T Cell Receptor Diversity, Cancer and Sex: Insights from 30,000 TCRβ Repertoires

H. Jabran Zahid[1*], Ruth Taniguchi[2], Marco Garcia Noceda[2], Harlan Robins[2] and Julia Greissl[1]

[1]Microsoft Research - 14820 NE 36th St, Redmond, WA 98052

[2]Adaptive Biotechnologies - 1165 Eastlake Ave E, Seattle, WA 98109

[*]To whom correspondence should be addressed; E-mail: hzahid@microsoft.com

Abstract

Cancer is a widespread disease claiming millions of lives each year, yet the relationship between cancer risk and age-related immune system decline (i.e., immunosenescence), specifically the ability of T cells to detect and eliminate cancerous cells, remains poorly understood. Here we analyze T cell receptor (TCR) β repertoires from $\sim 30,000$ subjects demonstrating that TCR diversity—a key measure of immunocompetency—systematically declines between the ages of 20 and 80 years. Notably, this decline occurs 11 years later in females compared to males and coincides with their lower incidence of cancer, reinforcing the link between TCR diversity and cancer risk. Thus, using data from the Surveillance, Epidemiology, and End Results (SEER) database, we correlate our measured TCR diversity to cancer risk by modeling rising cancer incidence with age as the combined effect of accumulating mutations and the increasing probability that cells harboring these mutations evade immune detection due to loss of TCR diversity. Our analysis suggests that both of these

processes significantly contribute to the risk of cancer and that lower cancer incidence in females is due to the delayed onset of immunosenescence. Our results highlight the importance of understanding immunosenescence in carcinogenesis, potentially providing new opportunities for immunotherapies targeting an aging immune system as well as personalized care leveraging TCR diversity as a biomarker.

Introduction

The human population is rapidly aging, with the number of people over 60 years old expected to double by 2050 (*1*). This demographic shift is predicted to result in a commensurate increase in cancer incidence and mortality (*2*). As the need for cancer prevention and treatment grows, understanding why the risk of cancer rises dramatically with age becomes more urgent[1]. Over a century ago it was suggested that the immune system is critical in preventing cancer (*4*), yet the precise role of immunosenescence in rising cancer incidence with age remains ambiguous. An essential feature of a healthy immune system is high TCR diversity, which enables it to respond to a broad range of immunological threats including cancer (*5–14*). Here we explore the relationship between declining TCR diversity and rising cancer incidence with age.

Alpha-beta T cells are the most abundant T cell subtype and a key component of the adaptive immune system, playing a central role in immune surveillance including the elimination of cells harboring potentially deleterious mutations via TCR-mediated antigen recognition (*15*, *16*). A large pool of T cells with diverse, randomly generated TCRs (*17*) is produced by the human body mostly during childhood and adolescence, before thymic involution (*18*). One consequence of aging is the functional decline of the immune system, often referred to as im-

[1]We use cancer incidence and cancer risk interchangeably throughout, as the distinction between the two is minimal in the context of first-time cancer diagnoses and does not impact our analysis (*3*).

munosenescence, which is characterized by a range of effects (*19–24*) including the loss of TCR diversity (*25–29*). In elderly individuals this loss may significantly contribute to higher disease incidence, poor health outcomes and inferior vaccine response (*30–36*). Accurate measurements of TCR diversity as a function of age, and other relevant biological factors, are important for quantifying immunosenescence and its impact on human health.

Sex is a genetically defined biological factor known to influence immune response and health outcomes (*37*). Males are more susceptible than females to many infectious diseases (*38*, *39*) and they have poorer response to vaccines (*40*). Similarly, males exhibit higher incidence rates for many cancers and poorer survival outcomes (*41*, *42*). While studies report that the T cell repertoires decline more rapidly in males than in females (*43*, *44*), significant uncertainties remain due to small sample sizes. Thus, sex-linked differences in immunosenescence, perhaps due to some combination of genetics, hormones and differences in thymic output (*37*, *45*, *46*), may explain the lower incidence of cancer in females. However, a precise quantitative relationship between sex-linked disparities in immunosenescence and cancer incidence—based on large-scale measurements—has yet to be established.

Mathematical models of carcinogenesis attempt to identify and quantify the dominant processes by which cancers arise. An early influential model posits that cancer results from a series of discrete cellular mutations (*47*); later models incorporate additional genetic and cellular mechanisms (*48*). However, the role of immunosenescence is often not treated or is incorporated in a limited and/or implicit manner. In contrast, Palmer et al. (*49*) model cancer incidence explicitly adopting an exponentially declining immune system based on the timescale of thymic atrophy, suggesting that this decline may be the dominant contributing factor to the increasing risk of cancer with age and attributing the difference in cancer incidence between sexes to sex-dependent variations in thymic output. We extend their approach by leveraging direct measurements from sequenced T cell repertoires.

T cell repertoire sequencing is a powerful tool for understanding and characterizing the adaptive immune system (*50–52*). High-throughput methods (*53*, *54*) now allow for sequencing the TCRs of more than 10^6 T cells obtained from a single blood draw. This depth of sampling represents a small fraction of the estimated $\sim 5 \times 10^{11}$ T cells in the human body (*55*) which have undergone thymic selection (*56*). Despite capturing only a fraction of the full repertoire, this approach provides an unbiased window into systematic changes in the immune system across large cohorts. Here we measure the relationship between TCR diversity, age and sex and use these measurements to model cancer incidence. Our study aims to provide new insights into immunosenescence and cancer, with potential relevance to the development of personalized medicines, robust immunotherapies, vaccines of high efficacy and other efforts aimed at combating declining health with age (*57–60*).

Results

We perform a cross-sectional analysis of a large homogeneously processed cohort of 30,430 unique subjects with TCRβ repertoires sequenced as part of the T-Detect COVID test. This test was granted Emergency Use Authorization by the Food and Drug Administration[2] for identifying the T cell immune response to SARS-CoV-2. We statistically analyze repertoires as a function of age and sex; the sequencing protocol is independent of these two variables. We interpret the total number of T cells sequenced, S, and the number of unique clonotypes (i.e., richness), D, as relative estimates of total repertoire size and total TCR diversity, respectively. These quantities are log-normally distributed and base 10 logarithms (referred to simply as "log") are used throughout the analysis.

[2]https://www.fda.gov/media/146481/download

Repertoire Size and TCR Diversity Depend on Age and Sex

Figures 1A and 1B show that the median repertoire size and TCR diversity decline by a factor of 1.3 (0.1 dex[3]) and 1.8 (0.25 dex), respectively, between the ages of 20 and 80 years. While the decreasing size of repertoires contributes to the loss of TCR diversity with age (a factor of 1.3 between the ages of 20 and 80 years), a more significant driver is the increasing clonality of T cells (a factor of 1.4 or 0.15 dex; Figure 1C). After accounting for measurement uncertainty, we estimate that the peak-to-peak intrinsic biological scatter in TCR diversity for the central 90% of subjects increases by a factor of 2 to 5 (0.3 dex to 0.7 dex) between the ages of 20 and 80 years, respectively. Notably, at any given age, the variation in TCR diversity across individuals is greater than the systematic change with age, particularly in older individuals.

