Tumor mutation burden signatures and relation to cancer risk in humans: A pan-cancer analysis

Xin Li[1] and D. Thirumalai[1]

Department of Chemistry, University of Texas, Austin, Texas, 78712, United States

(Dated: 20 April 2020)

Cancers, resulting in uncontrolled cell proliferation, are driven by accumulation of somatic mutations. Whole-genome sequencing has produced a catalogue of millions of somatic mutations, which contain the evolutionary history of the cancers. However, the connection between the extent and rate of mutation accumulation and disease development and risks are poorly understood. We analyzed data from 5,000 cancer patients with more than 1,200,000 mutations for 16 cancer types using The Cancer Genome Atlas (TCGA) database to unearth markers of tumor evolution. A strong correlation between Tumor Mutation Burden (TMB) and the Patient Age at Diagnosis (PAD) for cancers with low TMB (mean value less than 3 mutations per million base pairs) is absent in cancers with high TMB. The finding, demarcating cancers according to the TMB, resolves controversies relating the dynamics of mutation accumulation and disease risks for different cancers. We also find that the differences in cancer risk between the sexes are mainly driven by the disparity in mutation burden. A simple model, explaining these results, also show that immunotherapy could work well for elderly melanoma patients.

I. INTRODUCTION

Cancer is a complex disease caused by a gradual accumulation of somatic mutations, which occur through several distinct biological processes. Mutations could arise spontaneously during DNA replication[1], through exogenous sources, such as ultraviolet radiation, smoking, virus inflammation and carcinogens[2] or could be inherited[3]. During the past two decades, a vast amount of genomic data have been accumulated from patients with many cancers, thanks to developments in sequencing technology[4]. The distinct mutations contain the fingerprints of the cancer evolutionary history. The wealth of data have lead to the development of mathematical models[5–8], which provide insights into biomarkers that are reasonable predictors of the efficacy of cancer therapies[9]. Our study is motivated by the following considerations:

At what stage do most mutations accumulate? The relation between mutation accumulation during cancer progression and disease risks in different tissue types and populations are not only poorly understood but also have resulted in contradictory conclusions. For instance, it has been suggested that half or more of the genetic mutations in cancer cells are acquired before tumor initiation[10]. In contrast, analyses of colorectal cancer tumors[11] suggest that most of the mutations accumulate during the late stages of clonal expansion. In the cancer phylogenetic tree, representing tumor evolution, most mutations would appear in the trunk in the former case[10] while the branches would contain most of the mutations if the latter scenario holds[11]. Therefore, simple measures that discriminate between the two scenarios would be invaluable either for effective drug screening (given more targetable trunk mutations) or prediction of resistance (presence of more branched mutations)[12].

The determinants of cancer risk: The number of mutations accumulated in cancer cells is likely related to great variations in cancer risks among different human tissue types[13]. According to bad-luck theory[14], cancer risk across tissues is determined by the total number of adult stem cell (ASC) divisions during a lifetime, with enhanced mutations accumulating in those that divide frequently. The implication is that some tissue types are at a greater disease risk than others. In contrast, it has been shown that the ASCs gradually accumulate genetic mutations at a steady rate (about 40 per year) irrespective of the tissue type[15].

Patient age: Although cancer cells have evolved different mechanisms to escape immune surveillance[16], immunotherapy is fast becoming a very promising treatment for metastatic

cancer patients[17,18]. However, a large fraction of patients do not respond to immunotherapy positively, and suffer from the inflammatory side effects, and likely miss out on other alternative treatments[19,20]. Therefore, finding biomarkers that could be used to separate patients who show response from those who do not[19,21] is crucial. The number of non-synonymous mutations is one such independent biomarker[22,23]. Although mutations accumulate continuously in somatic cells with increasing age, the function of immune system also declines with age[24]. Thus, it is unclear whether the age of patients is a determining factor in deriving benefits from immunotherapy[24].

Motivated by the consideration described above, we developed computational methods to analyze the accumulation dynamics of somatic mutations[14,15] during cancer progression through a systematic pan-cancer study. We used the publicly available data produced by The Cancer Genome Atlas (TCGA) project, which contains a large and well-characterized patient cohorts from multiple cancer types[25]. In recent years, mutations in many cancers[26,27] have been analyzed at the base pair level, to arrive at a number of signatures, which are combination of mutation types that drives different mutational process. Here, we developed a simple coarse-grained measure of mutation load and a theoretical model that explains the distinct role of mutations in different cancers. We arrive at four major conclusions. (1) In the 16 cancers, we discovered two distinct signatures for the accumulation of somatic mutations. There is a strong correlation between the tumor mutation burden (TMB) and the patient age as long as the overall mutation load is below a critical value of ~ 3 mutations/Mb. In sharp contrast, the correlation is weak for cancers with TMB that exceeds the critical value. (2) Using a simple model, we show that if TMB is less than ~ 3 mutations/Mb, the majority of the mutations appear before the tumor initiation[10]. In the high TMB limit, most mutations accumulate during the tumor expansion stage, especially for some of the most lethal cancer types. (3) Although the mutation burden is one of the predominant factors that determines cancer risk, it only accounts for less than 50% of the differences in cancer risks among different tissues. Factors, such as the number of driver mutations required for each cancer, immunological and sex-hormone differences, also play a vital role in determining cancer risk. After excluding these factors, we show that the cancer risk disparity between the sexes is explained by the mutation load differences. (4) A byproduct of our work suggests that elderly patients could also benefit from immunotherapy.

II. RESULTS

Low Tumor Mutation Burden (TMB) strongly correlates with Patient Age at Diagnosis (PAD): We define TMB as the total number of non-synonymous and synonymous mutations per megabase (Mb) of the genome. Two mutation signatures emerge across many cancers (16 types are considered here with data taken from the TCGA database[25,28], see Table I in the Supplementary Information (SI)). A strong positive correlation between TMB and PAD is found for cancers with low overall mutation load (TMB < 3 mutations/Mb). However, such a correlation is absent for cancers with much higher TMB. The distinct mutation signatures are a reflection of different underlying evolutionary dynamics among these distinct cancer types.

Two key inferences can be drawn from Fig. 1. (1) The linear behavior, for the six cancers in Fig. 1, between TMB and PAD reveals strong correlation with a Pearson correlation coefficient (ρ) varying between 0.66 to 0.93. The strong correlation is also evident directly from the violin plots shown in Figs. S1-S3, which present the full mutation distribution for the cancers as a function of PAD. Note that if the overall mean TMB is less than 3 mutations/Mb, as is the case for LAML, THCA, KIRC, and KIRP cancers, the Pearson correlation coefficient is ≥ 0.85. Interestingly, the correlation is weaker as the TMB approaches and exceeds 3 mutations/Mb, as is the case for GBM and KICH (see Figs. 1e and 1f). The TMB in these cancers becomes higher as the PAD increases, which is similar to the behavior in normal tissues (see especially Fig. 1 here and Fig. 2b in[15]). (2) The mutation rates calculated from the slope of the solid lines in Fig. 1 vary from 13 (LAML) to 81 (KIRC) per year. The range is in line with the mutation rate observed in normal tissues (40 to 65 mutations per year for normal colon, small intestine, liver and oesophageal epithelia cells)[11,15,29]. Therefore, the cell mutation rate is approximately stable, and does not depend strongly on the tumor evolutionary process among the group of low TMB cancers (Fig. 1).

Large fraction of mutations accumulate before tumor initiation in low TMB cancers: The accumulation of mutations in cells in a tissue may be pictorially represented as a fish with the head, body and tail corresponding to the tissue development, self-renewal and tumor formation, respectively (see Fig. 1 in[10]). A prediction of the theory based on the "fish-like" picture is that the self-renewal stage (with a stable cell number) of a tissue is proportional to the patient age at tumor initiation. Therefore, a linear relation between the

number of somatic mutations and the PAD must exist, assuming that a fixed time elapses from tumor initiation to diagnosis[10]. This important prediction is supported by our analyses for low TMB cancers as shown in Fig. 1. Indeed, we can calculate the fraction of mutations accumulated in cells before tumor initiation once we know the average time it takes for tumor detection from the initiation time. Consider the thyroid carcinoma (THCA) as an example (Fig. 1b). The estimated latency period (τ_L, from tumor initiation to diagnosis) is 5 to 20 years[30]. The number of mutations that accumulate in this time frame would be 150 to 600 (assuming an average accumulation rate of 30 mutations per year, see Fig. 1**b**). The median age of THCA patient at diagnosis is 46, and the number of accumulated mutations at that age would be around 2100, which is the product of TMB and the human genome length. By noting that the TMB is around 0.7/Mb at 46 years (Fig. 1**b**), and the human genome length 3,000Mb we arrive at 2100 mutations. Thus, for τ_L=5 years only $\frac{150}{2100} \approx 7\%$ of the mutations occur during the period of tumor initiation till it is detected. This, implies that 93% and 71% of the mutations must have appeared before the initiation of the thyroid carcinoma for τ_L=5 and 20 years, respectively. In this case, the time from tumor initiation to detection is usually much shorter than the age of patients at diagnosis. From a similar analysis (see Table II in the SI), we conclude that the majority of mutations appear before the tumor initiation for cancers shown in Fig. 1. In the same figure, we also observe that the correlation between the TMB and patient age becomes weaker as the overall TMB increases. As noted above, the ρ value decreases from 0.9 to 0.7 as the overall mean value (averaged over all patients for different ages) of TMB increases from 0.4 mutations/Mb to 3.0 mutations/Mb.

