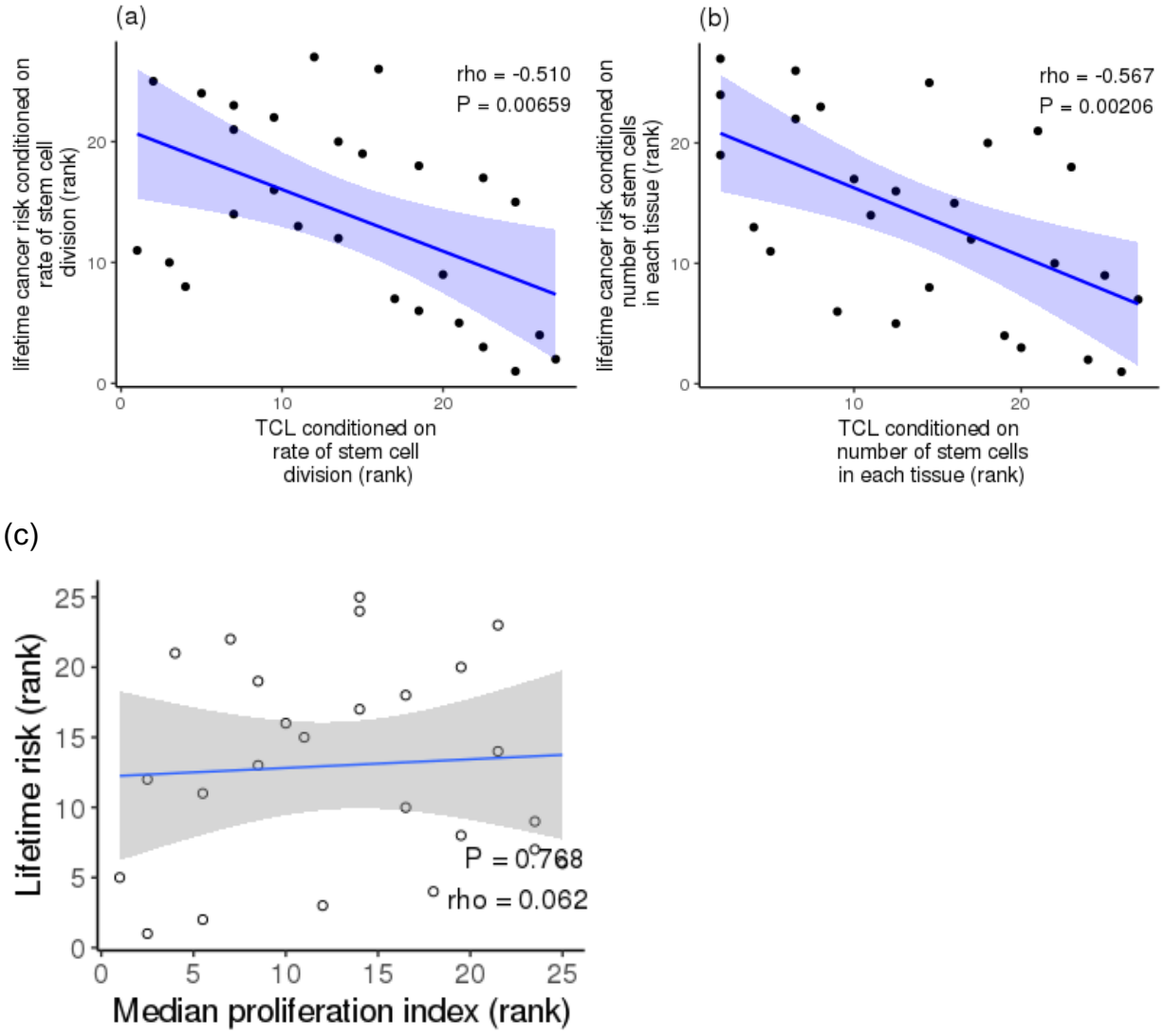


figure S10. Correlation between cSL load and lifetime cancer risk is independent from stem cell division. *Scatter plots showing partial Spearman correlations (rho) between lifetime cancer risk and tissue cSL load (TCL) conditioned on: (a) the rate of tissue stem cell division; (b) the number of stem cells residing in each normal human tissue. (c) A scatter plot showing the Spearman correlations (rho) between lifetime cancer risk and the median proliferation index for each type of the normal tissues from the GTEx dataset.*