Repertoire size and TCR diversity also vary systematically by sex (Figures 2A and 2B). Females younger than 45 years old have similar repertoire size and slightly higher TCR diversity than males; at older ages, females have a slower decline in both measures. Similar trends in T cell fractions by age and sex are reported (*61*, *62*), further reinforcing the robustness and generalizability of these findings. Figures 2C and 2D show that the repertoire size and TCR diversity of females are comparable to those of males who are 11.4 ± 0.9 (1σ) years younger, suggesting a delayed onset of immunosenescence in females. This delayed decline is evident not only in the median derived from tens of thousands of subjects but also in the consistent scatter between sexes shown in Figures 2C and 2D. The consistency of the scatter suggests that the observed differences in repertoire size and TCR diversity between sexes may not be a consequence of a mechanism impacting only a subset of the male or female population but a systematic factor, such as more enduring naive T cell production in females (*45*). We observe similar trends in a smaller, independent cohort (see Supplementary Material Figure S2).

[3] A difference of x dex indicates a change by a factor of 10^x.

A Model Linking Immunosenescence, Cancer Incidence and Sex

Using data from the SEER database, we model the relationship between TCR diversity and cancer incidence. These measures show opposing trends with respect to age and sex (compare Figures 2B and 3A). Cancer incidence increases with age and is lower in females, whereas TCR diversity declines with age and is higher in females. Given the central role of T cells in eliminating cancerous cells, rising cancer incidence with age may result from the loss of TCR diversity (*49*, *63–66*). As a corollary, the lower incidence of cancer observed in females may be due to their delayed onset of immunosenscence. We test this hypothesis by developing a biologically grounded model to identify fundamental factors driving carcinogenesis, prioritizing simplicity and interpretability over statistical optimization. We link the age- and sex-dependent rise in cancer incidence to a declining immune system, focusing solely on the initial immune evasion by cancerous cells and not the complex dynamics of subsequent disease progression.

We mathematically model two key competing biological processes that contribute to carcinogenesis: (1) the accumulation of cancerous mutations with age and (2) the declining ability of the immune system to eliminate cells harboring these mutations. We adopt the Armitage and Doll model for the accumulation of mutations, which is described by a simple power-law function of age, where the power-law index is interpreted as quantifying the number of discrete mutations required for carcinogenesis (*47*). For the immune component, the probability of cancerous cells evading immune detection is modeled as an exponential function of TCR diversity (*67*, *68*). To understand this dependence, let p represent the average probability that a random TCR will respond to a given antigen. Then the probability that an antigenic mutation escapes recognition by $\mathcal{D}$ unique TCRs in an individual's repertoire is $(1-p)^{\mathcal{D}}$ which approximates to $e^{-p\mathcal{D}}$ for small p; in other words, the probability that an antigen is not detected by the immune system increases exponentially with decreasing TCR diversity. Our model is a product of these

two biological processes given by

$$I(t, D) = A\, t^{\gamma}\, e^{-D/D_0}, \tag{1}$$

where $I(t, D)$ is cancer incidence as a function of age in years, t, and the TCR diversity that we measure, D, and A, γ and D_0 are free parameters. The normalization of D by D_0 means that our measurements of TCR diversity only need to be relatively—not absolutely—accurate.

We incorporate TCR diversity into our model using the median relationships measured separately for males and females (Figure 2B) to simultaneously fit the cancer incidence data for both sexes. This simple approach yields results consistent with those obtained from more complex individual-level modeling (see Materials & Methods). Importantly, sex is not an explicit input variable of the model and a single set of parameters is optimized to fit the cancer incidence data for both males and females. All sex differences in the model arise solely from the sex-dependence of TCR diversity, which is a model input. Both the accumulation of mutations and declining TCR diversity contribute to rising cancer incidence with age, making their effects potentially degenerate in the model. However, because only TCR diversity depends on sex, it uniquely accounts for the observed dependence of cancer incidence on sex. This sex-dependence breaks the degeneracy between model components, enabling us to isolate the risk attributable to declining TCR diversity with age.

We constrain model parameters using non-linear least squares optimization, yielding best-fit values with 95% confidence intervals: $A = 0.26^{+0.24}_{-0.17}$, $\gamma = 2.45^{+0.19}_{-0.11}$ and $D_0 = 1.22^{+0.18}_{-0.11} \times 10^5$ (Figure 3A). The primary source of residuals in our fit is the modeling assumption that cancer incidence increases monotonically with age, though it is known to decline at very old ages (*69*, *70*). Excluding the oldest age bin significantly improves the fit, as does adding an additional term to account for this decline (Supplementary Material Section S4, Figure S5). However, in both cases, the core parameter estimates remain consistent, supporting the robustness of our

model. Finally, because immune evasion precedes diagnosis, cancer incidence at time t may be expected to depend on the immune status of individuals at an earlier time $t - t_0$. While we do not explicitly account for this offset here, we note that our results are insensitive to its inclusion in our model (see Materials & Methods).

We factor cancer risk associated with accumulation of mutations and immunosenescence, respectively, by noting that cancer incidence fit by a single power-law scales as $t^{4.6}$, cf. Palmer et al. (*49*). In our model, risk associated with accumulation of mutations scales as $t^{2.45}$ and immunosenescence accounts for the remainder of the risk. Palmer et al. assume a more rapidly declining immune system than what we measure, leading them to attribute greater risk to its decline, though our main conclusions are consistent.

We partition cancer risk into two independent factors: (1) risk solely attributable to aging, which is uniform for an age group and includes contributions from accumulation of mutations and immunosenescence, and (2) variable risk among individuals of a fixed age group due to differences in their TCR diversity, which can be large (see scatter in Figure 1B). The accumulation of mutations is explicitly modeled as a function of age. For TCR diversity the dependence on age is implicit in our model and derived empirically from our measurements. To partition the risk, we calculate cancer risk as a function of age and TCR diversity using our best-fit model and then relate TCR diversity at a given age to its percentile within the overall distribution of TCR diversity for that age group. For instance, while the median TCR diversity declines with age, it represents the 50th percentile of the TCR diversity distribution within any age group, meaning a given percentile retains the same relative position within the distribution despite the overall decline in TCR diversity with age. Thus, risk associated with both the accumulation of mutations and the general decline in TCR diversity (median curve in Figure 1B) is attributed to aging. In contrast, within any age group, variable risk results from differences in TCR diversity among individuals.