High TMB and PAD are weakly correlated: As the overall TMB exceeds 3 mutations/Mb, there is no meaningful correlation between TMB and PAD (Fig. 2 and Figs. S4-S7). Interestingly, the results in Fig. 2 show that the TMB even decreases in certain cancer types as the PAD increases (negative correlation if one exists at all), see Fig. 2**f** for lung adenocarcinoma - LUAD as an example). This finding contradicts the conclusions reached in previous studies[10,31]. Clearly, the findings for cancers with high TMB cannot be explained by the fish-like model because of the absence of a linear relation between the TMB and PAD, which holds only for cancers with low TMB (Fig. 1).

A simple model explains the TMB data: To understand the drastically different behavior between cancers with low and high TMBs in Figs. 1 and 2, we propose a simple theoretical model. It has been noted that the mutation rate can increase by a few orders

of magnitude in tumor cells compared with normal cells caused by factors such as DNA mismatch repair deficiency or microsatellite instability[32,33]. As a consequence, a very high TMB would be detected in the tumor cells at the time of diagnosis. Therefore, the number of accumulated mutations (AM) in cells during the tumor formation stage could be comparable or even larger than the number during the tissue self-renewal stage. Let μ_1 be the mutation rate during normal DNA replication process, and μ_2 the enhanced mutation rate in tumor cells. The value of μ_2 would be zero if the mutation rate does not change during tumor progression. In this case, the number N_t of AM in tumor cells is given by $N_t = \mu_1 T$ where T is the age of the patient in years. The linear relation holds if the TMB is low (Fig. 1). However, the mutation rate increases in high TMB cancers (Fig. 2). Then, the number of mutations N_t is given by,

$$N_t = \mu_1 T + \mu_2 (T - T_I), \tag{1}$$

where the first term is the AM in the normal replication process, and the second term arises due to the accelerated mutations generated during the tumor formation stage. Note that T_I corresponds to the time at tumor initiation. Because the latency period $T - T_I$ for tumors from the initiation till diagnosis is likely to be similar for patients with the same type of cancer[10], the second term in Eq. (1) is likely to be a constant subject to minor variations. If the mutation rate $\mu_2 \gg \mu_1$, then the first term in Eq. (1) is negligible, which leads to the weak dependence of TMB on the patient age as found in Fig. 2.

It is instructive to calculate the fraction of accumulated mutations before tumor initiation for a cancer with high TMB. For the hepatocellular carcinoma (LIHC) the median age of patients at diagnosis is 61, so the number $N_I = \mu_1 T_I$ of AM before tumor initiation is less than $\mu_1 T \approx 3000$ (with rate $\mu_1 \approx 50$ mutations/year[11,15]). From Fig. 2**b**, which shows that the TMB is ≈4 mutations/Mb at the same age and taking the genome length to be 3,000Mb, the total number of mutations accumulated at age 61 is about 12,000 (4×3000). Thus, the fraction of AM before tumor initiation should be less than 25%. We surmise that in cancers with high TMB (Fig. 2) most of the somatic mutations occur during the late stages of clonal expansion (see Table III in the SI). This conclusion agrees with the observations in colorectal cancers[11], which we conclude must be a cancer type with high TMB (> 3 mutations/Mb).

Extrinsic Factors dominate mutation load in certain cancers: So far we have focused on somatic mutations arising from the intrinsic process due to DNA replication. The tissues of several cancers, such as head and neck, stomach, and skin, are constantly

exposed to ultraviolet (UV) radiation and chemical carcinogens[34]. These factors directly increase the mutation rate (the "extrinsic origin") of normal cells in the absence of DNA replication during the cell lifetime[35]. To account for extrinsic factors, we modify Eq. (1) by adding an additional term $N_{ext} = \mu_{ext}T$. As long as $\mu_{ext} \gg \mu_2$, a linear relation between TMB and patient age holds (see Fig. 3(a)-(c) and Figs. S8-S9 in the SI). In these cases, just as in Fig. 1, more mutations appear at the trunk of tumor phylogenetic trees types (see Table II in the SI and Fig. 3(d)). The strong linear correlation, with high Pearson correlation coefficient in Fig. 3(a)-(c), shows that in these cancers μ_{ext} far exceeds μ_1 and μ_2. It might be relatively easy for the immune system or drugs to clear cancer cells from patients with such cancers because the number of mutations are shared among all the cancer cells[12].

In contrast, the majority of the mutations are expected to be present in low frequencies and appear at the branches and leaves of the tumor phylogenetic tree (see Fig. 3(d)) for the cancers shown in Fig. 2. This results in immune surveillance escape and drug resistance[11]. In support of this finding, we found that the most deadly cancers (ones with the lowest 5-year survival rate), such as pancreatic cancer (8.5%), liver hepatocellular carcinoma (17.7%), and lung cancer (18.6%)[36] (see Fig. S10 in the SI), appear in Fig. 2.

Two universal patterns in cancers with low and high TMB: Our analyses above show two distinct signatures for TMB across many cancers. In order to elucidate if there are underlying universal behaviors in the cancers demarcated by the TMB, we first rescaled the TMB based on linear regression over the PAD for cancers listed in Figs. 1 and 3. Most interestingly, the data from the nine cancers fall on a simple linear curve, $\hat{\text{TMB}} \equiv (\text{TMB} - b_0)/a_0 = \text{PAD}$ (see Fig. 4**a**), which clearly demonstrates a clock-like mutational process in cancer evolution[37]. For the remaining cancers in Fig. 2, we rescaled the TMB only by the mean value μ_0 because of the absence of a linear relationship. Surprisingly, the data from these cancers also fall into a straight line but with a nearly *zero* slope (see the Pearson correlation coefficient ρ, P-value and the green line in Fig. 4**b**), which supports the absence of *any* correlation between TMB and PAD for these cancer types.

Next, we investigated the TMB distribution for each cancer type. Surprisingly, we found two universal TMB distribution patterns. A Gaussian distribution is found for all cancers analyzed in Figs. 1 and 3 (see Fig. S11 in the SI). After rescaling, the TMB distributions from the nine cancer types collapse onto a single curve described by the standard normal distribution (see Fig. 4**c**). The universal TMB distribution found in Fig. 4**c** for the nine

cancers can be explained by the Gaussian distribution of PAD (see Figs. S19 and S20 in the SI) and the linear relation between TMB and PAD (see Figs. 1 and 3). By simulating a population of patients with such properties, we obtain a Gaussian distribution for TMB (see Fig. S13 and the related discussion in the SI). We find it remarkable that for tumor evolution, which is a complex process, essentially a *single easily calculable parameter (TMB) from database* explains a vast amount of data for low TMB cancers.

Surprisingly, the distribution of cancers with large TMB is governed by a power-law (see Fig. S12 in the SI and Fig. 4**d**) (Fig. 2), except for PAAD, whose TMB is close to the critical value, and LUSC which have a Gaussian distribution. In these cancers, majority of the mutations accumulate after tumor initiation and the dynamics is likely to be highly complicated with many interrelated mutational processes contributing to clonal expansion. Similar signatures are found for cancers with low and high TMB, respectively, after we separate the patient data into two groups according to sex (see Figs. S14-S16 in the SI). The entirely different distributions of TMB found for the two groups of cancer show that cancers with high and low TMB follow distinct evolutionary process. In low (high) TMB cancers, most of the mutations occur before (after) disease initiation.

Taken together, we discovered two distinct universal signatures for TMB across many cancers, which could be understood in terms of the mean/median TMB (~ 3 mutations/Mb, see the green dashed line in Fig. 4**e** as the boundary and Figs. S17 and S18 in the SI for further support). The linear relation between TMB and PAD, and the universal Gaussian distribution of TMB in the first group of cancers imply a clock-like mutation dynamics holds[37]. In contrast, the power-law distribution of TMB and lack of linear relation between TMB and PAD suggest a more complex mutation dynamics in the other group of cancer.

