# Supporting Information

## Palmer et al. 10.1073/pnas.1714478115

### SI Theory

**Apparent Power Laws: The Overlap of IM-II and PLM.** For many cancer types, IM-II and the PLM overlap and in some cases are essentially indistinguishable (Fig. S11). The equation describing IM-II can appear very similar to a power law, especially for ages close to the pivot age. To better assess this phenomenon and analyze when this happens, we perform a Taylor series expansion around the pivot age and equate coefficients with a binomial expansion up to and including the third-order term to show that

$$\frac{1}{e^{e^{-\alpha(t-\tau)}}-1} \propto \left(\alpha(t-\tau)+\frac{e-1}{e-2}\right)^{\frac{e}{e-2}}+\mathcal{O}\left((t-\tau)^4\right).$$

Therefore, the IM-II equation approximates a power law when $\tau=(e-1)/\alpha(e-2)\approx 53.2$ y, with exponent $e/(e-2)\approx 3.78$. For these values, the IM-II and PLM agree to within 5% for the age interval of 33–82 y.

Intriguingly, it is precisely this narrow range of overlap of IM-II and PLM that fits the majority of cancer types. The median pivot age for all cancer types is $\tau = 49.9$ y and the median exponent for the PLM is 3.51. This explains how so many cancer types (e.g., colon and rectum cancer) can fit both IM-II and PLM.

**Stem Cell Power-Law–Type Model.** To estimate the number of driver mutations in carcinogenesis, we apply the PLM to the data collected in ref. 19. Although some of the data may be making invalid assumptions about stem cell behavior, the values span many orders of magnitude and we therefore believe can be informative about overall trends.

We assume each stem cell division has a constant probability $p$ of acquiring a driver mutation, and we estimate the probability of at least one stem cell acquiring $n$ such mutations. A very similar model was used in a recent paper to study cancer risk (38). In this approach, all cancer types are assumed to have the same number of driver mutations, the same probability for mutation, and the same probability for evading the immune system and progressing to cancer incidence. Despite the strength of these assumptions, the results still give powerful insights into the most prominent factors involved in cancer risk.

If we assume that, once a hit occurs it cannot be reversed, the probability for a single stem cell to acquire $n$ given mutations in $d$ asymmetric divisions is given by the following:

$$P(n,d,s=1)=\left(1-(1-p)^d\right)^n. \quad \text{[S1]}$$

For a population of $s$ stem cells, we make a mean field approximation and assume that each stem cell is independent. This is essentially assuming that all stem cell divisions are asymmetric, producing one stem cell and one differentiated cell. The probability of at least one stem cell acquiring $n$ driver mutations in $d$ divisions is then as follows:

$$P(n,d,s)=1-\left(1-\left(1-(1-p)^d\right)^n\right)^s\approx p^n\, s\, d^n, \quad \text{[S2]}$$

where the power law approximation above is valid for $p\,d \ll 1$. Values for $s$ and $d$ are taken from ref. 19. In our calculations, we have ignored stem cell divisions from early development, since their effect is negligible. One exception is ovarian germ cell cancer, with $d = 0$ and all divisions arising from embryonic development. We have excluded this cancer from the above calculation but included it in Fig. S6 *A* and *C* as part of the blue group. Furthermore, we have excluded brain cancers, since, in our estimation, the value of $d = 0$ used in ref. 19 is not well supported by the reference cited therein.

Fitting to the data, we find $n = 0.91$ (Fig. S6*E*). Although $n < 1$ driver mutations does not make sense biologically, this is a strong indication that the number of driver mutations is small, and possibly exactly 1. In the absence of other factors, such as immune system decline, this would imply constant incidence of cancer with age and is therefore at odds with observed incidence data.

We also analyzed values for $s$ and $d$ separately and found that the various organs separate naturally into two distinct groups (Fig. S6 *B* and *D*). We found that the organs with high $d$ do not have higher cancer risk than those with high $s$ (Fig. S6*C*). If accumulation of mutations was an important factor in cancer risk, one would expect lifetime risk $L$ to show a high, nonlinear dependence on $d$, which is not observed, suggesting that the number of replication-induced driver mutations is small and hence that additional factors are fundamental to the increase in risk with age. We note that some authors define driver mutations as mutations that increase fitness, which is subtly different from those considered above. Here, driver mutations are mutations that correspond to rate-limiting steps in carcinogenesis. In this way, the immunological model is not incompatible with having several mutations necessary for carcinogenesis, as long as these are not rate-limiting independent events.

**Parameter Estimates.** To estimate values for the biological parameters, we assume only one driver mutation is required and therefore that the rate of attempts is proportional to the rate of stem cell divisions:

$$r = p\, s\, d, \quad \text{[S3]}$$

where $p$ is the probability of an oncogenic mutation per division. Following ref. 38, we use the estimated value $P = 10^{-8}$. For colon cancer for example, this gives $r \sim 100$/y and $T_0 \sim 10^6$ cells (for all available cancer types, see Dataset S1).