Figure 3B shows our partitioning of relative cancer risk as a function of age and TCR diversity percentile. At the median TCR diversity (Figure 1B), a 75 year-old faces $\sim$13-fold higher risk of cancer than a 40 year-old; accumulation of mutations account for a 4.7-fold increase, while the loss of TCR diversity accounts for a 2.7-fold increase. Conversely, a 75-year-old in the tenth percentile has a factor of $\sim$3 lower TCR diversity than a peer in the ninetieth percentile, resulting in a $\sim$7-fold higher risk of cancer, solely due to a difference in TCR diversity. If individuals were able to maintain the median TCR diversity observed at age 45 into old age (e.g., through medical interventions), our model predicts a $\sim$50% reduction in cancer incidence at age 70. Conversely, in the hypothetical absence of all T cells ($D = 0$), our model predicts cancer incidence would be 2,200 and 11,000 per 100,000 person-years for a 40 and 75 year-old, respectively, which is significantly higher than what is observed in the general population. While no individual completely lacks T cells, studies report that immunocompromised or thymectomized subjects have higher risk of cancer (*71–73*).

Given the strong link between TCR diversity and cancer incidence, measurements of TCR diversity could help identify individuals at elevated risk, potentially warranting preventative medical intervention. The risk quantification shown in Figure 3B enables straightforward identification of such individuals.

Discussion

We measure a decline in repertoire size and TCR diversity with age which we interpret as a quantification of immunosenescence. This decline occurs 11 years later in females than in males, leading us to hypothesize that lower cancer incidence in females may be due to their delayed onset of immunosenenescence. We investigate this hypothesis by modeling cancer incidence solely as a function of age and TCR diversity. The nearly identical age at which TCR diversity and cancer incidence diverge between males and females serves as a key quantitative

constraint in our model, enabling us to link these two factors. Our biologically interpretable model quantifies the independent contributions of aging and TCR diversity loss to cancer risk and attributes sex-linked disparities in cancer incidence to the dependence of TCR diversity on sex. Consistent with our findings, independent studies report that T cell fractions vary by sex (*61*) and are lower in individuals with cancer (*62*), further reinforcing the link between immunosenescence and increased cancer risk. However, TCR diversity—while strongly correlated with T cell fraction—more precisely reflects the immune system's ability to recognize and eliminate cancerous mutations, suggesting that its decline may be more directly linked to increased cancer risk.

Mathematical models of carcinogenesis can provide valuable insights to guide research and influence future clinical practice (*74*, *75*). While cancer prevention has primarily focused on reducing risk linked to carcingogenic mutations (*76*), the risk associated with a declining immune system has received less attention (*77*). Strategies addressing immune system decline may offer new opportunities for disease prevention and care but robust molecular biomarkers are needed (*78*). Our results suggest that TCR diversity may be a sensitive biomarker which can be used prior to disease onset to identify patients with heightened risk due to their immune state. Moreover, higher TCR diversity has also been linked to positive responses in cancer immunotherapy (*79–85*), suggesting that TCR diversity may also be a key biomarker for treatment success. Accordingly, measurements of TCR diversity could help guide treatment decisions when multiple options are available. Future work aimed at understanding the role of TCR diversity as a biomarker could help advance personalized care by identifying patients who may benefit from interventions focused on disease prevention and treatment.

While our analysis presents a significant correlation between cancer incidence and immunosenescence, it also highlights the need for further research to address key limitations. Determining whether immunosenescence is causually linked to cancer risk is paramount, requiring

longitudinal studies combining immunosequencing with clinical outcomes. Immunosenescence is a complex process (*22*, *86*) which may variably impact different immune components (e.g., $CD4^+$ or $CD8^+$ T cells). Repertoire-based analyses of the clone size distribution (*27*) along with robust predictions of past infections and HLAs (*87–92*) could be combined with single-cell sequencing (*93*) of subjects to provide deeper insights into how aging, and other factors, impact the immune system's various components. We adopt a simplified treatment of cancer statistics, combining data from heterogeneous types and limiting our analysis to ages where cancer incidence increases monotonically (but see Supplementary Material Section S5 and Figure S5), though it declines at very old ages due to reasons that remain poorly understood (*69*, *70*). Our measurements of TCR diversity could be incorporated into more sophisticated models accounting for additional biological processes and cancer heterogeneity, potentially providing a more nuanced understanding of the relationship between immunosenescence and cancer. Ultimately, understanding the molecular mechanisms driving this relationship will be essential, likely requiring more large-cohort multi-omic studies of aging (*94–96*). Our results may inform and guide such research.

Cancer claims ten million lives each year and is a leading cause of death in a majority of countries (*97*, *98*). As the world's population ages, cancer incidence and mortality are expected to rise dramatically. Identifying the risk factors and causes of cancer is critical for developing medical interventions aimed at preventing and treating this widespread disease. We have quantified cancer risk associated with declining TCR diversity, a key component of immunosenescence, to show that an aging immune system may be a significant risk factor. Understanding the causes of immunosenescence and its role in disease susceptibility, onset, progression and treatment is likely to be instrumental for developing strategies to reduce mortality. This understanding has potential for significant translational impacts, including advances in predictive models of disease risk, immuno-prevention of cancer, personalized treatments and novel

immunotherapies—all of which are urgently needed to address the growing burden of cancer in our aging world.

Materials & Methods

Sequencing of Human Samples

Sequencing data from human samples used in this study were aggregated from a cohort of subjects who took the T-Detect Covid diagnostic test which received Emergency Use Authorization from the FDA. All human subjects provided informed consent and the study was approved by the WIRB Copernicus Group Institutional Review Board. Age and sex were recorded for all participants and the influence of these variables was explicitly analyzed and reported in the study.

A multi-plexed PCR method was used to sequence the CDR3 of TCR β chains of T cells, typically from 18μg of genomic DNA (*53*, *99–101*). Our results and conclusions are robust to small variations in the amount of input genomic DNA. Using the T-Detect Covid diagnostic test—which has high clinical performance (*101*)—we identify $9,952$ subjects in our sample with past or present SARS-CoV-2 infection, finding that past infections do not impact repertoire size or TCR diversity.

The median total number of productive TCRs sequenced (i.e., sequencing depth) for T-Detect Covid samples is 518,618 TCRs with 95% of subjects having a sequencing depth between 222,082 and 853,647. Self-identified males and females comprise 47.2% and 52.5% of subjects, respectively. The median age of the cohort is 50 years with 95% of subjects ranging between 20 and 74 years.

We estimate uncertainty in TCR diversity measurements from repeat measurements of the same subject. We identify 396 subjects with repeat samples on the basis of overlapping TCRs observed in their repertoires. Repeat samples from the same donor are identified by generating

the Jaccard index of the 100 highest frequency clones in any two repertoires with a set threshold on the overlap. We further match these repeat samples on age and sex. The method is found to be 100% accurate in 446 samples, correctly categorizing these samples as belonging to 223 known unique donors. From repeat measurements of the same subject, we estimate a 1σ measurement uncertainty of 0.10 dex in the log of TCR diversity for a single sample (see Supplementary Material Figure S1). These random variations average out in our summary statistics.