Cancer risk and TMB: It is thought that the continuous accumulation of mutations drives the transformation of normal cells to cancerous cells. However, there is a controversy on the origin of the great variations in cancer risk among different tissues. A strong correlation is found between cancer risk in different tissues and the total number of adult stem cell divisions[14]. Therefore, the variations of cancer risk might be mainly determined by the mutation burden in the tissues, which seems to be consistent with the extreme variations of TMB observed in different cancer types (see Fig. 4**e**). We now assess the correlation between the cancer risk and the mutation burden in cancer patients directly by taking data for the 16 cancers considered above from the TCGA and SEER database (see Tables IV, V and the

materials in the SI). The cancer risk for males and females vary greatly (see Table V in the SI), which we discuss further in the following sections. Here, we include the data for both the sexes. The median age for all cancer patients regardless of sex differences at diagnosis is around 66 years[13]. In order to rationalize the age disparity in different cancers for both the sexes (see Figs. S19-S20, and Table VI in the SI), we adjusted the value of TMB for both the sexes. We assumed that both the sexes accumulate mutations in a similar way (see Figs. S21-S23 and Section VIII in the SI for validation). We investigated the relation between cancer risk and TMB (see Fig. S24 (a) and (b)). The Pearson correlation coefficient ρ is 0.6 with a P value of 2.67×10^{-4} ($\rho = 0.7$ with $P = 2.70 \times 10^{-5}$) between cancer risk and TMB for 16 (15) cancers. Additional results are shown in Fig. S24 (c) and (d) without age-adjustment. A similar Pearson correlation coefficient ($\rho = 0.66$) is also found between cancer risk and mutation frequency across 41 cancers (using double-logarithmic coordinates with the data from a uniform sequencing pipeline) in a recent study[38]. Therefore, the cancer risk and TMB are correlated.

Because TMB accounts for less than 50% ($\approx 0.7^2$) of differences in cancer risks among different tissues other factors must be taken into account to explain more quantitatively cancer risks among all tissues. These include the driver mutation number[39], immunological[40] and sex-hormone differences[41]. Nevertheless, the mutation burden is one of the predominant factors for predicting cancer risks across different tissues. If we only compare cancer risks of the same type for both the sexes, excluding gender-specific cancers (with 16 types of cancers considered here as shown in Fig. 1-3), we might be able to exclude the influences of these factors on cancer incidence, which would allow us to focus only on TMB.

TMB explains the role of sex in cancer risk: The risk ratio for the male to female changes from 0.39 (thyroid) to 28.73 (Kaposi sarcoma)[42]. For most types of cancer, men have a higher risk compared to women. Although breast and prostate cancers are gender-specific, driven by sex steroid hormones[41], the enhanced risk in men is still a puzzle for many other non-gender-specific cancers. There are many potential factors leading to the disparity. These include but are not limited to genetic and expression variation[43], sex hormones[44], environmental exposures[45].

We investigated the relation between cancer risk and the mutation burden for both the sexes for each cancer separately (see Fig. 5(a)-(b)). Surprisingly, the cancer risk can be explained solely by the TMB score in 13 out of the 16 cancers. Nine of them, which show

high cancer risk, also have high TMB score (HNSC, KIRC, KIRP, LIHC, THCA, BLCA, LUAD, LUSC, and SKCM), and four of them show almost the same risk and have similar TMB (GBM, LAML, PAAD, and READ). The negative correlation between cancer risk and TMB in KICH in Fig. 5 might be because the sample size (66 in total) is small. In this case there is a large deviation in the median age of patients between men and women populations (see Fig. S20g). The lower risk, but higher TMB in women in STAD and COAD, could be caused by other factors such as the immune response and sex-hormone. Inflammation is frequently observed in these two cancer types, which results in the activation of the immune system[46]. Because the efficiency of the immune system declines faster in men[40], and estrogen suppresses tumor growth in female patients with COAD[47], they might decrease the cancer risk in women with STAD and COAD.

Taken together our analyses show that the risk disparity between men and women for the same type of cancer can be explained mainly through the mutation burden in the cancer cells (see the schematic illustration in Fig. 5(c)). However, other factors on cancers have to be considered to explain the risk among different tissues as shown in Fig. S24 and in previous studies[14,38]. The number (N_d) of driver mutations required for cancer emergence is an important factor[39]. Because it takes a long time for a normal cell to accumulate driver mutations[48] it would reduce the cancer risk for tissues with a higher threshold of N_d. A well-known example is the African elephant[49]. Consider THCA for instance. It is thought that a single driver mutation is required for the transition from a normal cell to a cancerous cell[39]. This might explain the relatively higher risk (see Fig. 5**a**) of THCA with low TMB compared to other cancers with lower risk but higher mutation burden, such as GBM which might require 6 driver mutations[39]. Although the mutation burden for SKCM is much higher than other cancers such as LUAD (HNSC), the large number of drivers (11) required for SKCM emergence could be the critical factor resulting in a similar risk as LUAD (HNSC), which might only require 4 (5) driver mutations (see Fig. 5**b**)[39].

TMB and immunotherapy for aging patients: We have shown that high TMB is frequently related to high cancer risk for both the sexes for the same type of cancer, as illustrated above. Therefore, high TMB is a foe of cancer patients. Although not desirable, high TMB could also elicit enhanced expression of a large number of neoantigens in cancer cells, which could be recognized by T-cells for apoptosis[50]. However, cancer cells have evolved different escape pathways, and even disarm immune surveillance[16]. As we gain

more knowledge about the mechanisms of these critical processes, the immunotherapy may be modified to truly tame many metastatic cancers[51] regardless of the TMB. At present, the FDA approved treatments are successful in melanoma, non-small-cell lung, head and neck, colorectal, urothelial cancers and so on[18]. Most of these cancers have high TMB as observed in Figs.2-3.

Efficiency of the the immune system declines with age[52], which might be a reason why the pivotal clinical trials leading to FDA approval are conducted mainly on relatively young patients (with the median age of 57.5)[53,54]. However, the majority (about two-thirds) of cancers occur in humans over 65 years[55]. Therefore, it is important to know whether older patients may also benefit from immunotherapy. In this context, the preclinical data are controversial. Some studies indicate a less-efficient response for older patients[56,57], while others show no overall difference or even the opposite trend[58,59].

We investigated the influence of age on patient response to immunotherapy. As an example, we first examined the age distribution for all patients and the ones who show favorable response to immune checkpoint blockade from a clinical study for metastatic melanoma[60] (see Fig. 6**a**). The age varies considerably in both the cases. The median age for patients who show response is 65 years while it is 61.5 years for the whole population. We also find that the old patients (≥ 65) show a similar response (27%) as young patients (22%) even though there are more young patients (< 65) in this study. The number of non-synonymous mutations for patients for the same cancer also varies greatly, as illustrated in Fig. 6**b**. The median value (309) of mutations for patients showing response is much higher than the value (197) of the whole population. A significant difference exists in the TMB between patients with positive and negative response to immunotherapy[60]. In order to remove the influence of this factor and focus solely on age, we compared the patient age with different responses with similar TMB. For a similar mutation burden level, (<5% difference), patients showing favorable response can be either young or old as shown in Fig. 6**c**. However, in this group we find a much higher fraction of old patients compared to young patients (63% versus 37%). A similar conclusion was reached by analyzing the data from other immunotherapy studies[23,61] with 75% and 100% fraction of older patients showing response but the sample sizes are smaller (12 and 6 cases, respectively).

We explain the data in Fig. 6**c** using the dynamics of accumulation of mutations discussed above. The number of mutations shows a strong linear relation with the age of melanoma

patients (see Fig. 3**c**). A similar time period (τ_d) for all the patients would be expected from the initiation to the detection of melanoma[10]. Consider two patients A and B with ages T_A and T_B with the same TMB, denoted by N_T. The fraction ($F_i^{A/B}$) of mutations accumulated before the tumor initiation for both the patients is,

$$F_i^{A/B} = \frac{N_T - \alpha_{A/B}\tau_d}{N_T}, \tag{2}$$

where $\alpha_{A/B}$ is the rate for the mutation accumulation in patient A or B. If $T_A > T_B$, then the mutation rate obeys $\alpha_A < \alpha_B$ because of N_T for both the patients is the same. Therefore, the older patient would accumulate more mutations ($F_i^A > F_i^B$) in the trunk of the tumor phylogenetic tree. A better response to immunotherapy would be expected if all tumor cells contain the neoantigens recognized by T-cells compared with the case that these neoantigens only appear in a tiny fraction of cell populations. This model explains the higher fraction of older patients showing response to immunotherapy with the same TMB, noted in the clinical studies cited above.