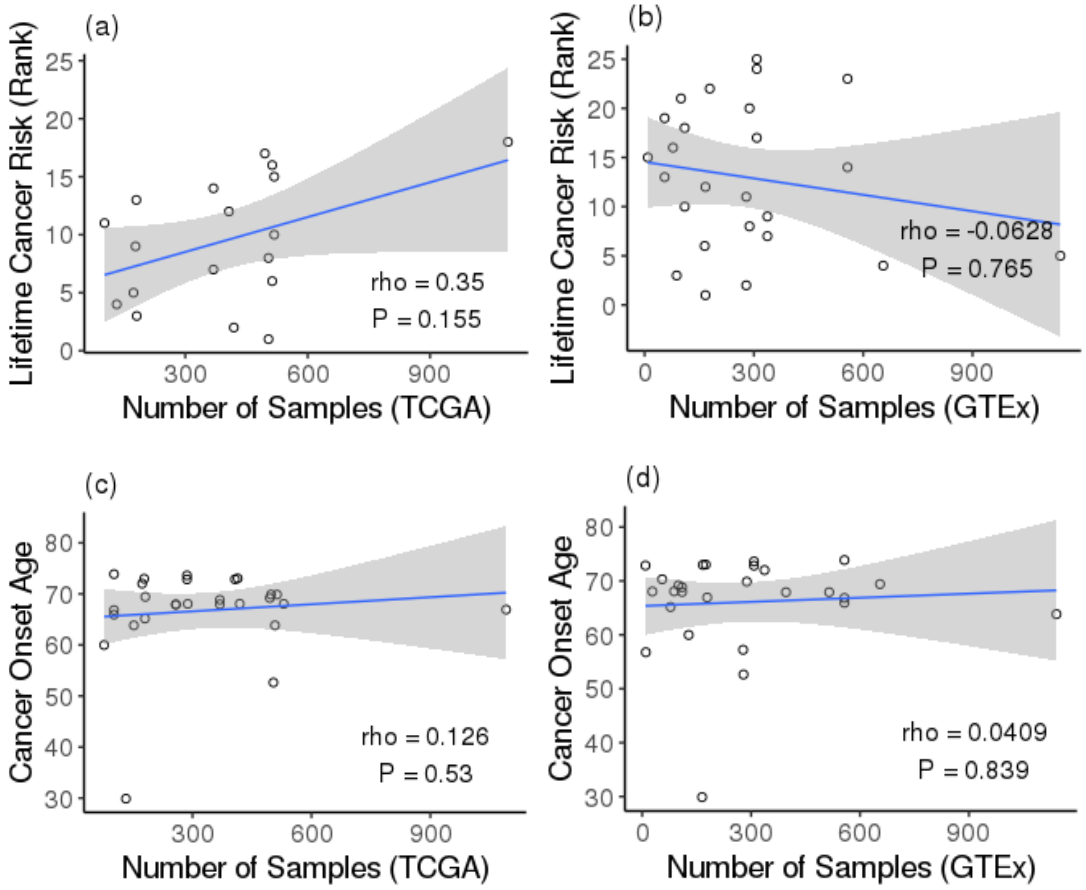


figure S11. Correlation between cSL load and lifetime cancer risk or cancer onset age is not confounded by the number of samples available for each tissue. *The number of samples of each tissue in the TCGA or GTEx datasets is not a confounding factor for correlation with cancer risk or onset age. The four scatter plots show all the pairwise correlation between the number of TCGA or GTEx samples for each tissue type and lifetime cancer risk (ranked) or cancer onset age. Spearman's correlation coefficient and the corresponding P value are shown.*