**Assessing the Goodness of Fit of the Models.** Fitting different models to data and comparing their performance is not a trivial task. The most widely used methods rely on comparing the sum of the residuals potentially, with additional adjustments if a different number of parameters is used in the different models. In our analysis, we focused on $R^2$, which is an easily interpretable quantity that compares how well a model performs with respect to the mean, with larger values indicating a better performance. Although $R^2$ can be unreliable when used to evaluate different nonlinear models (39), especially if small differences are present, it still provides a reasonable quantification due to its standardized range, 1 being the maximum value it can take. Moreover, since both IM-II and PLM have the same number of parameters, no additional adjustments are necessary due to a different complexity of the model. It is worth noting that performing the analysis using the Akaike information criterion (AIC) (40) produces similar results.

When fitting a model to data, it is also important to prevent the model from fitting small statistical fluctuations present in most real-word datasets (a phenomenon usually named overfitting). To avoid overfitting, it is important that the number of data points being fitted ($d$) is significantly larger than the number of parameters

of the model ($p$), with a "rule of thumb" suggesting that $d$ should be at least three times larger than $p$. In our study, when analyzing common cancers, $d = 66$ and $1 < p < 4$, thus suggesting that overfitting is very unlikely to play a role. Consistently, for colorectal cancer, the IM-II–adjusted $R^2$ is 0.99, while the PLIM-adjusted $R^2$ is also 0.99.

**Further Analysis.** Ideally, evidence for a causative link between thymic output and disease risk in humans would come from a prospective study examining TREC DNA or T cell telomeres. As far as we are aware, no such study exists. An association between short leukocyte telomeres and cancer risk has been shown in a prospective study, which holds even after taking age into account (41). Similarly, patients with head and neck cancer have significantly lower TREC content (42). However, this could be due to an immune response. Furthermore, obesity is associated with both low TREC counts and increased risk of cancer and infectious diseases (43). Moreover, in mice, thymus regeneration has also been shown to extend life span (44).

We note that there appears to be a period between ages 5 and 20 where male TRECs are higher than female TRECs (4) and a period between ages 25 and 40 where male risk is lower than female risk in colon cancer, possibly suggesting a 20-y delay between immune exhaustion and incidence.

We can also attempt to take into account the effect of an expansion of the memory T cell pool with age. This will result in a dilution of TRECs, and therefore the rate of decline of T cell production will be slightly smaller than that estimated from TRECs. One experiment attempted to address this issue by measuring TRECs from naive T cells separately (3), which decreased exponentially with exponent $\alpha = 0.033\ y^{-1}$. This figure is still quite close to the value used above and gives similar fitting results, suggesting that the expansion of the memory T cell pool has a minimal effect.

It is possible that the initially high risk of cancer and infectious disease in infants is due to a lack of memory T cells and, therefore, that incidence in adults arises despite functional memory T cells. In this way, our model suggests that, for a given clone of memory T cells, consequent changes with age are minimal and that the change in disease incidence with age is mainly dictated by the changes in naive T cell production.

**T Cell Heterogeneity Model.** The declining output of naive T cells from the thymus results in T cells that are closer to replicative senescence, but also a decrease in T cell heterogeneity (at any given time but not over long timescales). The immunological model presented above does not involve an increase in risk due to the decline in heterogeneity, suggesting perhaps that the timescales for which reduced heterogeneity may be a problem are shorter than the timescales of cancer progression.

To elucidate this further, we have constructed a simple mathematical model describing an increase in risk with age due to decreasing T cell heterogeneity. We assume a critical window of time exists, in which a T cell with an appropriate receptor must detect the immunogenic cells.

Let $H$ be the size of the homeostatic T cell pool and $L$ be the rate of T cell production, which decreases as $L = L_0\, e^{-\alpha t}$. Then the size of each clone of T cells with a given receptor is $H/L$. Let $p$ be the probability a given T cell receptor can detect the antigenic cells.

Assuming the average number of T cells encountering a group of cells in a given window of time is constant with age, given by $c$, the probability that none of the T cells has an appropriate receptor is given by the following:

$$P = (1-p)^{\frac{cL}{H}},$$

which gives a risk profile of the following form:

$$R = Ae^{-e^{-\alpha(t-\tau)}}.$$

This model has the same number of fitting parameters as IM-II and the PLM, but fits significantly worse. This suggests that the decline in T cell heterogeneity does not explain the rising risk of disease incidence with age.

**Fig. S1.** Thymic involution. Data taken from ref. 2. An exponential decay with a half-life of 16 y can be seen in thymus volume (2) and also in TREC measurements (2–4).

Fig. S1

**Fig. S2.** Many cancers rise with age inversely proportional to T cell production. Upon fitting an exponential curve, regions with $R^2 > 0.9$ are shaded in yellow. The blue line indicates an exponent $\alpha = 0.044\ y^{-1}$ with shaded region 0.04–0.06 $y^{-1}$. The top five best-fitting cancers (as measured by AIC) have median exponent of 0.043 $y^{-1}$.