III. DISCUSSION

We investigated the evolutionary signatures focusing on the mutation burden in cancer patients, and their consequences on disease risks, variations between the sexes, and age related responses to treatment. By analyzing the data for 16 cancers, we found that two distinct scenarios for the accumulation of somatic mutations during cancer progression is determined by a single parameter. If the overall mutation burden is low (< 3 mutations/Mb), a strong positive correlation exists between the number of mutations and the patient age at diagnosis, which is absent for high TMB cancers. If the TMB is suitably rescaled then the distribution is a Gaussian for cancers with low mutational load. In contrast, a power-law distribution of TMB is found for high TMB cancers. We also found a few exceptions with high TMB where the positive linear relation between these two quantities is recovered for cancers such as HNSC, STAD, and SKCM. Interestingly, all these types of cancer are strongly influenced by environment[34], which we explained by including the effect of extrinsic factors on mutations within the framework of a simple model.

For cancers that show a linear relation between the TMB and PAD, half or more mutations appear before tumor initiation[10], which might be due to the relatively short time from

tumor initiation to diagnosis compared to the duration of the tissue self-renewal stage. In sharp contrast, the majority of the mutations appear during the tumor expansion stage for other types of cancer, which exhibit weak correlation between the TMB and PAD, such as colorectal cancer[11]. In other words, the majority of the somatic mutations appear in the trunk of the cancer phylogenetic tree for cancers with low TMB while most of them would arise in the branches for high TMB cancers.

The evolution of mutations in cancer cells, quantified using TMB, should serve as a valuable prognostic biomarker, allowing for a more informed treatment. One example from immunotherapy is discussed here. The fraction of patients showing response to immunotherapy is much higher for aging patients than young patients with similar TMB. This might be due to more mutations that appear in the trunk of the phylogenetic tree in old patients. Therefore, aging patients should also benefit from this new treatment method as well as or even better than young patients, even if they have a compromised immune system.

The cancer risk for different tissues has been reported to be strongly correlated with the number of the corresponding stem cell divisions[14] while a recent research claims that the cancer risk is strongly correlated with the mutation load[38]. We also find that the mutation burden is indeed correlated with the cancer risk. Interestingly, mutation burden alone explains the disparity of cancer risk found between the sexes after we excluding sex-hormone-related cancers. However, the present and previous studies[38] with data from a uniform sequencing pipeline, show that the mutation load alone cannot account for all the differences in cancer risk among different tissues. Therefore, factors such as the number of driver mutations, immune responses, and sex hormones must be important in understanding the variations in the cancer risks among different tissues.

A deeper understanding of the evolutionary process is the key to cure cancer. Although, we have investigated this process for only 16 cancers covering a large population of patients, it would be valuable to assess if our findings are also applicable to other cancer types. Additional data will be needed to verify the correlation between cancer risk and the mutation load for KICH and other cancers between the genders. We have assumed that the mutation load found in cancer patients reflects the mutations accumulated in ancestral cells[38]. Therefore, we need more mutation profiles from normal and precancerous lesions to further validate the conclusions reached here[11,15,29]. In addition, the influences of other types of mutations such as copy number variations, and aneuploidy should also be considered in future. Due to

the complexity of cancer diseases and presence of many different risk factors simultaneously, it is valuable to evaluate one factor at a time (even if additivity principle does not hold), if at all possible, to better understand their roles in cancer progression. It is estimated that 73% of cancer patients will be 65 years and older by 2040[62]. Nevertheless, elderly patients are often excluded from promising treatments such as the immunotherapy. It is urgent and necessary to have more clinical studies supplemented with computational models to appraise the response of elderly patients to all new cancer treatment strategies.

Acknowledgments: We are grateful to Sumit Sinha for discussions and comments on the manuscript. This work is supported by the National Science Foundation (PHY 17-08128), and the Collie-Welch Chair through the Welch Foundation (F-0019).

REFERENCES

[1]Thomas A Kunkel and Katarzyna Bebenek. Dna replication fidelity. *Annual review of biochemistry*, 69(1):497–529, 2000.

[2]Errol C Friedberg, Lisa D McDaniel, and Roger A Schultz. The role of endogenous and exogenous dna damage and mutagenesis. *Current opinion in genetics & development*, 14(1):5–10, 2004.

[3]Eric R Fearon. Human cancer syndromes: clues to the origin and nature of cancer. *Science*, 278(5340):1043–1050, 1997.

[4]Levi A Garraway and Eric S Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, 2013.

[5]Philipp M Altrock, Lin L Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12):730, 2015.

[6]Abdul N Malmi-Kakkada, Xin Li, Himadri S Samanta, Sumit Sinha, and D. Thirumalai. Cell growth rate dictates the onset of glass to fluidlike transition and long time superdiffusion in an evolving cell colony. *Physical Review X*, 8(2):021025, 2018.

[7]Martin A Nowak, Franziska Michor, and Yoh Iwasa. The linear process of somatic evolution. *Proceedings of the national academy of sciences*, 100(25):14966–14969, 2003.

[8]Xin Li and D. Thirumalai. Share, but unequally: a plausible mechanism for emergence and maintenance of intratumour heterogeneity. *Journal of the Royal Society Interface*, 16(150):20180820, 2019.

[9]Marta Łuksza, Nadeem Riaz, Vladimir Makarov, Vinod P Balachandran, Matthew D Hellmann, Alexander Solovyov, Naiyer A Rizvi, Taha Merghoub, Arnold J Levine, Timothy A Chan, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*, 551(7681):517, 2017.

[10]Cristian Tomasetti, Bert Vogelstein, and Giovanni Parmigiani. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences*, 110(6):1999–2004, 2013.

[11]Sophie F Roerink, Nobuo Sasaki, Henry Lee-Six, Matthew D Young, Ludmil B Alexandrov, Sam Behjati, Thomas J Mitchell, Sebastian Grossmann, Howard Lightfoot, David A Egan, et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, 556(7702):457, 2018.

[12]Nabil Amirouchene-Angelozzi, Charles Swanton, and Alberto Bardelli. Tumor evolution as a therapeutic target. *Cancer discovery*, 7(8):805–817, 2017.

[13]N Howlader, AM Noone, M Krapcho, N Neyman, R Aminou, SF Altekruse, CL Kosary, J Ruhl, Z Tatalovich, H Cho, et al. SEER Cancer Statistics Review, 1975-2009 (vintage 2009 populations), National Cancer Institute. Bethesda, MD. *MD, USA*, 2012.

[14]Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

[15]Francis Blokzijl, Joep De Ligt, Myrthe Jager, Valentina Sasselli, Sophie Roerink, Nobuo Sasaki, Meritxell Huch, Sander Boymans, Ewart Kuijk, Pjotr Prins, et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260, 2016.

[16]Ryungsa Kim, Manabu Emi, and Kazuaki Tanabe. Cancer immunoediting from immune surveillance to immune escape. *Immunology*, 121(1):1–14, 2007.

[17]Suzanne L Topalian, Charles G Drake, and Drew M Pardoll. Immune checkpoint blockade: a common denominator approach to cancer therapy. *Cancer cell*, 27(4):450–461, 2015.

[18]Michael A Postow, Robert Sidlow, and Matthew D Hellmann. Immune-related adverse events associated with immune checkpoint blockade. *New England Journal of Medicine*, 378(2):158–168, 2018.

[19]Suzanne L Topalian, Janis M Taube, Robert A Anders, and Drew M Pardoll. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nature Reviews Cancer*, 16(5):275, 2016.

[20]Padmanee Sharma, Siwen Hu-Lieskovan, Jennifer A Wargo, and Antoni Ribas. Primary,

adaptive, and acquired resistance to cancer immunotherapy. *Cell*, 168(4):707–723, 2017.

[21]Padmanee Sharma and James P Allison. The future of immune checkpoint therapy. *Science*, 348(6230):56–61, 2015.

[22]Aaron M Goodman, Shumei Kato, Lyudmila Bazhenova, Sandip P Patel, Garrett M Frampton, Vincent Miller, Philip J Stephens, Gregory A Daniels, and Razelle Kurzrock. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Molecular cancer therapeutics*, pages molcanther–0386, 2017.

[23]Matthew D Hellmann, Tavi Nathanson, Hira Rizvi, Benjamin C Creelan, Francisco Sanchez-Vega, Arun Ahuja, Ai Ni, Jacki B Novik, Levi MB Mangarin, Mohsen Abu-Akeel, et al. Genomic features of response to combination immunotherapy in patients with advanced non-small-cell lung cancer. *Cancer cell*, 33(5):843–852, 2018.

[24]Amaury Daste, Charlotte Domblides, Marine Gross-goupil, Camille Chakiba, Amandine Quivy, Valérie Cochin, Erwan de Mones, Nicolas Larmonier, Pierre Soubeyran, and Alain Ravaud. Immune checkpoint inhibitors and elderly people: A review. *European Journal of Cancer*, 82:155–166, 2017.

[25]John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[26]Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[27]Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259, 2013.