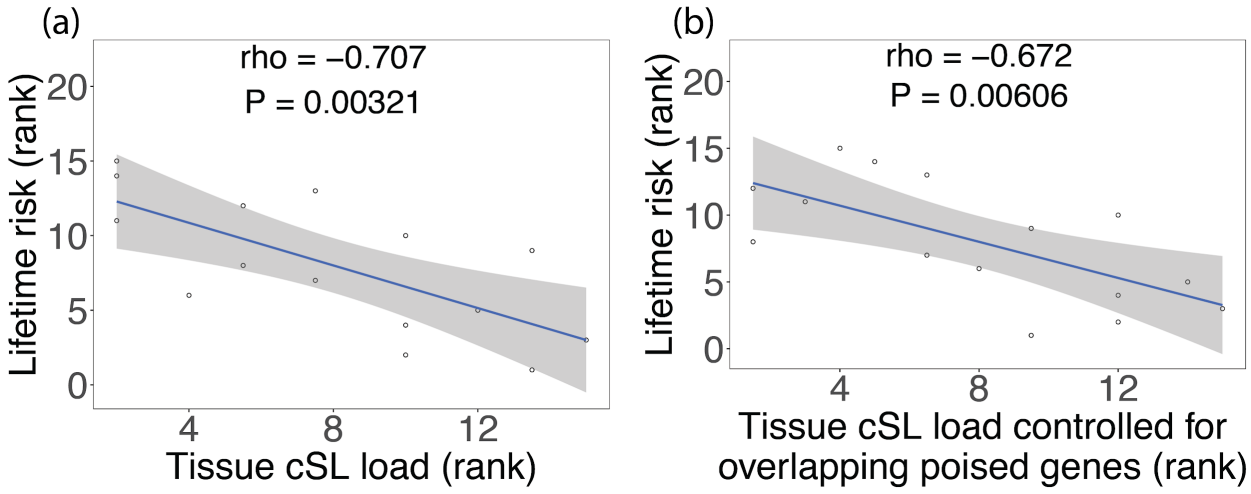


figure S12. Correlation between cSL load and lifetime cancer risk is not confounded by the number of poised genes in each tissue. *Scatter plots showing the Spearman's correlations across tissues between lifetime cancer risk and: (a) TCL computed for the older population (age ≥ 50 years) by considering only tissues for which we have data for the number of poised genes (15 data points); (b) TCL controlled for the number of poised genes the number of poised genes*