TCRs are heterodimers and the same TCRβ chain may randomly pair with different TCRαs on different T cells. Furthermore, TCRβ repertoires significantly undersample the total number of unique clones in subjects, which is unknown, but is estimated to be $\mathcal{O}(10^7)$ (*99*). However, we are primarily concerned with systematic changes and thus our analysis only requires *relative* accuracy in TCR diversity, which our data provide. Down-sampling repertoires to a fixed number of T cells yields qualitatively consistent results but introduces systematic bias due to prevalence of Cytomegalovirus infections (see Supplementary Material).

Cancer Incidence Data

Cancer incidence data were retrieved via the SEER*Stat Software (*102*) from the Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 22 Registries, Nov 2023 Sub (2000-2021) - Linked To County Attributes - Time Dependent (1990-2022) Income/Rurality, 1969-2022 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2024, based on the November 2023 submission. We collect incidence data sorted by sex for all cancer types excluding sex-linked cancers (e.g., reproductive system cancers, breast cancer, etc) and lung cancer due to known sex-linked behavioral differences (i.e., smoking). We restrict our analysis of cancer incidence to individuals between 20 and 80 years old.

Intrinsic Scatter Calculation

We estimate the 1σ intrinsic scatter in TCR diversity, which appears to be log-normally distributed, as

$$\sigma_i = \sqrt{\sigma_o^2 - \sigma_m^2}, \tag{2}$$

where σ_i is the intrinsic scatter, σ_o is the observed scatter containing the central 68% of samples and σ_m is the measurement uncertainty of 0.10 dex estimated from repeat samples. To estimate the intrinsic scatter at different quantiles of the distribution, we scale σ_m accordingly. For example to estimate the intrinsic scatter of the central 90% of samples, we calculate the half-confidence interval containing the central 90% of the data, and subtract σ_m scaled by 1.645 in quadrature (for a normal distribution, 90% of samples are within $\pm 1.645\sigma$) and multiply by 2.

Age Offset Between Male and Female Repertoires

We derive the age offset between male and female repertoires (Figures 2C and 2D) and estimate statistical uncertainty by independently bootstrapping the male and female repertoires. For each bootstrapped realization, we generate the median repertoire size and TCR diversity (i.e., Figures 2A and 2B) as function of age and determine the age offset that minimizes the sum of squared residuals between the male and age-offset female curves. In order to match the age grid between male and age-offset female curves, we interpolate the curves for males and age-offset females between ages of 18 and 70 in 1 year increments. We limit the maximum interpolated age to 70 years to avoid significant extrapolation of the age-offset female curves. We report the mean and standard deviation of 10,001 realizations as the age offset and 1σ statistical uncertainty, respectively.

Model Fitting

We fit cancer incidence as a function of age and TCR diversity using the Python routine curve_fit implemented in the SciPy package (*103*). We input the standard errors of the cancer incidence data to weight the fit but derive parameter uncertainties via bootstrapping to account for the significantly larger uncertainties associated with our measurements of TCR diversity which can not be straightforwardly accounted for in standard non-linear least squares fitting. When fitting the cancer incidence we interpolate the median relation of $D(t)$ for males and females separately (Figure 2B) using CubicSpline implemented in SciPy. Our results are insensitive to the polynomial degree of the spline. Interpolation of D is necessary because the age grid of D is not coincident with the cancer incidence data. Fitting a model to the log of cancer incidence yields consistent results.

We are unable to use standard error propagation because we have errors in both cancer incidence and TCR diversity data and the latter are treated as a dependent variable in our fit. To estimate the errors of our parameters, we bootstrap the repertoire data prior to generating the median $D(t)$ for males and females (i.e., Figure 2B) via the binning procedure described in the text. Each bootstrapped realization is fit using the procedure described above. The dominant source of statistical uncertainty in our fit is from our measurement of D and our bootstrapping procedure effectively propagates these uncertainties to the model parameters.

We generate the 95% confidence interval of our fit shown in Figure 3A by propagating the 2σ uncertainties associated with D (see Table 1) through the best-fit model. Meaning, when generating the upper confidence interval, we evaluate our best-fit model adopting values of D that are 2σ larger, where σ is the error associated with each point in Figure 2B. We treat the lower bound similarly and interpolate D as before. Adopting 95% confidence from the bootstrapped fits yields similar intervals but fails to account for covariance.

For simplicity, we adopt the median TCR diversity as a function of age when fitting cancer

incidence. One potential issue with using the median relation is that it does not account for the fact that in our model, individuals with lower TCR diversity have higher risk. We test whether adopting the median TCR diversity, rather than using TCR diversity of individual subjects, impacts our results. We calculate the expectation value for cancer incidence, $\langle I \rangle$, in each decade age bin as

$$\langle I \rangle = \frac{1}{N} \sum_{i=1}^{N} A\, t_i^{\gamma}\, e^{-D_i/D_0} \tag{3}$$

where t_i and D_i are age and TCR diversity for a given subject, respectively, and the sum is over all subjects in the age bin. A, γ and D_0 are free parameters, as before. We calculate $\langle I \rangle$ for males and females independently and interpolate $\langle I \rangle$ across age bins to match the binning of the cancer incidence data. We use the same procedure described above to generate best-fit parameters and 95% confidence intervals yielding $A = 0.41^{+0.73}_{-0.27}$, $\gamma = 2.41^{+0.19}_{-0.17}$ and $D_0 = 0.94^{+0.17}_{-0.13} \times 10^5$. The use of individual measurements has a marginal effect on D_0, and the parameters derived here should be applied when working with individual measurements. However, our results and conclusions are robust to use of median TCR diversity instead of individual measurements.

For completeness, we attempt to account for the fact that initial immune evasion leading to a cancer diagnosis at time t depends on accumulated mutations and TCR diversity at some earlier time. This can be parameterized as a time offset, t_0, added to the right side of equation 1 as follows:

$$I(t, D) = A\, (t - t_0)^{\gamma}\, e^{-D(t-t_0)/D_0}, \tag{4}$$

where $D(t)$ which we measure is evaluated at some earlier time $t - t_0$. However, we are unable to properly constrain this parameter with our data due to covariance between A and t_0 and therefore model cancer incidence without this time offset parameter.

Calculating TCR diversity percentiles

We derive the 1σ intrinsic scatter of log D as a function of age (Figure 1B) as described in the text. We then use this 1σ intrinsic scatter and assume a normal distribution to relate values of D to their corresponding percentiles which we then evaluate using our best fit model to generate a prediction of cancer incidence. Interpolation is necessary to generate the smooth map shown in Figure 3B and introduces minimal errors that do not impact our conclusions. To estimate relative risk, we normalize all predictions to a 40 year-old subject with median TCR diversity.

Acknowledgments: We thank Mary Carrington, Margaret Geller, Matthew Noakes, Ash Chapfuwa, Miah Wander and Gheath Alatrash for reading the manuscript and providing constructive comments, Beryl Crossley for developing the method to identify repeat samples from the same donor and Rebecca Elyanow for its implementation in our software pipeline.