[28]Mehmet Kemal Samur. RTCGAToolbox: a new tool for exporting TCGA Firehose data. *PloS one*, 9(9):e106397, 2014.

[29]Akira Yokoyama, Nobuyuki Kakiuchi, Tetsuichi Yoshizato, Yasuhito Nannya, Hiromichi Suzuki, Yasuhide Takeuchi, Yusuke Shiozawa, Yusuke Sato, Kosuke Aoki, Soo Ki Kim, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*, 565(7739):312, 2019.

[30]AF Tsyb, EM Parshkov, VV Shakhtarin, VF Stepanenko, VF Skvortsov, and IV Chebotareva. Thyroid cancer in children and adolescents of Bryansk and Kaluga Regions. *Age*, 4(10-14):15–19, 1996.

[31]Dmitriy I Podolskiy, Alexei V Lobanov, Gregory V Kryukov, and Vadim N Gladyshev. Analysis of cancer genomes reveals basic features of human aging and its role in cancer development. *Nature communications*, 7:12157, 2016.

[32]Christoph Lengauer, Kenneth W Kinzler, and Bert Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643, 1998.

[33]Lawrence A Loeb. A mutator phenotype in cancer. *Cancer research*, 61(8):3230–3239, 2001.

[34]Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, 2017.

[35]Song Wu, Scott Powers, Wei Zhu, and Yusuf A Hannun. Substantial contribution of extrinsic risk factors to cancer development. *Nature*, 529(7584):43, 2016.

[36]AM Noone, N Howlader, M Krapcho, D Miller, A Brest, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, et al. SEER Cancer Statistics Review, 1975-2015, National Cancer Institute. Bethesda, MD, 2018.

[37]Ludmil B Alexandrov, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. Clock-like mutational processes in human somatic cells. *Nature genetics*, 47(12):1402, 2015.

[38]Dapeng Hao, Li Wang, and Li-jun Di. Distinct mutation accumulation rates among tissues determine the variation in cancer risk. *Scientific Reports*, 6:19458, 2016.

[39]Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041, 2017.

[40]Sam Palmer, Luca Albergante, Clare C Blackburn, and TJ Newman. Thymic involution and rising disease incidence with age. *Proceedings of the National Academy of Sciences*, 115(8):1883–1888, 2018.

[41]Gail P Risbridger, Ian D Davis, Stephen N Birrell, and Wayne D Tilley. Breast and prostate cancer: more similar than different. *Nature Reviews Cancer*, 10(3):nrc2795, 2010.

[42]Michael B Cook, Sanford M Dawsey, Neal D Freedman, Peter D Inskip, Sara M Wichner, Sabah M Quraishi, Susan S Devesa, and Katherine A McGlynn. Sex disparities in cancer

incidence by period and age. *Cancer Epidemiology and Prevention Biomarkers*, 18(4):1174–1182, 2009.

[43]Yuan Yuan, Lingxiang Liu, Hu Chen, Yumeng Wang, Yanxun Xu, Huzhang Mao, Jun Li, Gordon B Mills, Yongqian Shu, Liang Li, et al. Comprehensive characterization of molecular differences in cancer between male and female patients. *Cancer cell*, 29(5):711–722, 2016.

[44]Shelia Hoar Zahm and Joseph F Fraumeni Jr. Racial, ethnic, and gender variations in cancer risk: considerations for future epidemiologic research. *Environmental health perspectives*, 103(Suppl 8):283, 1995.

[45]Frederica P Perera. Environment and cancer: who are susceptible? *Science*, 278(5340):1068–1073, 1997.

[46]Lisa M Coussens and Zena Werb. Inflammation and cancer. *Nature*, 420(6917):860, 2002.

[47]M Tevfik Dorak and Ebru Karpuzoglu. Gender differences in cancer susceptibility: an inadequately addressed issue. *Frontiers in genetics*, 3:268, 2012.

[48]Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[49]Lisa M Abegglen, Aleah F Caulin, Ashley Chan, Kristy Lee, Rosann Robinson, Michael S Campbell, Wendy K Kiso, Dennis L Schmitt, Peter J Waddell, Srividya Bhaskara, et al. Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. *JAMA*, 314(17):1850–1860, 2015.

[50]Ton N Schumacher and Robert D Schreiber. Neoantigens in cancer immunotherapy. *Science*, 348(6230):69–74, 2015.

[51]Ira Mellman, George Coukos, and Glenn Dranoff. Cancer immunotherapy comes of age. *Nature*, 480(7378):480, 2011.

[52]Phyllis Jean Linton and Kenneth Dorshkind. Age-related changes in lymphocyte development and function. *Nature immunology*, 5(2):133, 2004.

[53]F Stephen Hodi, Steven J O'day, David F McDermott, Robert W Weber, Jeffrey A Sosman, John B Haanen, Rene Gonzalez, Caroline Robert, Dirk Schadendorf, Jessica C Hassel, et al. Improved survival with ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine*, 363(8):711–723, 2010.

[54]Caroline Robert, Luc Thomas, Igor Bondarenko, Steven O'day, Jeffrey Weber, Claus Garbe, Celeste Lebbe, Jean-François Baurain, Alessandro Testori, Jean-Jacques Grob, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *New England Journal of Medicine*, 364(26):2517–2526, 2011.

[55]A Jemal, R Siegel, E Ward, Y Hao, J Xu, T Murray, and MJ Thun. Cancer statistics, 2005. *CA Cancer J Clin*, 55:10–30, 2005.

[56]Robert L Ferris, George Blumenschein Jr, Jerome Fayette, Joel Guigay, A Dimitrios Colevas, Lisa Licitra, Kevin Harrington, Stefan Kasper, Everett E Vokes, Caroline Even, et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *New England Journal of Medicine*, 375(19):1856–1867, 2016.

[57]Robert J Motzer, Bernard Escudier, David F McDermott, Saby George, Hans J Hammers, Sandhya Srinivas, Scott S Tykodi, Jeffrey A Sosman, Giuseppe Procopio, Elizabeth R Plimack, et al. Nivolumab versus everolimus in advanced renal-cell carcinoma. *New England Journal of Medicine*, 373(19):1803–1813, 2015.

[58]Vanna Chiarion Sileni, Jacopo Pigozzo, Paolo Antonio Ascierto, Antonio Maria Grimaldi, Michele Maio, Lorenza Di Guardo, Paolo Marchetti, Francesco De Rosa, Carmen Nuzzo, Alessandro Testori, et al. Efficacy and safety of ipilimumab in elderly patients with pretreated advanced melanoma treated at italian centres through the expanded access programme. *Journal of Experimental & Clinical Cancer Research*, 33(1):30, 2014.

[59]Antoni Ribas, Omid Hamid, Adil Daud, F Stephen Hodi, Jedd D Wolchok, Richard Kefford, Anthony M Joshua, Amita Patnaik, Wen-Jen Hwu, Jeffrey S Weber, et al. Association of pembrolizumab with tumor response and survival among patients with advanced melanoma. *JAMA*, 315(15):1600–1609, 2016.

[60]Eliezer M Van Allen, Diana Miao, Bastian Schilling, Sachet A Shukla, Christian Blank, Lisa Zimmer, Antje Sucker, Uwe Hillen, Marnix H Geukes Foppen, Simone M Goldinger, et al. Genomic correlates of response to CTLA4 blockade in metastatic melanoma. *Science*, 350:207–211, 2015.

[61]Naiyer A Rizvi, Matthew D Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J Havel, William Lee, Jianda Yuan, Phillip Wong, Teresa S Ho, et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science*, 348(6230):124–128, 2015.

[62]Shirley M Bluethmann, Angela B Mariotto, and Julia H Rowland. Anticipating the "Silver

Tsunami": prevalence trajectories and comorbidity burden among older cancer survivors in the United States. *Cancer Epidemiol Biomarkers Prev*, 25:1029–1036, 2016.