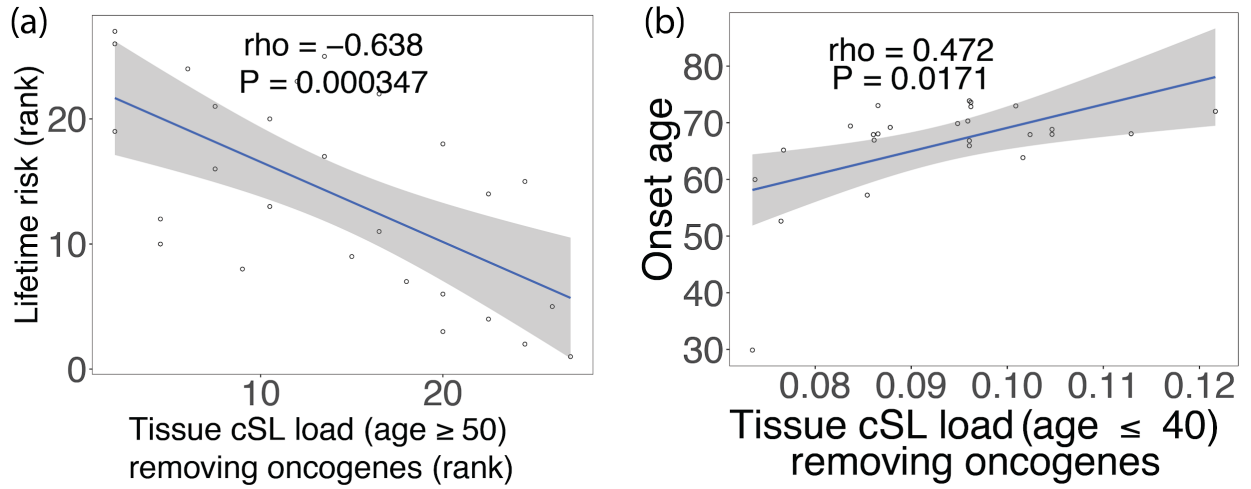


figure S13. Correlation between cSL load and lifetime cancer risk is not confounded by the expression levels of oncogenes in each tissue. *cSL pairs with any oncogenes are removed, and tissue cSL load (TCL) is recomputed for all tissues. (a) Scatter plot showing the Spearman's correlations between lifetime cancer risk and recomputed TCL computed for the older population (age ≥ 50 years). (Ranked values are used as lifetime*

cancer risk spans several orders of magnitude.) (b) Scatter plot showing the Spearman's correlations between cancer onset age and recomputed TCL (age ≤ 40 years).

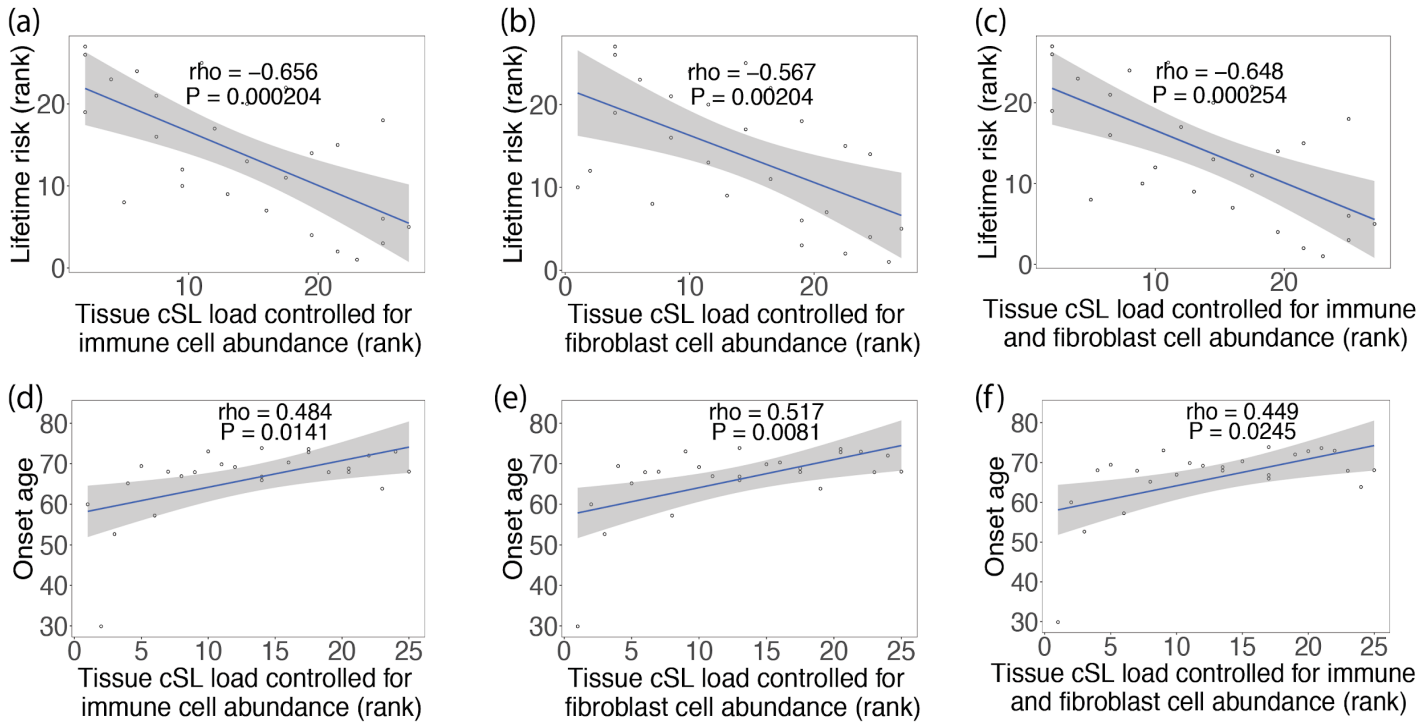
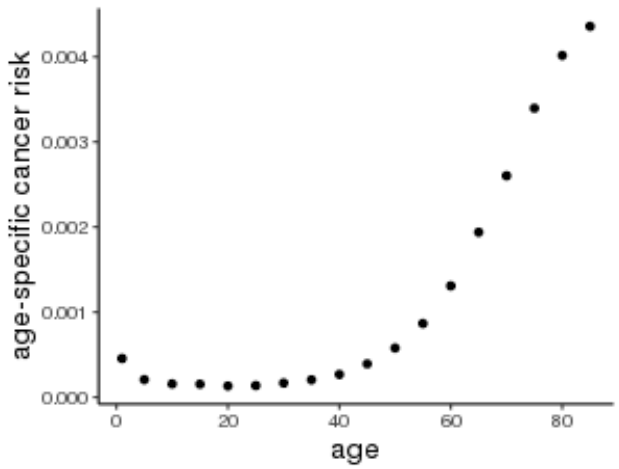