Funding Statement: The work was funded by Microsoft Corporation and Adaptive Biotechnologies.

Competing Interests: HJ Zahid, J Greissl have employment and equity ownership with Microsoft. MG Noceda, HS Robins have employment and equity ownership with Adaptive Biotechnologies. The authors declare no other competing interests.

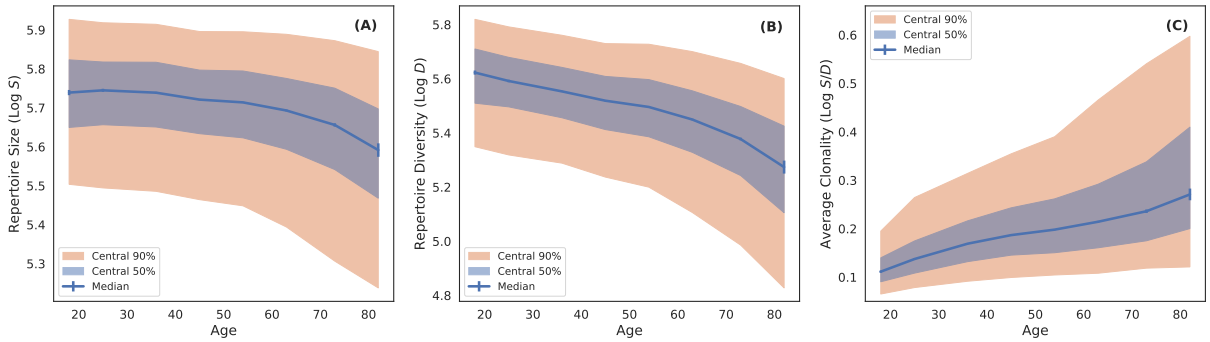

Figure 1: **Repertoire size, TCR diversity and average clonality as a function of age.** (A) Log of the total number of productive TCRs sequenced (Log S) in each repertoire as a function of age. The blue curve is the median in decade wide age bins. Error bars are bootstrapped and blue and orange shaded regions indicate the distribution of the central 50 and 90% of the data, respectively. (B) Log of TCR diversity (Log D) as a function of age. The binning procedure and shaded regions are the same as in (A). (C) Log of the average clonality of T cells in a repertoire as a function of age. For a given repertoire, average clonality is given by the ratio of TCR diversity to repertoire size (Log S/D). The binning procedure and shaded regions are the same as in (A). The binned data and errors are provided in Table 1.

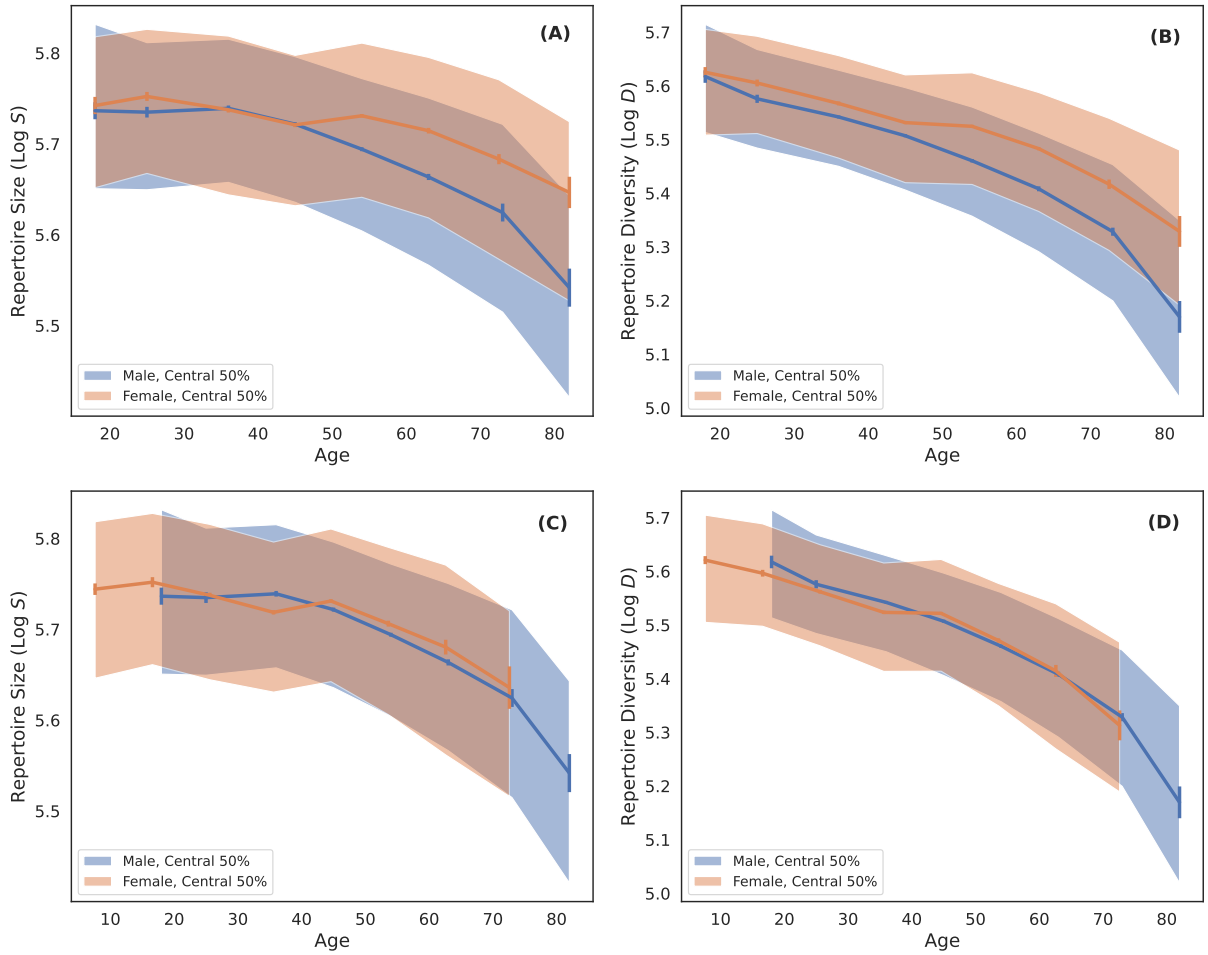

Figure 2: **Repertoire size and TCR diversity as a function of age and sex.** (A) Log of the total number of productive TCRs sequenced (Log S) as a function of age and sex. Blue and orange curves are the median in decade wide age bins for males and females, respectively. Error bars are bootstrapped and shaded regions indicate the distribution of the central 50% of the data points. (B) Log of the TCR diversity as a function of age (Log D) for male (blue) and female (orange) subjects. The binning procedure and shaded regions are the same as in (A). (C) and (D) are the same as (A) and (B), respectively, but with female curves shifted to younger ages by 11.4 years. Binned medians and errors are provided in Table 1.