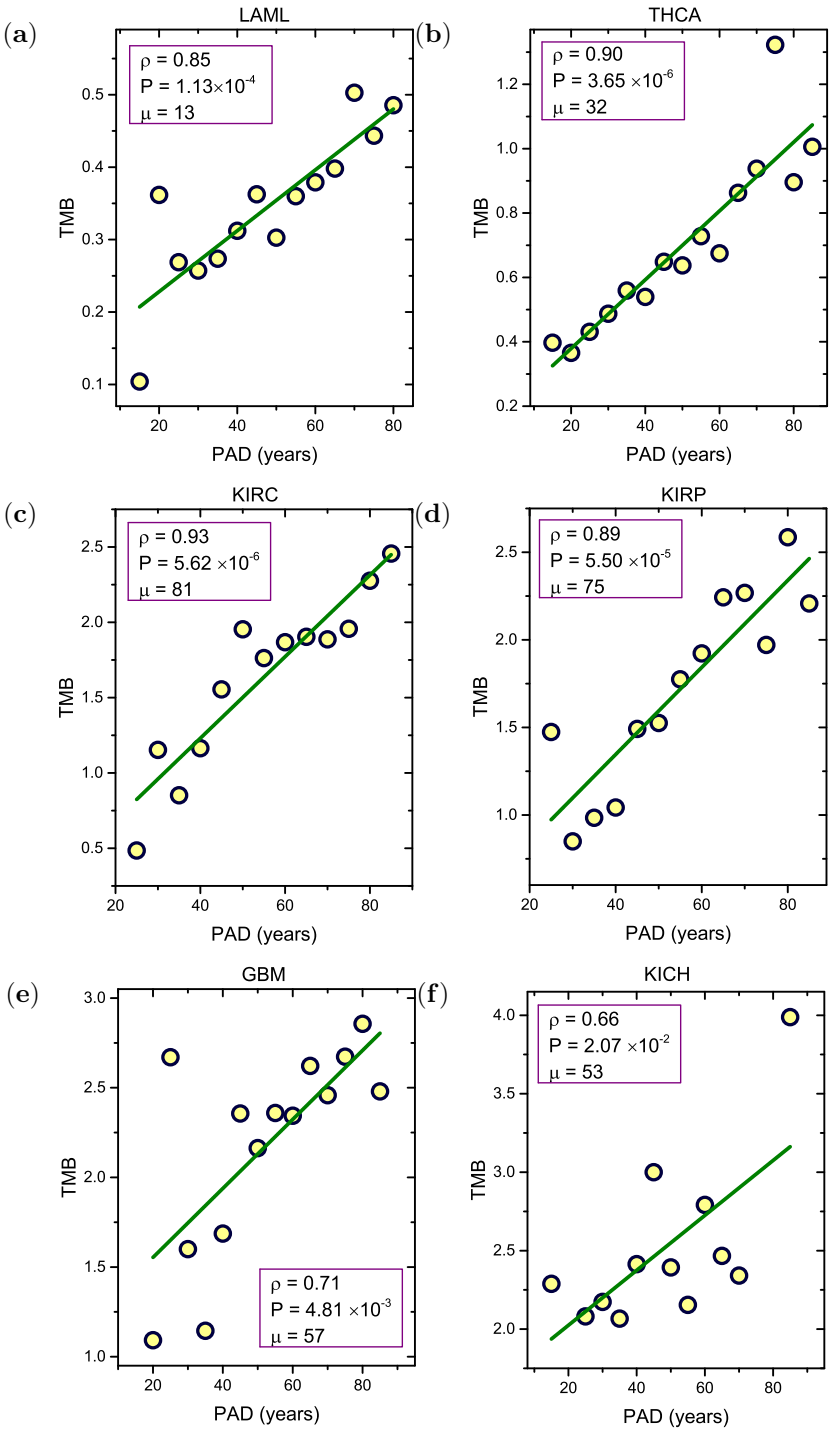

FIG. 1: The tumor mutation burden, TMB, (total number of non-synonymous and synonymous mutations per megabase) as a function of patient age at diagnosis (PAD). The mean value of TMB is used over a 5-year period. The green line shows the regression line. The Pearson correlation coefficient ρ, the significant P value from an F test and also the mutation rate μ (number of mutations per year) are listed in the figures. Good correlation holds only if TMB is low. Data are taken from TCGA[25].

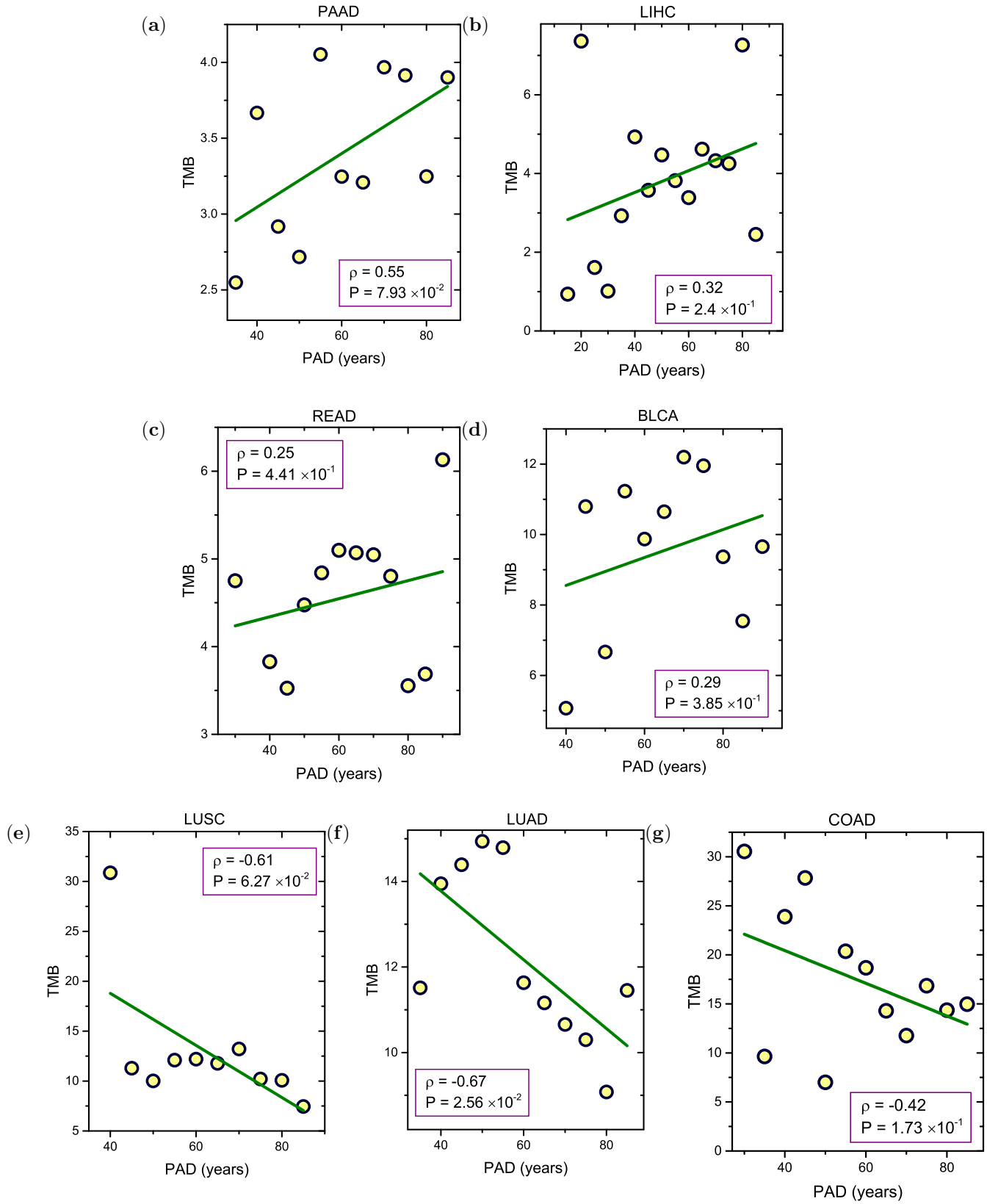

FIG. 2: Same as Fig.1 except the data are for cancers with high TMN. The green regression lines show a lack correlation between TMB and patient age.