figure S14. Correlation between cSL load and lifetime cancer risk or cancer onset age is not confounded by the abundance of immune cells or fibroblasts in each tissue. *Scatter plots showing the Spearman's correlations between lifetime cancer risk and TCL computed for the older population (age ≥ 50 years): (a) by controlling for the predicted immune cell abundance estimates across tissues; (b) by controlling for the predicted fibroblast (stromal) cell abundance estimates across tissues. (c) by controlling for both the predicted immune and fibroblast cell abundance estimates across tissues. Scatter plots showing the Spearman's correlations between cancer onset age and TCL computed for the younger population (age ≤ 40 years): (d) by controlling for the predicted immune cell abundance estimates across tissues; (e) by controlling for both the predicted fibroblast cell abundance estimates across tissues; (f) by controlling for the predicted immune and fibroblast cell abundance estimates across tissues.*

(a)



(b)

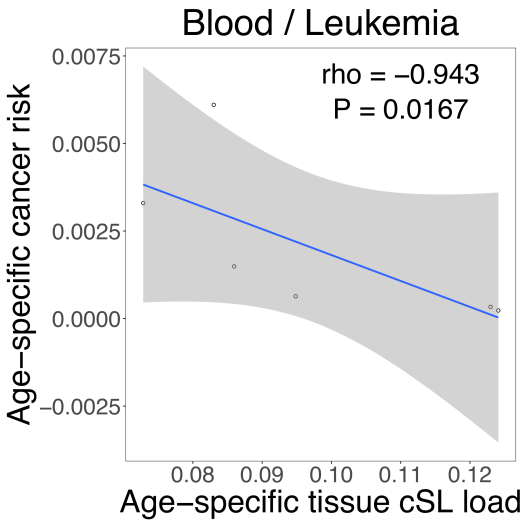


figure S15. The case of leukemia where tissue cSL load is correlated with cancer risk by age. *(a) Scatter plot showing the variation of age-specific cancer risk variation with age in leukemia. (b) Scatter plot showing the variation of age-specific cancer risk variation in leukemia with age-specific tissue cSL load (TCL). Spearman's correlation (rho) and p-values are shown.*

REFERENCES AND NOTES

1. Surveillance, Epidemiology, and End Results (SEER) Program, "SEER*Stat Database Incidence - SEER 9 Regs Research Data, Nov 2017 Sub (1973–2015) - Linked To County Attributes - Total U.S., 1969–2016 Counties," National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission (2018); www.seer.cancer.gov.