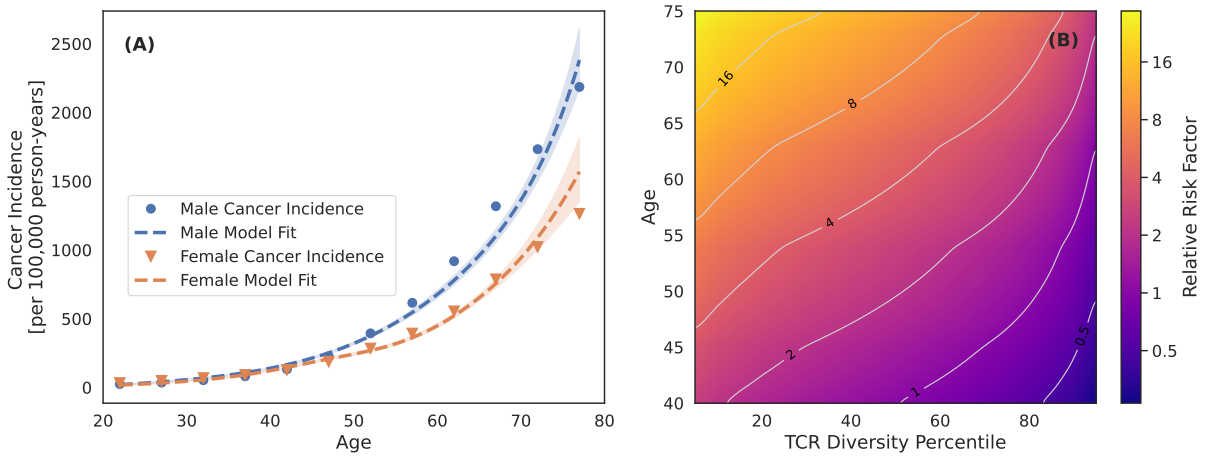

Figure 3: **Modeling cancer incidence as a function of age and TCR diversity.** (A) Blue and orange points are cancer incidence (per 100,000 person-years) for males and females, respectively. Cancer incidence is modeled as a function of age and TCR diversity using Equation 1 and the median TCR diversity measured for each sex (Figure 2B). Dotted lines are the best-fit model and shaded regions are 95% confidence intervals. (B) A heatmap showing relative cancer risk as a function of age and TCR diversity percentile. TCR diversity percentiles indicate the fraction of subjects at a given age who have TCR diversity values below that specific value. Relative cancer risk is calculated as the cancer incidence derived using our best-fit model normalized to cancer incidence of a 40 year old with median TCR diversity. Contours of constant relative risk are shown in gray.

N Subjects	Median Age	$\text{Log}_{10}S$	Error $\text{Log}_{10}S$	$\text{Log}_{10}D$	Error $\text{Log}_{10}D$
		All Subjects			
561	18	5.739	0.007	5.623	0.008
1591	25	5.745	0.004	5.592	0.005
5043	36	5.739	0.002	5.553	0.002
7608	45	5.721	0.002	5.519	0.002
8252	54	5.714	0.002	5.496	0.002
5491	63	5.693	0.002	5.449	0.003
1647	73	5.656	0.004	5.377	0.005
237	82	5.592	0.017	5.274	0.023
		Male Subjects			
272	18	5.736	0.009	5.617	0.012
754	25	5.735	0.006	5.576	0.007
2564	36	5.739	0.003	5.542	0.003
3642	45	5.722	0.002	5.507	0.003
3777	54	5.694	0.002	5.460	0.003
2475	63	5.664	0.003	5.408	0.005
779	73	5.624	0.010	5.328	0.008
108	82	5.542	0.020	5.170	0.030
		Female Subjects			
288	18	5.742	0.010	5.625	0.009
833	25	5.752	0.005	5.605	0.007
2461	36	5.738	0.003	5.567	0.004
3950	45	5.721	0.003	5.532	0.003
4448	54	5.731	0.002	5.525	0.003
3010	63	5.715	0.003	5.483	0.003
866	72	5.683	0.006	5.417	0.009
128	82	5.647	0.017	5.329	0.030

Table 1: **Repertoire size and TCR diversity as a function of age.** The first and second columns give the number of subjects in the bin and median age, respectively. The third and fifth columns give the median log of repertoire size, S, and log TCR diversity, D, respectively. The fourth and sixth columns give the errors on the median log repertoire size and log TCR diversity, respectively. Errors are 1σ uncertainties determined from 10,001 bootstrapped realizations. Measurements for all subjects, male subjects and female subjects are given in the table in descending order.

During the preparation of this work the author(s) used ChatGPT in order to assist in editing the manuscript for clarity and readibility. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

(*1*) Kanasi, E., Ayilavarapu, S., and Jones, J. (2016). The aging population: demographics and the biology of aging. *Periodontology 2000 72*, 13–18.

(*2*) J, F. et al. Global Cancer Observatory: Cancer Tomorrow (version 1.1) ed. by for Research on Cancer, I. A., `https://gco.iarc.who.int/tomorrow` (accessed Sept. 10, 2024).

(*3*) Sasieni, P. et al. (2011). What is the lifetime risk of developing cancer?: the effect of adjusting for multiple primaries. *British journal of cancer 105*, 460–465.

(*4*) Burnet, M., *Immunological surveillance*; Elsevier: 2014.

(*5*) Davis, M. M. et al. (1998). Ligand recognition by (alpha)(beta) T cell receptors. *Annual review of immunology 16*, 523.

(*6*) Nikolich-Žugich, J., Slifka, M. K., and Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology 4*, 123–132.

(*7*) Foster, A. D., Sivarapatna, A., and Gress, R. E. (2011). The aging immune system and its relationship with cancer. *Aging health 7*, 707–718.

(*8*) Martin, M. P., and Carrington, M. (2013). Immunogenetics of HIV disease. *Immunological reviews 254*, 245–264.

(*9*) Corthay, A. (2014). Does the immune system naturally protect against cancer? *Frontiers in immunology 5*, 197.

(*10*) Montgomery, R. A. et al. (2018). HLA in transplantation. *Nature reviews nephrology 14*, 558–570.

(*11*) Kovacs, A. A. et al. (2020). Association of HLA Genotype With T-Cell Activation in Human Immunodeficiency Virus (HIV) and HIV/Hepatitis C Virus–Coinfected Women. *The Journal of infectious diseases 221*, 1156–1166.

(*12*) Francis, J. M. et al. (2021). Allelic variation in class I HLA determines $CD8^+$ T cell repertoire shape and cross-reactive memory responses to SARS-CoV-2. *Science immunology 7*, eabk3070.

(*13*) Granadier, D. et al. In *Seminars in immunopathology*, 2021; Vol. 43, pp 119–134.

(*14*) Olafsdottir, T. A. et al. (2022). HLA alleles, disease severity, and age associate with T-cell responses following infection with SARS-CoV-2. *Communications Biology 5*, 914.