Fig. S2

**Fig. S3.** Schematic illustration showing five stages leading to cancer incidence. In our model, stage III is modeled as a random walk in which most cancers are eliminated. This gives an age dependence for cancer risk coming from stage IV, rather than stage II, as assumed in the PLM. Our analysis suggests that the IET might often be of the order of $\sim 10^6$ cells (*Methods*).

Fig. S3

**Fig. S4.** Tuberculosis in Cambodia. Tuberculosis prevalence in Cambodia (37) with IM-I and IM-II fitted in light and dark red, respectively.

Fig. S4

**Fig. S5.** T-lymphoblastic leukemia. T-lymphoblastic leukemia incidence is approximately constant in adults.

Fig. S5

**Fig. S6.** Number of driver mutations in carcinogenesis is small. Data taken from Tomasetti and Vogelstein (19). (*A*) Lifetime cancer risk ($L$) plotted against total number of stem cell divisions by tissue type with a trend line indicating approximately $L \propto (s\,d)^{0.4}$. (*B*) For each organ, the lifetime number of divisions per stem cell ($d$) is plotted against total number of stem cells ($s$), forming two distinct groups. Contour lines of constant $s \times d$ are plotted in gray, showing increasing total stem cell divisions toward the upper right. (*C*) Lifetime cancer risk plotted against total number of stem cell divisions with trend lines for each group, showing that high division rates do not contribute more than high stem cell numbers to lifetime risk. (*D*) Log scale of stem cell number $s$ divided by division rate $d$ with two groups separated at $s/d = s_{\max}/d_{\max}$, where $s_{\max}$ and $d_{\max}$ correspond to maximum values of $s$ and $d$. (*E*) Fitting the power-law–type model $L \propto s \cdot d^n$ to the data, gives $n = 0.912$ (*A*) or $n = 0.128$ (*B*), depending on whether fitting is performed on log-transformed values (*B*) or original values (*A*).

Fig. S6

**Fig. S7.** Survivability. Pivot age is inversely correlated with 5-y survival with Pearson's correlation $r = -0.58$ and value of $P = 8 \times 10^{-9}$. The IM-II parameter $B$, given by $B = K_0 \log(d/b)$, is inversely correlated with 5-y survival with Pearson's correlation $r = -0.68$ and value of $P = 4 \times 10^{-15}$.

Fig. S7

**Fig. S8.** Combined power law immunological model (PLIM). Level plots showing how well the PLIM fits in various areas of parameter space. The number of driver mutations, and the immune parameter $B$, indicate how steeply risk rises with age due to accumulating mutations, and immune system decline, respectively. The PLIM reduces to the PLM when the immune parameter $B$ goes to minus infinity. The IM-II region of parameter space is given by $\gamma = 0$, corresponding to 1 driver mutation. Within this region, IM-I corresponds to $B = 0$. The cancer types with exponential behavior have their best fits close to the IM-II (and IM-I) region of parameter space, whereas the cancer types with power law behavior have their best fits either close to the PLM region of parameter space or in between.

Fig. S8

**Fig. S9.** Cancer and infectious diseases show similar scaling behavior. (*A*) IM-I behavior. (*B*) IM-II behavior. Universal scaling functions for both genders (*C*) and gender separated (*D*).

Fig. S9

**Fig. S10.** Cancer incidence plotted on log–log scales and universal scaling functions using different measures. Log–log plots of incidence (per 100,000 person-years). Data taken from SEER (11). (*A* and *B*) Some cancer types rise exponentially fitting IM-I (*A*), while some cancer types rise like power laws, although can still be fit by IM-II (*B*). Fitting curves for IM-II and PLM are shown in red and green, respectively. (*C*) The top 20 best-fitting incidence curves as measured by Akaike information criterion (AIC) for IM-II. Universal scaling functions for the top 20 cancer types with best $R^2$ and the 22 cancer types, which are in the intersection of the top 30 best $R^2$ and top 30 best AIC. The decline in risk for late ages, discussed in the main text, can be seen in the center of each figure.

Fig. S10

**Fig. S11.** All cancer incidence curves. Log-linear plots of incidence (per 100,000 person-years) by age group for all cancer types. The fitting curves for PLM, IM-I, and IM-II are in green, light red, and dark red, respectively.

Fig. S11

## Dataset S1. All cancer types and infectious diseases

Dataset S1

Model fitting parameters, survivability, and gender bias exponents. Note that, for some cancers, the fitting parameter $B$ is negative, and the pivot age is undefined. For the cancer types where stem cell division data are available, the biological parameter estimates are also listed.

**Dataset S2. Cancer by tissue type**

Dataset S2

Stem cell numbers, division rates, and red and blue groupings, corresponding to high rate of division or high number of stem cells, respectively.