2. C. Tomasetti, B. Vogelstein, Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).

3. C. Tomasetti, L. Li, B. Vogelstein, Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).

4. S. Wu, S. Powers, W. Zhu, Y. A. Hannun, Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–47 (2015).

5. A. P. Feinberg, R. Ohlsson, S. Henikoff, The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7**, 21–33 (2006).

6. M. Klutstein, J. Moss, T. Kaplan, H. Cedar, Contribution of epigenetic mechanisms to variation in cancer risk among tissues. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2230–2234 (2017).

7. A. G. Sauer, R. L. Siegel, A. Jemal, S. A. Fedewa, Current prevalence of major cancer risk factors and screening test use in the United States: Disparities by education and race/ethnicity. *Cancer Epidemiol. Biomark. Prev.* **28**, 629–642 (2019).

8. P. Armitage, R. Doll, The age distribution of cancer and a multi-stage theory of carcinogenosis. *Int. J. Epidemiol.* **33**, 1174–1179 (2004).

9. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K.-S. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortés-Ciriano, D. C. Zhou, W.-W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Lian; The MC3 Working Group; The Cancer Genome

Atlas Research Network, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).

10. K. M. Haigis, K. Cichowski, S. J. Elledge, Tissue-specificity in cancer: The rule, not the exception. *Science* **363**, 1150–1151 (2019).

11. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

12. S. J. Elledge, A. Amon, The BRCA1 suppressor hypothesis: An explanation for the tissue-specific tumor development in BRCA1 patients. *Cancer Cell* **1**, 129–132 (2002).

13. G. Schneider, M. Schmidt-Supprian, R. Rad, D. Saur, Tissue-specific tumorigenesis: Context matters. *Nat. Rev. Cancer* **17**, 239–253 (2017).

14. C. B. Bridges, The origin of variations in sexual and sex-limited characters. *Amer. Nat.* **56**, 51–63 (1922).

15. A. Bender, J. R. Pringle, Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **11**, 1295–1305 (1991).

16. L. H. Hartwell, P. Szankasi, C. J. Roberts, A. W. Murray, S. H. Friend, Integrating genetic approaches into the discovery of anticancer drugs. *Science* **278**, 1064–1068 (1997).

17. C. J. Lord, A. Ashworth, PARP inhibitors: Synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017).

18. N. J. O'Neil, M. L. Bailey, P. Hieter, Synthetic lethality and cancer. *Nat. Rev. Genet.* **18**, 613–623 (2017).

19. S. Parameswaran, D. Kundapur, F. S. Vizeacoumar, A. Freywald, M. Uppalapati, F. J. Vizeacoumar, A road map to personalizing targeted cancer therapies using synthetic lethality. *Trends Cancer.* **5**, 11–29 (2019).

20. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

21. The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

22. J. S. Lee, A. Das, L. Jerby-Arnon, R. Arafeh, N. Auslander, M. Davidson, L. McGarry, D. James, A. Amzallag, S. G. Park, K. Cheng, W. Robinson, D. Atias, C. Stossel, E. Buzhor, G. Stein, J. J. Waterfall, P. S. Meltzer, T. Golan, S. Hannenhalli, E. Gottlieb, C. H. Benes, Y. Samuels, E. Shanks, E. Ruppin, Harnessing synthetic lethality to predict the response to cancer treatment. *Nat. Commun.* **9**, 2546 (2018).

23. X. Feng, N. Arang, D. C. Rigiracciolo, J. S. Lee, H. Yeerna, Z. Wang, S. Lubrano, A. Kishore, J. A. Pachter, G. M. König, M. Maggiolini, E. Kostenis, D. D. Schlaepfer, P. Tamayo, Q. Chen, E. Ruppin, J Silvio Gutkind, A platform of synthetic lethal gene interaction networks reveals that the GNAQ uveal melanoma oncogene controls the Hippo pathway through FAK. *Cancer Cell* **35**, 457–472.e5 (2019).