(*15*) Hedrick, S. M. et al. (1984). Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature 308*, 149–153.

(*16*) Yanagi, Y. et al. (1984). A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. *Nature 308*, 145–149.

(*17*) Hozumi, N., and Tonegawa, S. (1976). Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences 73*, 3628–3632.

(*18*) Aspinall, R., and Andrew, D. (2000). Thymic involution in aging. *Journal of clinical immunology 20*, 250–256.

(*19*) Walford, R. L. (1964). The immunologic theory of aging. *The Gerontologist 4*, 195–197.

(*20*) Aw, D., Silva, A. B., and Palmer, D. B. (2007). Immunosenescence: emerging challenges for an ageing population. *Immunology 120*, 435–446.

(*21*) Gruver, A., Hudson, L., and Sempowski, G. (2007). Immunosenescence of ageing. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland 211*, 144–156.

(*22*) Pawelec, G. (2018). Age and immunity: what is "immunosenescence"? *Experimental gerontology 105*, 4–9.

(*23*) Mittelbrunn, M., and Kroemer, G. (2021). Hallmarks of T cell aging. *Nature immunology 22*, 687–698.

(*24*) Liu, Z. et al. (2023). Immunosenescence: molecular mechanisms and diseases. *Signal transduction and targeted therapy 8*, 200.

(*25*) Naylor, K. et al. (2005). The influence of age on T cell generation and TCR diversity. *The Journal of Immunology 174*, 7446–7452.

(*26*) Goronzy, J. J., Lee, W.-W., and Weyand, C. M. (2007). Aging and T-cell diversity. *Experimental gerontology 42*, 400–406.

(*27*) Qi, Q. et al. (2014). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences 111*, 13139–13144.

(*28*) Britanova, O. V. et al. (2014). Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *The Journal of Immunology 192*, 2689–2698.

(*29*) Goronzy, J. J., and Weyand, C. M. (2017). Successful and maladaptive T cell aging. *Immunity 46*, 364–378.

(*30*) Fulop, T. et al. (2010). Potential role of immunosenescence in cancer development. *Annals of the new York Academy of Sciences 1197*, 158–165.

(*31*) Goronzy, J. J., and Weyand, C. M. (2013). Understanding immunosenescence to improve responses to vaccines. *Nature immunology 14*, 428–436.

(*32*) Gil, A. et al. (2015). Narrowing of human influenza A virus-specific T cell receptor α and β repertoires with increasing age. *Journal of virology 89*, 4102–4116.

(*33*) Pinti, M. et al. (2016). Aging of the immune system: focus on inflammation and vaccination. *European journal of immunology 46*, 2286–2301.

(*34*) Del Giudice, G. et al. (2017). Fighting against a protean enemy: immunosenescence, vaccines, and healthy aging. *npj Aging and Mechanisms of Disease 4*, 1.

(*35*) Liu, W. et al. (2020). Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease. *Chinese medical journal 133*, 1032–1038.

(*36*) Zhou, F. et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet 395*, 1054–1062.

(*37*) Klein, S. L., and Flanagan, K. L. (2016). Sex differences in immune responses. *Nature Reviews Immunology 16*, 626–638.

(*38*) Van Lunzen, J., and Altfeld, M. (2014). Sex differences in infectious diseases–common but neglected. *The Journal of infectious diseases 209*, S79–S80.

(*39*) vom Steeg, L. G., and Klein, S. L. (2016). SeXX matters in infectious disease pathogenesis. *PLoS pathogens 12*, e1005374.

(*40*) Klein, S. L., Jedlicka, A., and Pekosz, A. (2010). The Xs and Y of immune responses to viral vaccines. *The Lancet infectious diseases 10*, 338–349.

(*41*) Cook, M. B. et al. (2009). Sex disparities in cancer incidence by period and age. *Cancer Epidemiology Biomarkers & Prevention 18*, 1174–1182.

(*42*) Dong, M. et al. (2020). Sex differences in cancer incidence and survival: a pan-cancer analysis. *Cancer Epidemiology, Biomarkers & Prevention 29*, 1389–1397.

(*43*) Yan, J. et al. (2010). The effect of ageing on human lymphocyte subsets: comparison of males and females. *Immunity & Ageing 7*, 1–10.

(*44*) Kverneland, A. H. et al. (2016). Age and gender leucocytes variances and references values generated using the standardized ONE-Study protocol. *Cytometry part A 89*, 543–564.

(*45*) Pido-Lopez, J., Imami, N., and Aspinall, R. (2001). Both age and gender affect thymic output: more recent thymic migrants in females than males as they age. *Clinical & Experimental Immunology 125*, 409–413.

(*46*) Stankiewicz, L. N. et al. (2025). Sex-biased human thymic architecture guides T cell development through spatially defined niches. *Developmental Cell 60*, 152–169.

(*47*) Armitage, P., and Doll, R. (2004). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British journal of cancer 91*, 1983–1989.

(*48*) Vineis, P., Schatzkin, A., and Potter, J. D. (2010). Models of carcinogenesis: an overview. *Carcinogenesis 31*, 1703–1709.

(*49*) Palmer, S. et al. (2018). Thymic involution and rising disease incidence with age. *Proceedings of the National Academy of Sciences 115*, 1883–1888.

(*50*) Liu, X., and Wu, J. (2018). History, applications, and challenges of immune repertoire research. *Cell biology and toxicology 34*, 441–457.

(*51*) Bradley, P., and Thomas, P. G. (2019). Using T cell receptor repertoires to understand the principles of adaptive immune recognition. *Annual review of immunology 37*, 547–570.

(*52*) Cowell, L. G. (2020). The Diagnostic, Prognostic, and Therapeutic Potential of Adaptive Immune Receptor Repertoire Profiling in CancerAdaptive Immune Receptor Repertoire Profiling in Cancer. *Cancer research 80*, 643–654.

(*53*) Carlson, C. S. et al. (2013). Using synthetic templates to design an unbiased multiplex PCR assay. *Nature communications 4*, 1–9.

(*54*) Robins, H. (2013). Immunosequencing: applications of immune repertoire deep sequencing. *Current opinion in immunology 25*, 646–652.

(*55*) Sender, R. et al. (2023). The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences 120*, e2308511120.

(*56*) Klein, L. et al. (2014). Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology 14*, 377–391.

(*57*) Lang, P. O., Govind, S., and Aspinall, R. (2013). Reversing T cell immunosenescence: why, who, and how. *Age 35*, 609–620.

(*58*) Duggal, N. A. (2018). Reversing the immune ageing clock: lifestyle modifications and pharmacological interventions. *Biogerontology 19*, 481–496.

(*59*) Aspinall, R., and Lang, P. O. (2018). Interventions to restore appropriate immune function in the elderly. *Immunity & Ageing 15*, 1–8.