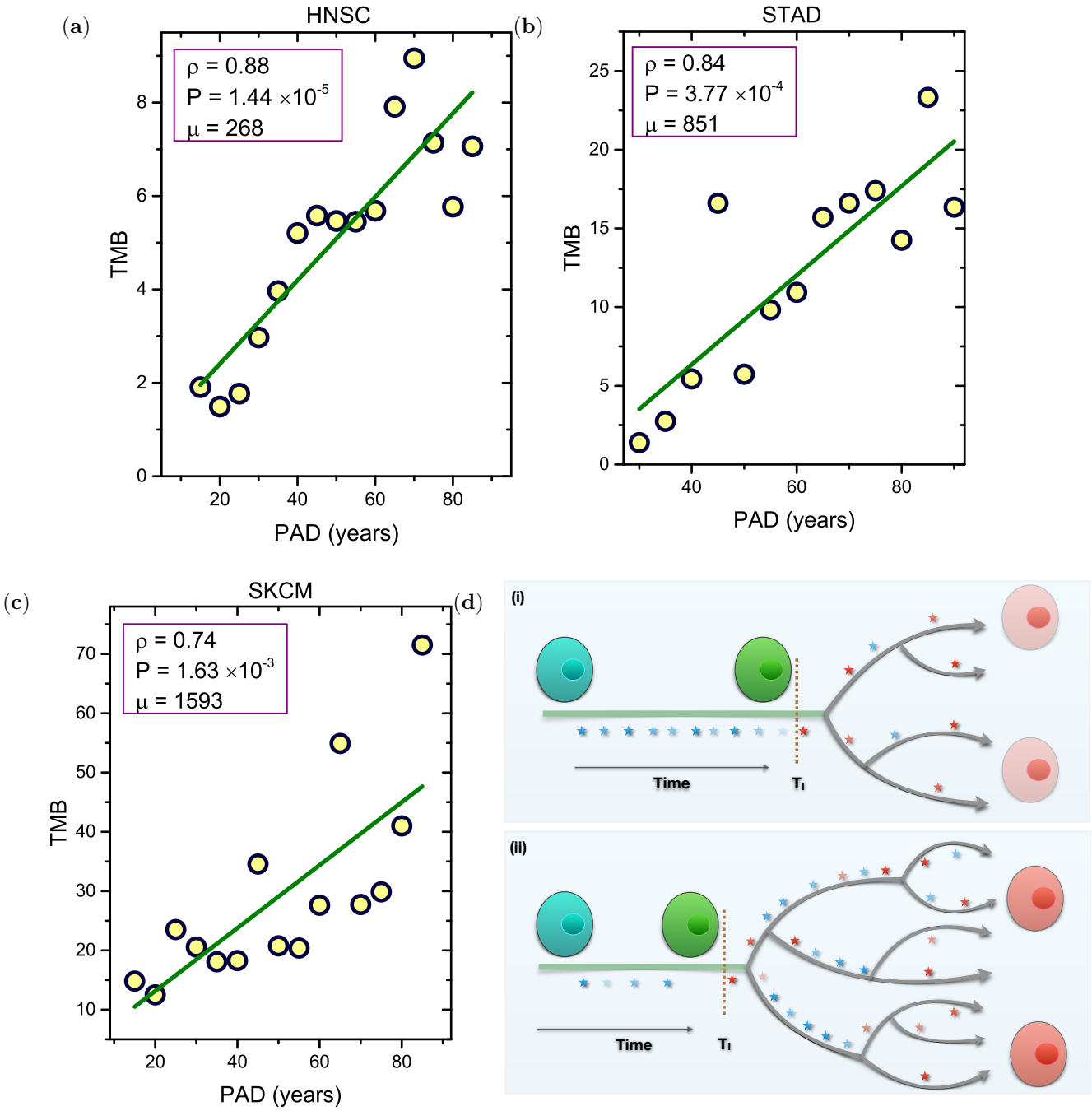

FIG. 3: (a)-(c) TMB as a function of age for cancers strongly influenced by environments. The regression lines are in green. The Pearson correlation coefficient ρ, the significant P value, and the mutation rate μ (number of mutations per year) are in the figures. (d) Schematic of the mutation accumulation in cells during cancer progression. (i) corresponds to the case in Fig. 1 and Fig. 3(a)-(c), and (ii) shows the case in Fig. 2. Mutations are indicated by stars, and T_I is the tumor initiation time. Normal cells are in blue/green colors without/with neutral mutations (blue stars) while cancer cells are in red color with driver mutations (red stars).

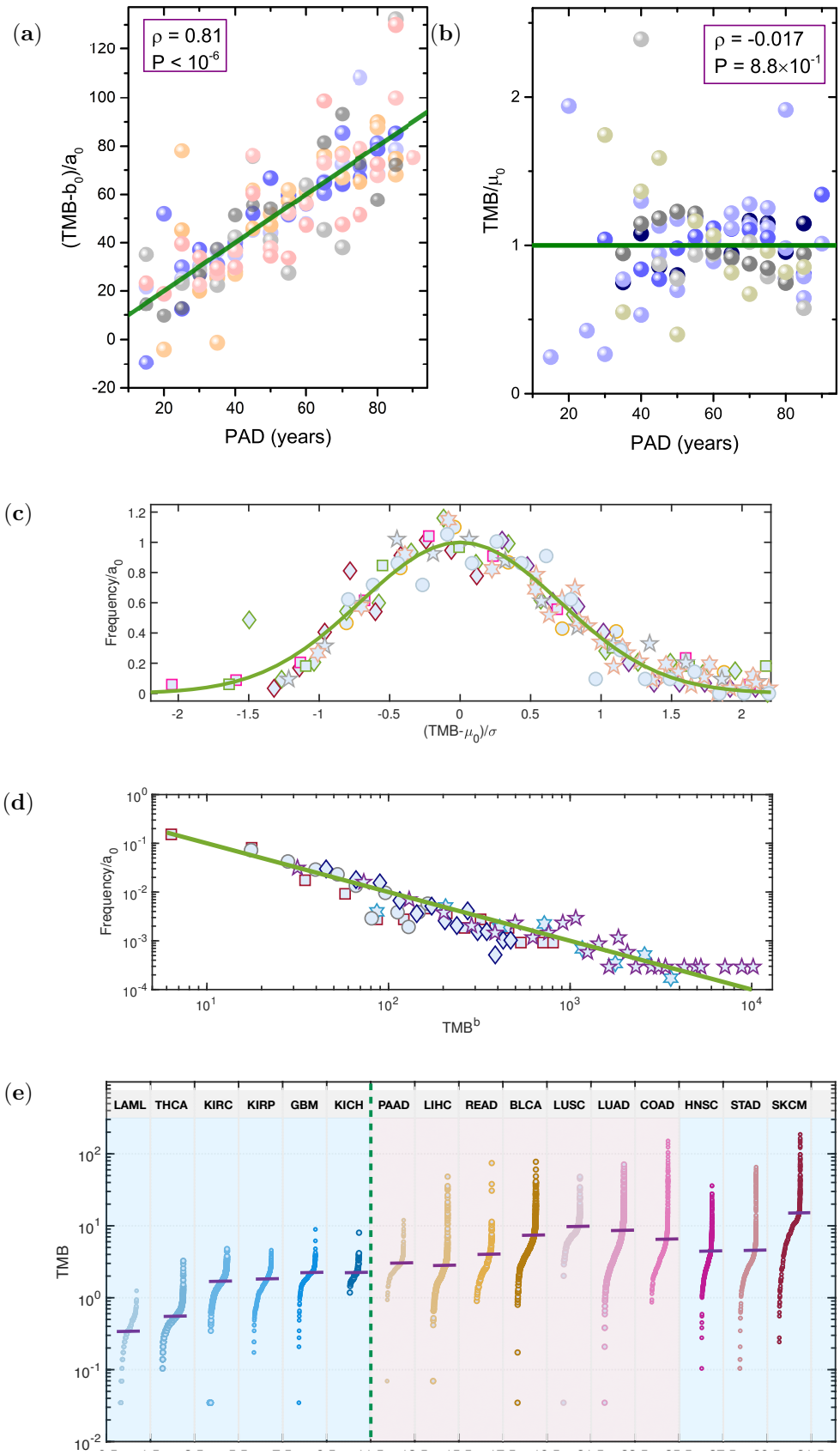

FIG. 4: See next page for caption.

FIG. 4: The rescaled TMB versus PAD for the 16 cancer types. (a)-(b) All nine types of cancer in Figs. 1 and 3 are displayed together in (a) after the TMB is rescaled by the slope (a_0) and intercept (b_0) of the corresponding linear regression lines. The other seven types of cancers (Fig. 2) are in (b) with TMB rescaled by the mean TMB value (μ_0) of each cancer. The green line is described by $y = x$ in (a) and $y = 1$ in (b). The Pearson correlation coefficient ρ, and the significant P value are given in the figure. (c) Rescaling the TMB distribution for each cancer type gives a Gaussian distribution (with the mean value μ_0, variance σ, and coefficient a_0) fits the data for cancers in Figs. 1 and 3. The rescaled TMB distributions for these cancers collapse into the standard normal distribution (see the green line). (c) Same procedure as in (b) leads to a fat tail power-law distribution ($P(\text{TMB}) = a_0 \times \text{TMB}^b$) is found for data analyzed in Fig. 2. The green linear line has a slope value of -1. Each symbol represents one type of cancer. (e) The TMB across the 16 types of cancer. Each dot represents the TMB of a single cancer patient. The purple solid lines mark the median TMB value for each cancer. The green dashed line shows the boundary between cancers with low and high TMB defined in Figs. 1 and 2. The three cancer types discussed in Fig. 3 are listed at the end.