24. S. M. B. Nijman, S. H. Friend, Cancer. Potential of the synthetic lethality principle. *Science* **342**, 809–811 (2013).

25. P. Schyman, R. L. Printz, S. K. Estes, T. P. O'Brien, M. Shiota, A. Wallqvist, Concordance between thioacetamide-induced liver Injury in rat and human in vitro gene expression data. *Int. J. Mol. Sci.* **21**, 4017 (2020).

26. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F.

Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W.-N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, T. R. Golub, A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).

27. B. Szikriszt, Á. Póti, O. Pipek, M. Krzystanek, N. Kanu, J. Molnár, D. Ribli, Z. Szeltner, G. E. Tusnády, I. Csabai, Z. Szallasi, C. Swanton, D. Szüts, A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol.* **17**, 99 (2016).

28. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, Biological agents. Volume 100 B. A review of human carcinogens. *IARC Monogr. Eval. Carcinog. Risks Hum.* **100** (Pt. B), 1–441 (2012).

29. L. R. Ferguson, A. E. Pearson, The clinical use of mutagenic anticancer drugs. *Mutat. Res.***355**, 1–12 (1996).

30. S. Gundy, M. Baki, I. Bodrogi, Vinblastine, cisplatin and bleomycin (VPB) adjuvant therapy does not induce dose-dependent damage in human chromosomes. *Neoplasma* **36**, 457–464 (1989).

31. C. Mascaux, M. Angelova, A. Vasaturo, J. Beane, K. Hijazi, G. Anthoine, B. Buttard, F. Rothe, K. Willard-Gallo, A. Haller, V. Ninane, A. Burny, J.-P. Sculier, A. Spira, J. Galon, Immune evasion before tumour invasion in early lung squamous carcinogenesis. *Nature* **571**, 570–575 (2019).

32. A. Magen, A. D. Sahu, J. S. Lee, M. Sharmin, A. Lugo, J. S. Gutkind, A. A. Schäffer, E. Ruppin, S. Hannenhalli, Beyond synthetic lethality: Charting the landscape of pairwise gene expression states associated with survival in cancer. *Cell Rep.* **28**, 938–948.e6 (2019).

33. A. Mullard, Synthetic lethality screens point the way to new cancer drug targets. *Nat. Rev. Drug Discov.* **16**, 736 (2017).

34. M. A. Horlbeck, A. Xu, M. Wang, N. K. Bennett, C. Y. Park, D. Bogdanoff, B. Adamson, E. D. Chow, M. Kampmann, T. R. Peterson, K. Nakamura, M. A. Fischbach, J. S. Weissman, L. A. Gilbert, *Cell* **174**, 953–967.e22 (2018).

35. M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, S. T. Sherry, The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).

36. M. Goldman, B. Craft, M. Hastie, K. Repečka, A. Kamath, F. McDade, D. Rogers, A. N. Brooks, J. Zhu, D. Haussler, The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv* 326470 (2019).

37. D. M. Parkin, C. S. Muir, S. J. Whelan, Y. T. Gao, J. Ferlay, J. Powell, *Cancer Incidence in Five Continents* (IARC, 1992).

38. A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The Reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649-D655 (2018).

39. H. W. Wendt, Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the U statistic. *Eur. J. Soc. Psychol.* **2**, 463–465 (1972).

40. D. Venet, J. E. Dumont, V. Detours, Most random gene expression signatures are significantly associated with breast cancer outcome. *PLOS Comput. Biol.* **7**, e1002240 (2011).

41. K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P. W. Laird, D. A. Levine, S. L. Carter, G. Getz, K. Stemke-Hale, G. B. Mills, R. G. Verhaak, Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).

42. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. N. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. Mc Manus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

43. M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, R. A. Irizarry, Minfi: A flexible and comprehensive bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).