(*60*) Lanfermeijer, J., Borghans, J. A., and van Baarle, D. (2020). How age and infection history shape the antigen-specific CD8+ T-cell repertoire: Implications for vaccination strategies in older adults. *Aging Cell 19*, e13262.

(*61*) Poisner, H. et al. (2024). Genetic determinants and phenotypic consequences of blood T-cell proportions in 207,000 diverse individuals. *Nature Communications 15*, 6732.

(*62*) Bentham, R. et al. (2025). ImmuneLENS characterizes systemic immune dysregulation in aging and cancer. *Nature Genetics*, 1–12.

(*63*) Manuel, M. et al. (2012). Lymphopenia combined with low TCR diversity (divpenia) predicts poor overall survival in metastatic breast cancer patients. *Oncoimmunology 1*, 432–440.

(*64*) Charles, J. et al. (2020). T-cell receptor diversity as a prognostic biomarker in melanoma patients. *Pigment cell & melanoma research 33*, 612–624.

(*65*) Kim, D. W. et al. Increased T-cell receptor repertoire diversity to predict better overall survival in gastrointestinal malignancies. 2021.

(*66*) Porciello, N. et al. (2022). T-cell repertoire diversity: friend or foe for protective antitumor response? *Journal of Experimental & Clinical Cancer Research 41*, 356.

(*67*) Wiegel, F. W., and Perelson, A. S. (2004). Some scaling principles for the immune system. *Immunology and cell biology 82*, 127–131.

(*68*) Zarnitsyna, V. I. et al. (2013). Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Frontiers in immunology 4*, 485.

(*69*) Harding, C., Pompei, F., and Wilson, R. (2012). Peak and decline in cancer incidence, mortality, and prevalence at old ages. *Cancer 118*, 1371–1386.

(*70*) Pavlidis, N., Stanta, G., and Audisio, R. A. (2012). Cancer prevalence and mortality in centenarians: a systematic review. *Critical reviews in oncology/hematology 83*, 145–152.

(*71*) Grulich, A. E. et al. (2007). Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *The Lancet 370*, 59–67.

(*72*) Ilham, S. et al. (2023). Cancer incidence in immunocompromised patients: a single-center cohort study. *BMC cancer 23*, 33.

(*73*) Kooshesh, K. A. et al. (2023). Health consequences of thymus removal in adults. *New England Journal of Medicine 389*, 406–417.

(*74*) Knudson Jr, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences 68*, 820–823.

(*75*) Sherr, C. J. (2004). Principles of tumor suppression. *Cell 116*, 235–246.

(*76*) Loomans-Kropp, H. A., and Umar, A. (2019). Cancer prevention and screening: the next step in the era of precision medicine. *NPJ precision oncology 3*, 3.

(*77*) Umar, A., Dunn, B. K., and Greenwald, P. (2012). Future directions in cancer prevention. *Nature Reviews Cancer 12*, 835–848.

(*78*) Sarhadi, V. K., and Armengol, G. (2022). Molecular biomarkers in cancer. *Biomolecules 12*, 1021.

(*79*) Postow, M. A. et al. (2015). Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. *Journal for immunotherapy of cancer 3*, 1–5.

(*80*) Kansy, B. A. et al. (2018). T cell receptor richness in peripheral blood increases after cetuximab therapy and correlates with therapeutic response. *Oncoimmunology 7*, e1494112.

(*81*) Guo, L. et al. (2020). Characteristics, dynamic changes, and prognostic significance of TCR repertoire profiling in patients with renal cell carcinoma. *The Journal of pathology 251*, 26–37.

(*82*) Han, J. et al. (2020). TCR repertoire diversity of peripheral PD-1+ CD8+ T cells predicts clinical outcomes after immunotherapy in patients with non–small cell lung cancer. *Cancer immunology research 8*, 146–154.

(*83*) Milano, F. et al. (2020). Impact of T cell repertoire diversity on mortality following cord blood transplantation. *Frontiers in Oncology 10*, 583349.

(*84*) Cardinale, A. et al. (2021). Thymic function and T-cell receptor repertoire diversity: implications for patient response to checkpoint blockade immunotherapy. *Frontiers in immunology 12*, 752042.

(*85*) Altan, M. et al. (2024). High peripheral T cell diversity is associated with lower risk of toxicity and superior response to dual immune checkpoint inhibitor therapy in patients with metastatic NSCLC. *Journal for immunotherapy of cancer 12*, e008950.

(*86*) Nikolich-Žugich, J. (2018). The twilight of immunity: emerging concepts in aging of the immune system. *Nature immunology 19*, 10–19.

(*87*) Emerson, R. O. et al. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature genetics 49*, 659–665.

(*88*) DeWitt III, W. S. et al. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife 7*, e38358.

(*89*) Snyder, T. M. et al. (2020). Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. *MedRxiv*.

(*90*) Greissl, J. et al. (2021). Immunosequencing of the T-cell receptor repertoire reveals signatures specific for diagnosis and characterization of early Lyme disease. *medRxiv*.

(*91*) Zahid, H. J. et al. (2024). Large-scale statistical mapping of T-cell receptor β sequences to Human Leukocyte Antigens. *BioRxiv*, 2024–04.

(*92*) Ortega, M. R. et al. (2024). Learning predictive signatures of HLA type from T-cell repertoires. *bioRxiv*, 2024–01.

(*93*) Mogilenko, D. A., Shchukina, I., and Artyomov, M. N. (2022). Immune ageing at single-cell resolution. *Nature Reviews Immunology 22*, 484–498.

(*94*) Valdes, A. M., Glass, D., and Spector, T. D. (2013). Omics technologies and the study of human ageing. *Nature Reviews Genetics 14*, 601–607.

(*95*) Rutledge, J., Oh, H., and Wyss-Coray, T. (2022). Measuring biological age using omics data. *Nature Reviews Genetics 23*, 715–727.

(*96*) Shen, X. et al. (2024). Nonlinear dynamics of multi-omics profiles during human aging. *Nature Aging*, 2024–01.

(*97*) Ferlay, J. et al. (2021). Cancer statistics for the year 2020: An overview. *International journal of cancer 149*, 778–789.

(*98*) Deo, S., Sharma, J., and Kumar, S. (2022). GLOBOCAN 2020 report on global cancer burden: challenges and opportunities for surgical oncologists. *Annals of surgical oncology 29*, 6497–6500.

(*99*) Robins, H. S. et al. (2009). Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood, The Journal of the American Society of Hematology 114*, 4099–4107.

(*100*) Robins, H. et al. (2012). Ultra-sensitive detection of rare T cell clones. *Journal of immunological methods 375*, 14–19.

(*101*) Dalai, S. C. et al. (2022). Clinical validation of a novel T-cell receptor sequencing assay for identification of recent or prior severe acute respiratory syndrome coronavirus 2 infection. *Clinical Infectious Diseases 75*, 2079–2087.

(*102*) National Cancer Institute SEER*Stat, version 8.4.3, 2024.

(*103*) Virtanen, P. et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17*, 261–272.