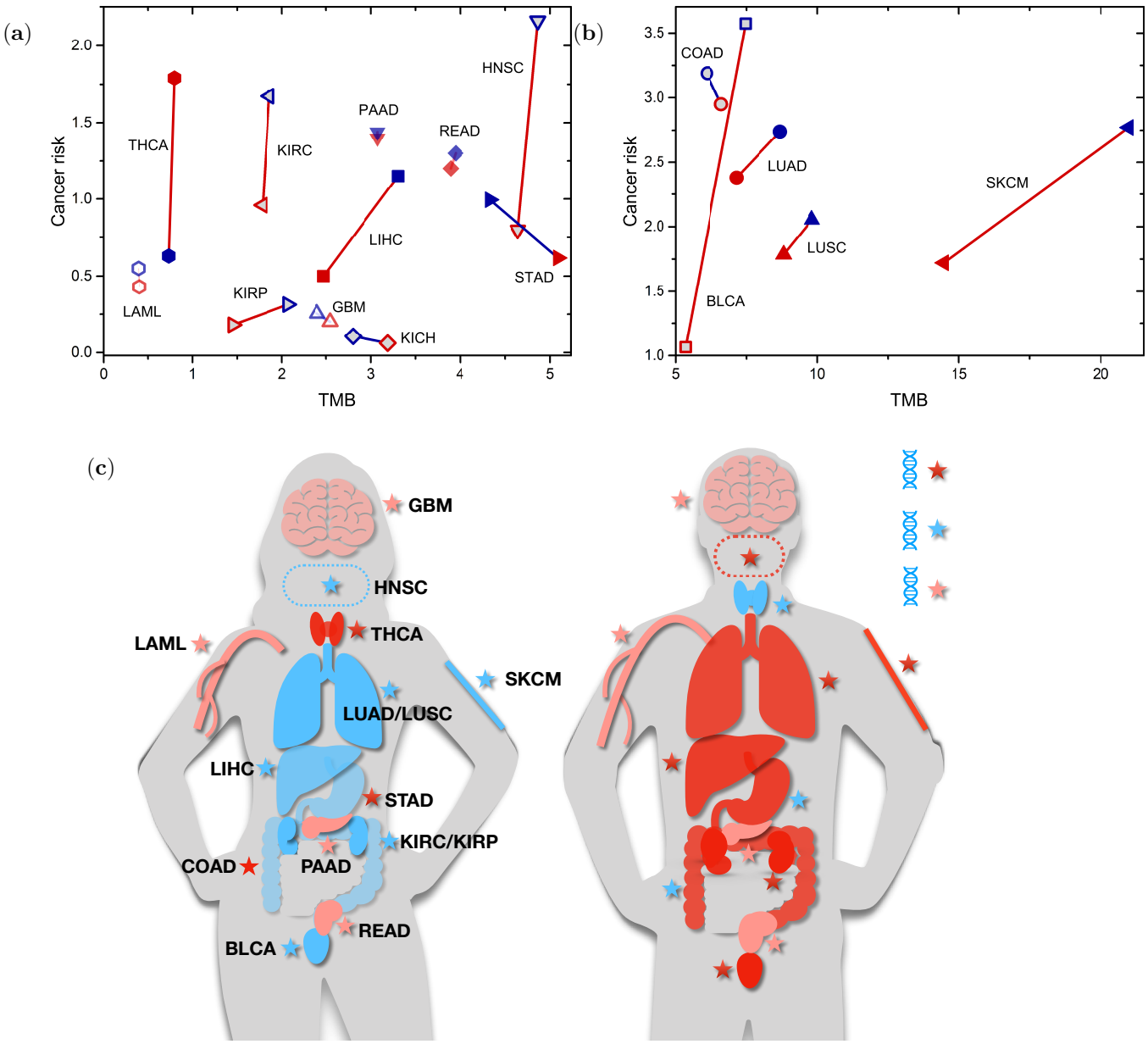

FIG. 5: The relationship between lifetime cancer risk (%) and TMB for both the sexes. (a)-(b) The data are the same as in Fig. S24 (a) and (b) in the SI, while the solid lines (red for a positive, navy for a negative correlation) connect the data for both the sexes for the same cancer type. The data are separated into two groups for better visualization. Red (navy) symbols show the data for females (males). (c) Schematic of the relative cancer risks and the TMB for females and males. The organs in red, blue, and pink colors indicate a higher, lower and similar cancer risk (relative to the other gender), respectively. Similarly, the red, blue and pink stars show a higher, lower and similar TMB in the tissue accordingly. For clarity, we just use three colors to illustrate the relative values of these two quantities between the two genders. The exact values are shown in (a)-(b).

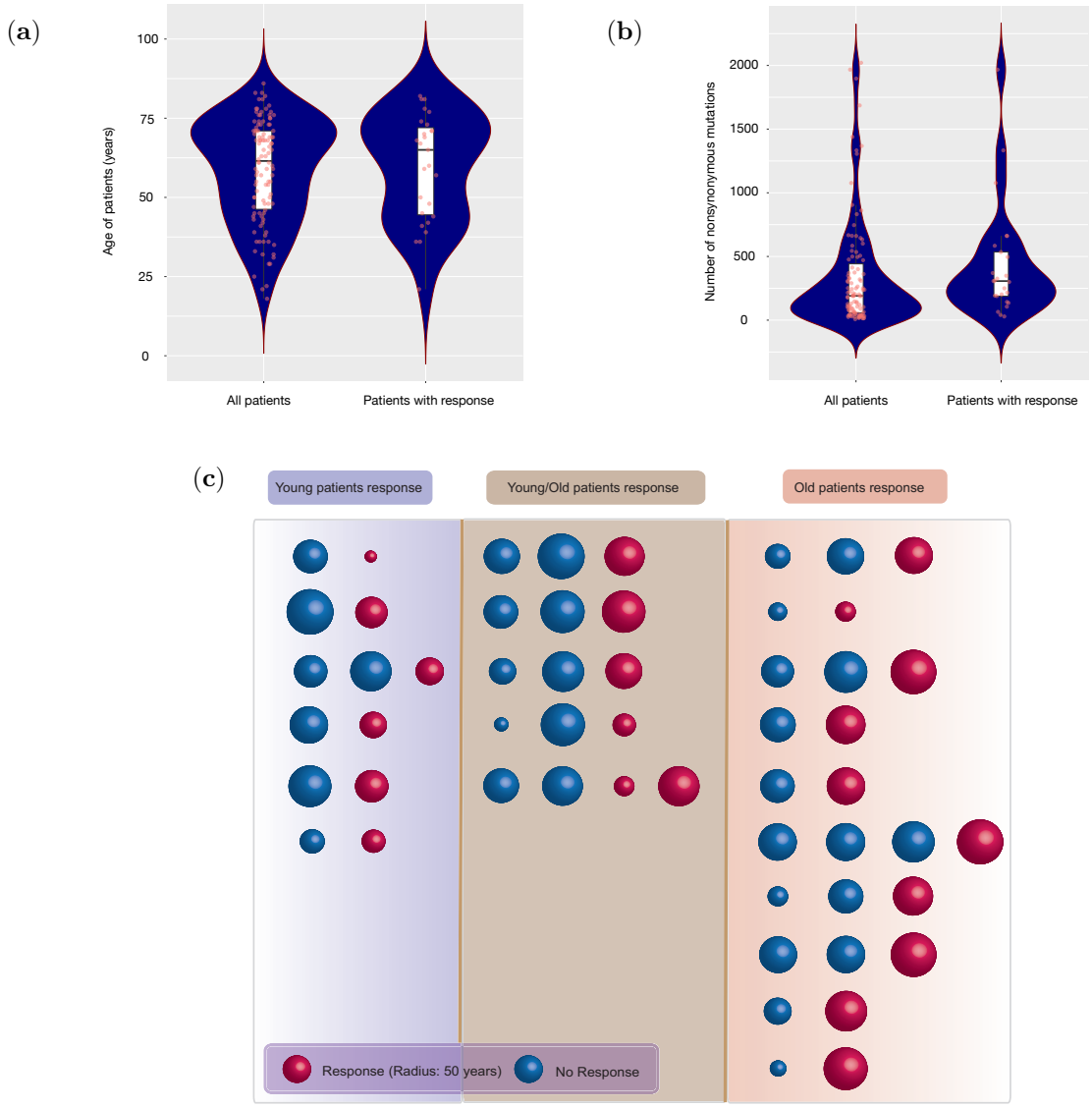

FIG. 6: (a) Violin plots for the age distribution of all patients and the patients showing response to the cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) blockade[60]. (b) Violin plots for the number of non-synonymous mutations for all patients and the patients responding to CTLA-4 blockade. Five data points are not shown in (b) because they are outliers. A box plot is overlaid on each violin plot, and all the data are shown in red dots in the same figure. (c) Schematic of the dependence of response of melanoma patients of differing ages to immunotherapy. The differences in values of the TMB across the rows is within 5%. Red sphere indicates patients who show response to treatment and patients who do not show response are illustrated by blue spheres. The size of the sphere is scaled with patients age as indicated in the lower left corner. The left (right) column gives all pairs in which young (old) patients respond to the treatment. The middle column shows all mixed cases where either young or old patients respond to the treatment.